Alignement automatique de textes parallèles français-japonais

Yayoi Nakamura-Delloye

UNIVERSITE PARIS. DIDEROT (Paris 7)
ECOLE DOCTORALE : Sciences du Langage
Laboratoire LATTICE – CNRS UMR 8094

DOCTORAT
Linguistique Théorique, Descriptive et Automatique

YAYOI NAKAMURA-DELLOYE

Alignement Automatique de Textes Parallèles Français-Japonais

Thèse dirigée par Catherine FUCHS

Soutenue le ? novembre/décembre 2007

JURY

Mme Catherine FUCHS	Directeur de recherche au CNRS, LATTICE (Directeur)
Mme Catherine GARNIER	Professeur à l'INALCO (Co-directeur)
M. Philippe LANGLAIS	Professeur à l'Université de Montréal (Rapporteur)
M. Pierre LE GOFFIC	Professeur à l'Université de Paris III
M. Yves LEPAGE	Professeur à l'Université de Caen (Rapporteur)
M. Pierre ZWEIGENBAUM	Directeur de recherche au CNRS, LIMSI

REMERCIEMENTS

À mes trois amours,
Guy, Noé et Olivier.

TABLE DES MATIÈRES

LISTE DES FIGURES ET TABLEAUX

INTRODUCTION

どのような芸術においても、初めから、人より上手にしようとするのではなく、違うようにしようと思うことである。

Dans tous les arts, il s'agit bien moins, au début, de faire mieux que les autres, que de faire autrement.
— Charles-Augustin Sainte-Beuve

L'alignement automatique consiste à trouver une correspondance entre des unités de « textes parallèles » – ensemble de textes de langues différentes, constitué d'un texte original et de ses traductions. L'alignement peut être réalisé à différents niveaux entre les textes : paragraphes, phrases, mots et expressions.

De nombreuses techniques d'alignement ont été proposées jusqu'à nos jours, le premier travail étant celui proposé par Kay & Röscheisen (1988). Alors que les travaux sur l'alignement des textes parallèles étaient réalisés au départ principalement dans le cadre de la traduction automatique, afin de stocker des exemples de traductions en vue de leur utilisation future, les applications des textes parallèles alignés – appelés parfois « bitextes » ou « multitextes » – sont aujourd'hui extrêmement diverses : constitution de mémoires de traduction, extraction de dictionnaires et de listes terminologiques bilingues, mais aussi extraction de connaissances pour la recherche d'informations multilingues, construction d'exemples pour l'enseignement assisté par ordinateur ou la linguistique constrastive, etc.

Objectifs d'étude et motivation

Dans le cadre de l'exploitation des textes parallèles, nos travaux sont consacrés à la réalisation d'un système qui procède à partir de **textes parallèles français-japonais** à l'alignement notamment au **niveau des propositions**. En effet, ce sujet reste encore un domaine peu exploré : d'une part, les travaux spécifiquement dédiés au traitement bilingue français-japonais sont extrêmement rares, et d'autre part, peu d'applications en traitement automatique des langues introduisent la notion de proposition, l'alignement ne constituant pas une exception.

Nous n'avons sans doute pas besoin de beaucoup de justification pour défendre l'intérêt de travaux bilingues portant sur le couple français-japonais. Étant

donné que chaque langue possède ses particularités, il est toujours intéressant d'étudier spécifiquement le traitement d'un couple de langue donné sans recourir à une langue pivot – typiquement l'anglais – même s'il peut exister déjà beaucoup de travaux à caractère bilingue traitant une des langues à traiter et la langue servant de pivot.

Mais, pourquoi la proposition ?
L'intérêt de l'introduction de cette unité nous semble également considérable. Dès que l'on aborde le traitement automatique des langues écrites sur des données réelles, on sent assez vite la nécessité d'une unité plus petite que la « phrase » (au sens du domaine du TAL : unité entourée de séparateurs graphiques).

En effet, les phrases « naturelles » sont souvent si longues que leur analyse automatique complète est difficile, voire impossible. Nous avons donc cherché une autre unité plus petite : la proposition nous a paru être un bon candidat comme nouvelle unité de traitement. Intuitivement, on sent que c'est une unité représentant un ensemble d'idées, construit mais plus basique que celui exprimé par une phrase.

Il existe des travaux de TAL déjà réalisés partant de cette intuition.

Takeishi & Hayashi (1992) proposent une méthode d'amélioration de la rédaction par segmentation des phrases longues – considérées comme « mauvaises » dans la théorie de la rédaction des documents techniques – en plusieurs propositions plus brèves. Cette idée de paraphrasage par segmentation en plus petites unités est le fondement même de la grammaire transformationnelle de Harris, qui part du constat que « les phrases contiennent d'autres phrases ; autrement dit, dans une phrase S_i, il peut être possible d'identifier une phrase S_j accompagnée de matériel supplémentaire X » (Z. Harris, 1976).

Maruyama et al. (2004) proposent le programme CBAP (*Clause Boundaries Annotation Program*) qui réalise la détection des frontières de propositions du japonais. Ce système a été développé en vue du traitement des monologues (e.g. nouvelles télévisées, conférence, présentation technique) dont les phrases sont considérées par les auteurs comme souvent très longues.

Kashioka et al. (2003) présentent la constitution d'un corpus parallèle avec un alignement au niveau des propositions réalisé avec ce système CBAP. Ils considèrent la proposition comme une unité idéale pour la traduction automatique des monologues et justifient leur choix par la complétude de la proposition aussi bien sur le plan syntaxique que sémantique.

Par ailleurs, dans le domaine de l'analyse discursive, la proposition est souvent considérée comme unité discursive élémentaire.

Certains psychologues ont, de leur côté, présenté les résultats d'expériences de lecture par des sujets, fournissant des éléments favorables à l'hypothèse de l'existence d'une correspondance entre les entités cognitives élaborées au cours de la lecture et les unités linguistiques qui sont des constituants syntaxiques, notamment les propositions (Gineste, 2003).

La segmentation des phrases en propositions est donc très intéressante et utile dans beaucoup de domaines. L'alignement des corpus parallèles au niveau de

cette unité est également profitable. Dans le cas de la constitution d'une mémoire de traduction, par exemple, la réutilisabilité des données est beaucoup plus élevée lorsqu'elles sont segmentées en propositions que quand elles sont constituées de phrases. Plus la phrase est longue, en fait, moins nous avons de chance de trouver une séquence identique dans le texte à traduire.

Nous défendons l'intérêt de l'alignement des propositions par rapport à celui des unités inférieures, en particulier des mots, par le fait que la relation de traduction semble plus fiable au niveau de la proposition entre les langues. En effet, endeçà de la proposition, plus l'unité est petite, plus la correspondance entre deux unités dépend de leur contexte, d'où une portabilité restreinte de leur correspondance.

Ce problème de portabilité restreinte des correspondances des petites unités lexicales s'apparente au problème de leur ambiguïté polysémique dans un contexte monolingue. La « signification [d'une unité polysémique] dépend de la phrase dans laquelle elle est insérée » (Fuchs & Victorri, 1993a). La différence dans le cas du contexte bi- ou multi-lingue est que s'il existait des unités de langues différentes ayant un schéma identique pour la polysémie, l'ambiguïté pourrait être conservée entre les langues et leur caractère polysémique ne poserait pas de problème pour l'alignement et la réutilisation des unités alignées. Or, il est rare, en particulier entre des langues non apparentées, qu'un tel schéma polysémique soit conservé entre les langues et deux unités lexicales en relation de traduction dans un contexte donné n'entretiennent pas forcément le même rapport dans un autre contexte.

Par exemple, le mot français « compte » recouvre plusieurs sens que les Japonais expriment par différents mots. Dans le grand dictionnaire français-japonais de Shogakukan-Robert[1], l'entrée « compte » comporte – en plus de six autres entrées pour les mots composés contenant ce mot – dix définitions constituées, pour la plupart, de noms japonais correspondants, souvent non interchangeables entre les différentes définitions.

Toutefois, lorsqu'il forme un syntagme avec le verbe « tenir », le sens véhiculé est beaucoup plus restreint. La correspondance avec le candidat traduction « 考慮する » (*kôryo suru*) dépend beaucoup moins du contexte, et la probabilité qu'il soit traduit par ce verbe japonais devient très élevée quel que soit le contexte.

Par ailleurs, au niveau du mot, la différence des structures morpho-syntaxiques intervient fortement dans la relation traductionnelle. L'adjectif français « économique », par exemple, trouve généralement comme correspondant dans les dictionnaires l'adjectif (ou qualificatif) japonais « 経済的な » (*keizai teki na*). Certes, « une raison économique » peut être traduite par « 経済的な理由 » (*keizai teki na riyû*). Mais, en japonais, la qualification d'un nom est aussi réalisable par la simple juxtaposition de substantifs en idéogrammes « *kanji* », ce qui permet éventuellement la traduction de l'adjectif « économique » par le simple substantif « 経済 » (*keizai*) comme dans « 経済協力 » (*keizai kyôryoku*, coopéra-

[1] *Shogakukan Robert, Grand dictionnaire français-japonais*. Shogakukan, 1988.

tion économique). Là encore, la correspondance des deux mots indépendants est moins stable que la séquence de plusieurs mots. Plus le contexte est large, plus la portabilité de la correspondance est importante.

Nous avons donc posé comme hypothèse que la proposition était l'unité qui présentait le meilleur équilibre entre réutilisabilité et portabilité de la correspondance. Aussi, la présente thèse est-elle consacrée à la conception d'un ensemble de systèmes réalisant l'alignement des textes parallèles français-japonais au niveau des propositions.

Problèmes à résoudre liés à la notion de proposition

Toutefois, lorsque nous nous attaquons à la tâche concrète de reconnaissance des propositions, nous sommes confrontés à une grande question : qu'est-ce qu'une proposition ?

En effet, quelle que soit l'impression de simplicité que sa familiarité nous donne, cette unité est difficile à identifier de manière automatique : elle n'a aucune indication physique au niveau des caractères, contrairement aux autres unités très utilisées dans les travaux de TAL telles que la phrase et le mot, unités marquées, quoique de façon non univoque parfois, par des moyens graphiques.

Certes, les frontières indiquées par des caractères considérés comme séparateurs ne correspondent pas toujours aux unités que nous croyons traiter, mais ils servent quand même à repérer certains éléments qui sont finalement assez proches de ceux que nous souhaitons manipuler. Cependant, dans le cas de la proposition, il n'y a pas de premier repère indiquant ses frontières *a priori*.

Il nous faut donc trouver d'autres moyens formels – par exemple la présence d'un mot d'une catégorie particulière – permettant de repérer ces unités dans une chaîne de caractères.

C'est pourquoi, dans le cadre de la présente thèse, nous avons tenté de cerner cette notion de proposition avant de réaliser les programmes informatiques la traitant automatiquement.

Schéma général de la thèse

La présente thèse (dont le schéma général est présenté page 23) comporte trois grandes parties : une consacrée aux travaux introducteurs permettant d'instaurer les bases nécessaires pour notre objectif principal, et deux qui constituent le noyau central de notre thèse qu'est l'alignement des textes parallèles français-japonais au niveau des propositions.

La partie dédiée aux travaux introducteurs (**partie I**) comporte l'étude des généralités sur l'alignement ainsi que les travaux consacrés à l'alignement des phrases, opération élémentaire de tout type d'alignement, qui conduisent à la réalisation d'un système d'alignement des phrases adapté au traitement des textes français et japonais.

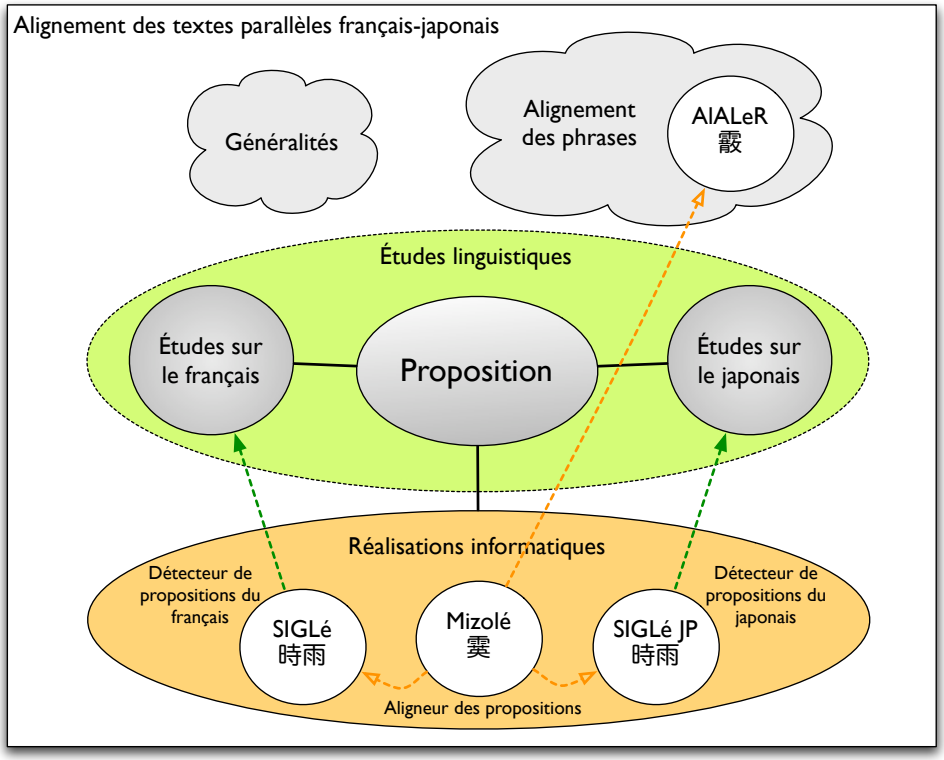


FIG. – Schéma général de la thèse

Le noyau de la thèse s'articule autour de la notion de proposition syntaxique. Il est composé de deux types de travaux, études linguistiques (**partie II**) et réalisations informatiques (**partie III**). Les études linguistiques se divisent elles-mêmes en deux sous-ensembles : la proposition en français et la proposition en japonais. Les réalisations informatiques, décrites dans la dernière partie, comportent trois tâches constituant au final l'opération d'alignement des propositions, incarnées par trois systèmes informatiques distincts : deux détecteurs de propositions (un pour le français et un pour le japonais), ainsi qu'un système d'alignement des propositions. Les deux systèmes de détection des propositions du français et du japonais ont été réalisés sur la base des études linguistiques. Le système d'alignement des propositions, fruit final de la présente thèse, recourt pour la phase de pré-traitement, aux trois autres systèmes développés dans le cadre de nos travaux.

Première partie

Éléments de base de l’alignement

PLAN DE LA PARTIE

La présente partie est consacrée à la présentation des éléments fondamentaux pour l'alignement. Elle comporte deux types de travaux : l'un consacré à l'introduction à l'alignement (**ch. 1**) et l'autre, à l'alignement des phrases, composé de l'étude des techniques existantes (**ch. 2**) et de notre propre réalisation d'un système d'alignement des phrases adapté au traitement des textes français et japonais, le système AlALeR (**ch. 3**).

CHAPITRE 1

GÉNÉRALITÉS SUR L'ALIGNEMENT AUTOMATIQUE DES TEXTES PARALLÈLES

Nous nous intéressons dans ce chapitre à l'ensemble des connaissances de base de l'alignement automatique en général. Pour commencer l'étude des généralités, nous nous intéressons aux ressources de l'alignement : les corpus parallèles (§ 1.1). Nous tenterons ensuite de cerner le concept d'alignement et quelque terminologie connexe (§ 1.2), avant d'aborder les principales applications de cette opération (§ 1.3). L'étude se poursuivra par la typologie de l'alignement (§ 1.4) pour déterminer différentes difficultés selon les classes. Enfin, la dernière partie du chapitre sera consacrée à la description des techniques d'alignement de tout type (§ 1.5).

1.1 Ressources : textes parallèles

Nous allons tout d'abord définir le terme « texte parallèle », que nous utiliserons tout au long de la présente étude pour désigner les données mêmes de l'alignement. Nous présenterons ensuite les caractéristiques et les problèmes de ces corpus, et finirons cette étude par un exposé sur la disponibilité des corpus parallèles notamment sur Internet.

1.1.1 Définition des termes : textes parallèles et comparables

Véronis consacre une des premières pages de son ouvrage « *Parallel text processing* » (Véronis, 2000c) à la définition du terme anglais « *parallel text* », source possible de confusion du fait de significations légèrement différentes selon le domaine où cette terminologie est traditionnellement employée.

Le terme équivalent en français, **texte parallèle**, semble également posséder l'ambiguïté présentée par Véronis pour l'anglais. Quoique, dans le domaine du traitement automatique des langues, le terme « textes parallèles » soit réservé pour désigner « deux ou plusieurs textes de langues différentes, comprenant un texte original et ses traductions », nous constatons parfois des emplois dans un sens proche de celui pour lequel les chercheurs en TAL réservent le terme **textes comparables**, qui fait référence à des textes de même domaine mais de langue différente, l'un n'étant pas une traduction de l'autre.

Dans le cadre de la présente thèse, nous employons le terme « textes parallèles » selon l'usage traditionnel dans le domaine du TAL, à savoir « textes multilingues constitués d'un original et de ses traductions » et le distinguons précisément du terme « textes comparables » désignant, lui, un ensemble de textes sur un même sujet dont aucun n'est traduction de l'un d'entre eux, ou encore un ensemble de textes multilingues sans préciser si l'un est une traduction de l'autre ou non.

Bitexte et multitexte

Par ailleurs, les textes parallèles sur lesquels l'alignement a été réalisé, sont appelés simplement « textes parallèles alignés », ou encore parfois **bitextes** ou **multitextes** (B. Harris, 1988a,b), mais la distinction entre un bitexte et un texte parallèle est encore moins nette dans la littérature.

La distinction de ces deux types d'ensembles de textes est cependant très importante pour l'alignement puisque l'un représente les données d'entrée de l'opération et l'autre le résultat du traitement. Nous conserverons donc strictement, encore une fois, le terme « textes parallèles » pour les documents non alignés, et utiliserons le terme « bitexte » pour désigner les documents déjà alignés.

Terminologie japonaise

En ce qui concerne le japonais, on trouve deux termes équivalents : une transcription phonétique dans un des syllabaires japonais, *katakana*, du terme anglais « *parallel text* » : パラレルテキスト (*parareru tekisuto*, texte parallèle), d'une part ; 対訳テキスト (*tai yaku tekisuto*, « textes avec traductions correspondantes » ou « textes parallèles »), d'autre part. De même, il y a deux équivalents à « corpus parallèle » : パラレルコーパス (*parareru kôpasu*, corpus parallèle), transcription phonétique en *katakana* du terme anglais « *parallel corpus* » ; 対訳コーパス (*tai yaku kôpasu*, « corpus avec traductions correspondantes » ou « corpus parallèles ») qui est l'équivalent du terme anglais « *translation corpora* ». Pour les « corpus monolingues » et « corpus multilingues », les termes, 単言語コーパス (*tan-gengo kôpasu*, mono - langue - corpus) et 多言語コーパス (*ta-gengo kôpasu*, plusieurs - langue - corpus) sont respectivement employés.

On constate également une distinction à l'intérieur des textes parallèles japonais-anglais selon la langue du document original. En effet, la structure des

phrases japonaises diffère considérablement lorsqu'il s'agit d'un texte traduit. La difficulté d'alignement (ou d'extraction d'information multilingue) varie selon la direction de traduction réalisée : avec les textes japonais traduits à partir d'un original anglais, l'analyse est plus facile du fait de la présence systématique de chacun des éléments de la phrase, omis souvent dans une phrase purement japonaise. Ainsi, les corpus japonais-anglais dont l'original est le texte anglais sont appelés 英日コーパス (*ei nichi kôpasu*, anglais - japonais - corpus, « corpus de textes anglais et leurs traductions en japonais »), tandis que les corpus parallèles dont l'original est le japonais sont désignés par le terme 日英コーパス (*nichi ei kôpasu*, japonais - anglais - corpus, « corpus de textes japonais et leurs traductions en anglais »)

1.1.2 Caractéristiques et problèmes des corpus parallèles

En dépit du nombre important de traductions, les textes parallèles compilés en corpus et disponibles dans le domaine public sont assez rares et surtout parmi un ensemble de langues très limitées (notamment l'ensemble des langues européennes et le chinois).

De plus, tous les textes traduits possèdent certaines particularités et nécessitent une certaine prudence lors de leur exploitation.

Premièrement, le type de traduction peut différer selon le type de texte. La traduction des documents ayant un caractère juridique est généralement très fidèle au texte original alors que celle d'autres documents tels que les textes publicitaires est parfois assez différente du texte original, voire une adaptation complète.

Deuxièmement, pour les textes parallèles d'un ensemble de langues données, les caractéristiques des textes peuvent varier selon le sens de traduction. Par exemple, un texte parallèle constitué d'un texte original français et de sa traduction japonaise, peut avoir des caractéristiques différentes des textes parallèles japonais-français dont les originaux sont en japonais.

Enfin, il existe toujours un risque de présence de fautes de traduction (omissions, mauvaises traductions, etc.). Ces erreurs peuvent être dues à l'utilisation de systèmes de traduction automatique ou de traduction assistée par ordinateur.

Lors de la réutilisation des données, il est indispensable de prendre en compte toutes ces caractéristiques et de savoir bien choisir les corpus adéquats. Le risque de présence de fautes est particulièrement problématique : l'inclusion de ces traductions erronées dans une mémoire de traduction entraînerait la reproduction de ces erreurs dans les textes traduits. Ces traductions pourraient à leur tour constituer des textes parallèles exploités pour la constitution d'une mémoire de traduction. La réutilisation de ces traductions pourrait ainsi constituer un cercle vicieux.

Dans le cadre de nos travaux, ces caractéristiques peuvent avoir de l'influence sur la qualité de l'alignement automatique. Il est donc important lors de l'évaluation des systèmes d'alignement, de tenir compte du sens de la traduction des corpus utilisés, pour déterminer correctement leurs performances et problèmes.

Les textes comparables sont débarrassés de tous ces inconvénients des textes parallèles.

Les textes sont « naturels » sans aucune influence d'autres textes et ils n'ont évidemment pas d'erreurs de traduction. L'atout le plus intéressant des textes comparables est leur très grande disponibilité.

Toutefois, l'alignement des textes comparables est beaucoup plus complexe que celui des textes parallèles. Il existe déjà des études sur l'alignement ou l'extraction de mots correspondants à partir de textes comparables et certains (Munteanu & Marcu, 2002) essayent même d'aligner les phrases – mais les résultats sont encore extrêmement limités.

1.1.3 Disponibilité des corpus parallèles

Le développement considérable d'Internet permet non seulement d'accéder à des corpus parallèles compilés, mais aussi de découvrir différents documents multilingues en nombre considérable. Étant donné qu'il existe déjà plusieurs études sur leur disponibilité, à commencer par le panorama présenté par Véronis (2000a,b), nous nous concentrons ici sur la présentation de la situation actuelle de la disponibilité des corpus parallèles comprenant des textes japonais.

Corpus compilés : français-japonais

- **European Corpus Initiative Multilingual Corpus I (ECI/MCI)**
  Le corpus est disponible sur CD-ROM et distribué par ELSNET. Il contient des données parallèles aussi bien dans la plupart des langues européennes que dans d'autres langues telles que le japonais.
- **OPUS**
  Corpus parallèle multilingue aligné, constitué de documents techniques de logiciels *Open Source* : Open Office, PHP Manual, KDE System, KDE Manual.

Corpus compilés : anglais-japonais

- **ATR Dialogue Database**
  Textes parallèles japonais-anglais, créés à partir de transcriptions de dialogues de conférences internationales.
- **Examples for Writing English Business Letter**
  Phrases parallèles d'exemples de lettres japonais-anglais.

Source de textes multilingues : français-japonais

Il est difficile de trouver des textes parallèles (d'un volume significatif) sur un même site. Cependant, on peut parfois constituer des textes parallèles français-japonais en récupérant séparément les documents en français et en japonais.

- **Journal « Le Monde Diplomatique »**
  disponible en vingt-six langues dont le japonais.
  Édition électronique en français : `http://www.monde-diplomatique.fr/`
  Édition électronique en japonais : `http://www.diplo.jp/`
- **Magazine « Label France » du Ministère des Affaires Étrangères**
  disponible en sept langues : français, allemand, espagnol, portugais, italien, russe, japonais.
  (`http://www.diplomatie.gouv.fr/label_france/index.html`)
- **Documents du Sommet d'Évian 2003 (G8)**
  Texte en français : sur le site du sommet Évian
  (`http://www.g8.fr/evian/francais`)
  Texte en japonais : sur le site du premier ministre
  (`http://www.kantei.go.jp/jp/koizumispeech/2003/06/02evian.html`)

Souce de textes multilingues : anglais-japonais

- **Rapports des Ministères**
  Certains ministères publient des Livres Blancs non seulement en japonais mais aussi en anglais, comme par exemple le Ministère de l'Économie, de l'Import/Export et de l'Industrie.
  Rapports japonais : `http://www.meti.go.jp/report/whitepaper/index.html`
  Rapports anglais : `http://www.meti.go.jp/english/report/index.html`
- **Journal Yomiuri**
  Les éditoriaux du quotidien Yomiuri et ceux de sa version anglaise *Daily Yomiuri* peuvent constituer des textes parallèles.
  Version japonaise : `http://www.yomiuri.co.jp/`
  Version anglaise : `http://www.yomiuri.co.jp/index-e.htm`
  Existe sur CD-ROM (`http://www.ndk.co.jp/yomiuri/kijideta/guidance/index.html`)

- ***Scientific Americain* et *Nikkei science***
  Les articles de la revue américaine *Scientific Americain* et de sa version japonaise *Nikkei science* sont utilisés pour l'évaluation d'un système d'alignement développé dans un laboratoire de NTT. Mais les articles en japonais de *Nikkei science* ne sont pas disponibles sur Internet.

1.2 Alignement des textes parallèles

Avant d'entrer dans la discussion sur l'alignement, nous allons tout d'abord présenter la terminologie que nous adoptons pour les données de l'alignement. Nous aborderons ensuite les hypothèses concernant la nature des textes paral-

lèles, sur lesquelles la plupart des méthodes d'alignement de textes parallèles s'appuient. L'exposé se poursuivra par la définition de l'alignement, ainsi que celle de l'unité « phrase », première unité élémentaire de tous nos travaux présentés dans cette thèse.

1.2.1 Conventions terminologiques

Nous appellerons **textes d'entrée** les textes parallèles sur lesquels l'opération d'alignement est appliquée. Comme nous venons de le voir, les textes parallèles sont deux ou plusieurs textes, comprenant un texte original et son/ses traduction(s). Nous désignons désormais le texte original des textes parallèles par **texte source** et son/ses traduction(s) par **texte(s) cible(s)**.

Lors de l'opération de comparaison, opération principale de tous les algorithmes d'alignement, un texte parmi les textes d'entrée sert de base. Toutefois, le choix du texte ne se réfère pas toujours au sens réel de la traduction. En d'autres termes, pour l'alignement des textes parallèles constitués d'un texte original en français et de sa traduction en anglais, le texte français n'est pas forcément utilisé comme base de l'opération. Ainsi, nous appellerons, indépendamment du sens de traduction, **texte de base** le texte servant de base et **texte(s) en regard** le(s) autre(s) texte(s) constituant les textes parallèles d'entrée.

Le terme **de base** peut également être utilisé pour d'autres éléments tels que **langue de base**, **phrase de base**, qui servent, tout comme le texte de base, de base à l'opération.

1.2.2 Hypothèse sur les textes parallèles : parallélisme

Langé & Gaussier (1995) ont défini le caractère de **parallélisme** des textes d'entrée comme condition nécessaire à la réalisation automatique de l'alignement.

Le parallélisme peut être vérifié par deux caractères concrets des textes d'entrée :

- **quasi-bijectivité** : toutes les phrases du texte source ont généralement un correspondant dans le texte cible ;
- **quasi-monotonie** : l'ordre des phrases cibles respecte en général celui des phrases sources.

Mais, la notion de « quasi- » montre la flexibilité de ces conditions. En effet, dans presque toutes les traductions, on constate des contre-exemples de ces principes de bijectivité et de monotonie. D'ailleurs, l'objectif des recherches les plus récentes est souvent l'amélioration de la robustesse des systèmes, afin de pouvoir traiter également de façon correcte les parties qui ne remplissent pas ces conditions de parallélisme.

1.2.3 Définition de l'alignement

À l'instar de la littérature publiée à ce jour, dans la présente thèse, un alignement désigne à la fois une opération et son résultat. Au besoin, le premier est ap-

pelé « opération d'alignement » et le second « résultat d'alignement » pour en préciser la nature exacte.

Opération d'alignement L'opération d'alignement est un ensemble de processus qui reçoit comme données deux ou plusieurs textes $T^{l_1},\ldots,T^{l_n}$ rédigés dans différentes langues $l_1,\ldots,l_n$ et qui produit comme résultat une liste d'ensembles $L=\{P_1,\ldots,P_m\}$ constitués chacun d'un élément $(E_i^{l_1},\ldots,E_j^{l_n})$ de chaque texte d'entrée. Cet élément est une unité ou une séquence d'unités linguistiques, différente selon les programmes, telle que les phrases, les mots ou les unités intermédiaires comme les propositions.

Par cette définition, l'alignement et les éléments d'alignement sont de manière formelle définis comme suit :

Texte d'entrée Soient T^l le texte d'entrée, l la langue du texte d'entrée, u^l l'unité composant le texte et k le nombre total (non nul) d'unités dans le texte. T^l est défini comme un ensemble ordonné de k unités u^l :

$$T^l=\{u_1^l,\ldots,u_k^l\}$$

Élément à aligner Soient E^l l'élément à aligner, n le nombre total d'éléments dans le texte (où $0<n\leq k$), E^l est constitué d'une ou plusieurs unités u^l appartenant à T^l.

$$E_i^l=\{u_j^l|u_j^l\in T^l\wedge 1\leq j\leq k\},\ \text{avec } 1\leq i\leq n$$

C'est donc un sous-ensemble de T^l :

$$E_i^l\subset T^l$$

Soit F^l l'ensemble des éléments à aligner de T^l :

$$F^l=\{E_i^l|E_i^l\in T^l\wedge 1\leq i\leq n\}$$

Toutes les unités appartenant au texte d'entrée doivent appartenir à un et un seul élément de F^l. F^l constitue donc une partition de T^l, et représente également le texte d'entrée mais segmenté de manière différente :

$$T^l=\bigcup_{i=1}^{n}E_i^l=F^l$$

Perle Soient T^l et T^m deux textes d'entrée à aligner écrits respectivement dans les langues l et m. On appelle **perle**[1] l'élément P résultant de l'alignement de deux

[1] Ce terme provient de la terminologie de Brown et al. (1991). Il est la traduction française de l'original en anglais *bead*.

éléments à aligner de chacun des textes d'entrée. On la note :

$$P_i^{lm} = (E_p^l, E_q^m) \quad \text{où} \quad \begin{array}{l} E_p^l \in F^l \\ E_q^m \in F^m \\ E_p^l \text{ ou } E_q^m \text{ est éventuellement nul.} \end{array}$$

On distingue six types de perles selon six modèles de traduction (substitution, suppression, insertion, contraction, extension et fusion[2]) de la langue l vers la langue m :

1. **perle-*lm*** : perle résultant d'une **substitution** d'une unité u^l par une unité u^m. E^l et E^m sont donc constitués d'une seule unité.
2. **perle-*l*** : perle résultant d'une **suppression** d'une unité u^l. Cette perle est composée d'un E^l contenant une seule unité et d'un élément vide représentant l'absence de E^m.
3. **perle-*m*** : perle résultant d'une **insertion** d'une unité u^m. Cette perle est composée d'un E^m contenant une seule unité et d'un élément vide représentant l'absence de E^l.
4. **perle-*l*$^+$*m*** : perle résultant d'une **contraction** par une seule unité u^m de plus d'une unité $u_1^l, u_2^l, \ldots, u_n^l$. Cette perle est composée d'un E^l contenant plusieurs unités et d'un E^m contenant une seule unité.
5. **perle-*lm*$^+$** : perle résultant d'une **extension** d'une unité u^l par plus d'une unité $u_1^m, u_2^m, \ldots, u_n^m$. Cette perle est composée d'un E^l contenant une seule unité et d'un E^m contenant plusieurs unités.
6. **perle-*l*$^+$*m*$^+$** : perle résultant d'une **fusion** de plus d'une unité $u_1^l, u_2^l, \ldots, u_n^l$ avec plus d'une unité $u_1^m, u_2^m, \ldots, u_n^m$. Cette perle est composée d'un E^l contenant plusieurs unités et d'un E^m contenant plusieurs unités.

Résultat d'alignement Un alignement L, résultat de l'opération d'alignement, est constitué d'un nombre t de perles P,

$$L^{lm} = \{P_i^{lm} | P_i^{lm} \in (F^l \times F^m) \wedge 1 \leq i \leq t\}$$

c'est donc un sous-ensemble du produit cartésien de la relation entre F^l et F^m :

$$L \subset F^l \times F^m$$

1.2.4 Définition de la phrase graphique

Nous définissons ici l'unité « phrase » uniquement par des critères graphiques qu'une machine peut traiter sans aucune connaissance particulière pré-acquise.

[2]Cette classification est basée sur l'hypothèse proposée par Gale & Church (1993).

Définition : phrase Simard (1998) présente la définition de la phrase utilisée pour la compilation du corpus BAF[3] comme suit :

> « A Sentence is a syntactically autonomous sequence of words, terminated by a full-stop punctuation. [...] Titles are sentences. [...] Enumerators are sentences. [...] Items of an enumeration are sentences. [...] Each cell in a table is a sentence. [...] »

Nous définissons la phrase, selon la langue dans laquelle elle est écrite, comme suit.

Dans un texte en langue français et anglais, une **phrase** est une séquence de caractères qui se termine par :

- un retour à la ligne (pour les titres et énumérations) ;
- un point d'interrogation ou d'exclamation ;
- un point final, sauf les cas où :
  - il est précédé par moins de 3 caractères (e.g. « 1. ») ;
  - il est suivi directement par un caractère imprimable qui n'est pas un séparateur (e.g. « 1.3 » (en anglais), « abc@xyz.fr ») ;
  - il est dans un sigle ou une abréviation de type « U.S.A. » ou « i.e. » ;
  - il est dans une des abréviations « etc. », « cf. » ou « ex. » ;
  - il est suivi d'un autre point final (il appartient à des points de suspension)[4] ;
- un deux-points ou un point-virgule lorsqu'ils sont suivis d'une espace ;
- un guillemet fermant précédé par un point.

Dans un texte en japonais, une **phrase** est une séquence de caractères qui se termine par :

- un retour à la ligne (pour les titres et énumérations) ;
- un point d'interrogation ;
- un point final japonais « 。 ».

Nous appelons cette unité **phrase graphique** ou simplement **phrase**.

1.3 Applications

L'alignement automatique constitue une sous-tâche de différentes applications et Véronis (2000a) présente le panorama de ces applications. Pour ne citer

[3] Le BAF (Bi-texte anglais français) est un corpus de bitextes anglais-français (disponible sur `http://www-rali.iro.umontreal.ca/arc-a2/BAF/`), c'est-à-dire un ensemble de paires de documents anglais et français, traductions les uns des autres, dont les phrases ont été alignées. Ce corpus a été constitué par l'équipe de traduction assistée par ordinateur (TAO) du CITI, dans le cadre de l'Action de recherche concertée (ARC) A2, coordonnée et financée par l'AUPELF-UREF. La plus grande partie du corpus est constituée de textes de nature institutionnelle (Hansard canadien, rapports de l'ONU, etc.), mais sont aussi inclus quelques articles scientifiques de même qu'une œuvre littéraire. Le tout représente environ 400 000 mots dans chaque langue.

[4] Afin d'éviter une multiplication des règles, nous avons décidé de ne pas prendre en compte les points de suspension apparaissant en milieu de phrase. Nous considérons donc le dernier point des points de suspension comme un indicateur de fin de phrase.

que les principaux, il est utilisé dans le domaine du TAL (recherche d'information multilingue), de la linguistique (lexicographie et terminologie, linguistiques comparatives et contrastives), de l'éducation (enseignement des langues), ou encore dans les recherches et l'étude de la traduction.

Nous pouvons encore y ajouter la traduction automatique (TA) et la traduction assistée par ordinateur (TAO). L'alignement fournit des outils et des ressources utiles dans ces travaux et peut intervenir dans l'automatisation de chacune des étapes : préparation, traduction et révision (Boitet, 2000). Dans la phase de préparation, l'alignement fournit des ressources. Dans les deux autres étapes, l'aligneur intervient en tant qu'outil (Isabelle, 1992). Il permet de visualiser les textes source et cible côte à côte. Il permet également de trouver des omissions de traduction, ou encore, il peut proposer, lorsque d'autres occurrences déjà traduites de la séquence à traduire sont présentes, la traduction correspondante.

Les applications de l'alignement peuvent être divisées d'un autre point de vue en deux catégories : celles utilisant des textes alignés comme données d'entrée, et celles mettant à profit la technique d'alignement elle-même, à l'intérieur d'un traitement global plus complet.

Les applications des textes parallèles alignés – celles de la première catégorie – sont extrêmement diverses : extraction d'information multilingue, constitution de mémoires de traduction, extraction de dictionnaires et de listes terminologiques bilingues, construction d'exemples pour l'enseignement assisté par ordinateur ou la linguistique contrastive.

La plupart de ces applications des textes alignés peuvent également faire appel à la technique d'alignement. En effet, un système d'extraction d'information peut lui-même être équipé d'un programme d'alignement afin d'exploiter directement des textes parallèles non-alignés. De même, certains systèmes d'aide aux traducteurs disposent également d'un aligneur qui initialise et met à jour une mémoire de traduction à partir des traductions passées.

Bien que, lorsqu'on parle de l'alignement, il vienne d'abord à l'esprit, ce premier type ne correspond pas nécessairement à des applications de l'alignement « automatique », car elles ne posent aucune condition quant à la façon de réaliser l'alignement des textes d'entrée : les corpus peuvent tout à fait être alignés à la main. Mais les applications du second type, intégrées dans un ensemble de systèmes, nécessitent un alignement automatique, représentant donc une véritable application de l'alignement automatique.

1.4 Typologie de l'alignement et difficultés de chaque classe

Il est possible de réaliser un alignement entre deux ou plusieurs textes à différents niveaux : paragraphes, phrases, mots et expressions. En d'autres termes, un système d'alignement peut être caractérisé par les unités qu'il envisage d'aligner. Selon l'unité à aligner, les problèmes rencontrés lors de la conception diffèrent

très largement. Nous allons maintenant étudier l'alignement à chaque niveau de façon plus précise.

1.4.1 Alignement de phrases

Depuis la première publication par Kay & Röscheisen (1988) d'un algorithme d'alignement de phrases, de nombreuses méthodes ont été présentées et l'alignement automatique de phrases donne déjà de bons résultats pour l'alignement entre certaines langues, qui ont déjà été beaucoup étudiées. La plupart des méthodes s'appuient sur des hypothèses rendues nécessaires pour des raisons d'efficacité :

- l'ordre des phrases dans les deux textes est identique ou très proche ;
- les textes contiennent peu de suppressions ou d'adjonctions ;
- les alignements 1:1 sont très largement prépondérants et les rares alignements m:n sont limités à de petites valeurs de m et n (typiquement 2).

Ces hypothèses sont cependant, d'une autre manière, source d'inefficacité du système lorsque les textes étudiés ont une structure très différente de ce modèle. Aussi, a été proposée une nouvelle approche (Fluhr et al., 2000) qui consiste à réaliser l'alignement par une méthode de recherche d'information (ou d'interrogation documentaire multilingue) en traitant les textes non pas séquentiellement mais en les transformant en bases de données. Cette méthode permet un meilleur support des structures contredisant ces hypothèses.

1.4.2 Alignement de mots

Beaucoup de méthodes d'alignement de phrases utilisent un alignement de mots. Cependant, dans le cadre de l'alignement de phrases, l'alignement des mots n'est pas le but premier. Lorsque celui-ci est le but premier, les techniques grossières utilisées pour l'alignement des phrases ne sont pas satisfaisantes.

Les mots grammaticaux sont également sources de problèmes : leur correspondance est encore moins nécessaire qu'entre les mots pleins. Néanmoins, il n'est pas possible de les ignorer totalement car ils peuvent faire partie d'une expression à repérer.

Les éléments complexes tels que les mots composés ou les locutions, qui sont largement présents dans les phrases, posent également des problèmes cruciaux. Ainsi, par exemple, l'alignement ou l'extraction de lexiques est théoriquement constitué de deux tâches : repérage dans chaque texte et mise en correspondance des termes extraits dans chaque langue. Mais ces tâches ne peuvent pas être totalement indépendantes car les expressions constituées d'un seul mot graphique dans une langue peuvent être exprimées par plusieurs mots graphiques dans l'autre langue.

Différentes méthodes statistiques ont été proposées, mais les méthodes purement statistiques se heurtent à des difficultés importantes qu'un modèle statistique, du moins simple, ne peut résoudre – comme par exemple les expressions

semi-figées qui supportent des variations de forme. Certains introduisent donc des connaissances linguistiques, mais étant donné leurs coûts relativement élevés et la dépendance à chaque langue, cette solution n'arrive pas à gagner l'approbation de tous les chercheurs.

Par ailleurs, pour les langues n'ayant pas de séparateurs graphiques telles que le japonais, les problèmes se posent d'une autre manière. Ce n'est non pas à un repérage d'unités discontinues que nous avons affaire, mais à la segmentation même de la phrase en unités lexicales. Tout comme pour d'autres applications de traitement automatique des langues, l'alignement ou l'extraction – non seulement d'unités composées, mais d'unités inférieures à la phrase en général – hérite de l'ensemble des problèmes pouvant apparaître dans l'étape de segmentation (voir le chapitre « Méthodes de segmentation » de Nakamura-Delloye (2003a)).

1.4.3 Alignement d'autres unités linguistiques

L'alignement de segments linguistiques supérieurs au terme et inférieurs à la phrase est intéressant dans différents domaines. Même pour l'alignement des mots, en partant de ces unités intermédiaires, il est sans doute possible d'obtenir un meilleur résultat qu'en partant des phrases alignées. Toutefois, l'alignement de ces unités intermédiaires est encore confronté à beaucoup de problèmes : la difficulté de détecter les frontières des unités dans chaque langue, la complexité de l'analyse syntaxique – même partielle –, les grandes divergences de structures entre langues, etc.

1.5 Techniques d'alignement

Nous nous intéressons à présent aux techniques d'alignement. Nous allons présenter de manière brève l'ensemble des techniques d'alignement (de phrases, de mots, et de propositions), dont les principaux algorithmes seront détaillés dans le chapitre 2.

1.5.1 Techniques d'alignement de phrases

Nous allons aborder les méthodes précurseurs, leurs méthodes dérivées et une nouvelle méthode tout à fait différente de ces dernières.

Méthodes précurseurs

La première méthode automatique d'alignement de textes parallèles a été développée par Martin Kay et Martin Röscheisen (Xerox) en 1984 à partir du constat suivant : lorsqu'une personne essaie de mettre en correspondances des phrases de deux textes parallèles, elle compare généralement les mots constituant chaque phrase. De cette intuition, Kay et Röscheisen ont conçu un algorithme d'aligne-

ment (Kay & Röscheisen, 1993) basé sur les informations de correspondance lexicale.

Après cette proposition de méthode exploitant les informations lexicales, Brown et al. (1991) ainsi que Gale & Church (1993) ont présenté leurs méthodes basées sur la corrélation des longueurs de phrases.

Ces deux premiers types de méthodes sont caractérisés par l'utilisation exclusive d'informations internes. Leurs concepteurs ont cherché avant tout la simplicité d'implémentation et de calcul.

Améliorations des premières méthodes

De nombreuses méthodes sont apparues depuis, mais la plupart de celles publiées appartiennent à l'une des deux classes ou combinent les deux méthodes proposées par ces précurseurs. Les méthodes dérivées proposent généralement une amélioration de leurs ancêtres par l'introduction de certaines connaissances linguistiques ou d'un modèle de traduction probabiliste. Enfin, d'autres types d'améliorations sont apportés par l'introduction d'informations externes, notamment les dictionnaires. Les chercheurs japonais y recourent également pour l'adaptation de l'alignement aux textes japonais.

La méthode basée sur la corrélation des longueurs est beaucoup moins efficace lorsque le texte contient beaucoup de phrases. Pour remédier à ce problème, les précurseurs ont introduit une étape de pré-découpage des textes en grandes parties, marquées par un signe quelconque, telles que les paragraphes. Beaucoup ont ensuite cherché une amélioration de cet ancrage.

Les chercheurs tels que Simard et al. (1992) proposent l'utilisation, en combinaison avec des méthodes d'alignement basées sur la corrélation des longueurs, d'un ancrage très simple, qui consiste en un repérage des éléments appelés « cognats ». Il s'agit de chaînes de caractères identiques ou ressemblantes graphiquement, telles que les chiffres, les symboles ou les mots apparentés comme « *language* » en anglais et « langue » en français. Néanmoins, la méthode des cognats ne permet d'obtenir qu'un résultat très limité lors de l'alignement de textes dans des langues non apparentées.

Dans le cadre de l'adaptation de la méthode de Gale au traitement des textes parallèles anglais-chinois, Wu (1994) a présenté une amélioration par l'utilisation d'une liste bilingue anglais-chinois de certains mots clés. Mais cette liste semble étroitement liée à certains corpus donnés, notamment le corpus utilisé, *Hong Kong Hansard*, actes du *Legislative Council* (LegCo). Par conséquent, cette méthode n'a pas apporté une plus grande généralisation que celle des cognats.

Contrairement à cette méthode recourant à une liste étroitement liée à certains corpus donnés, Debili & Sammouda (1992) essaient de profiter davantage d'informations lexicales grâce à l'utilisation d'un dictionnaire bilingue.

Enfin, d'autres chercheurs comme Chen (1993) essayent d'exploiter plus les informations lexicales et proposent l'utilisation d'un modèle de traduction probabiliste.

Mappage et méthodes combinant diverses techniques

Au fur et à mesure que les recherches avançaient, certains chercheurs se sont rendu compte de plus en plus des difficultés de l'alignement, de nature plutôt physique qu'algorithmique, mais fondamentales.

Premièrement, les textes d'entrée contiennent en fait souvent du bruit. C'est dû par exemple au formatage (OCR ou conversion de format, etc.) ou aux erreurs faites par le traducteur. Les différences entre les textes d'entrée provoquées par ce bruit perturbent énormément le programme d'alignement, nécessitant une étape de pré-traitement où est réalisée manuellement une retouche des textes.

Deuxièmement, la reconnaissance même des unités à aligner pose déjà un grand problème pour les développeurs de systèmes. Les symboles considérés généralement comme séparateurs graphiques d'une certaine unité donnée sont souvent polysémiques, empêchant ainsi parfois une segmentation correcte.

Church (1993) ayant remarqué très tôt cette difficulté liée au bruit propose un alignement au niveau des caractères, qui produit des résultats un peu différents de l'alignement classique. En fait, ce n'est pas un alignement proprement dit, mais un mappage qui donne comme résultat un ensemble de paires de points (x, y), où x et y se réfèrent à des localisations précises dans le premier et le second texte respectivement pour dénoter des parties de texte correspondant l'une à l'autre. Ces travaux de Church ont créé une nouvelle optique pour l'exploitation des textes parallèles, engendrant des travaux dérivés (Dagan et al., 1993 ; Fung & McKeown, 1994 ; Melamed, 1996) que Simard a regroupé sous le nom de *bi-text mapping*.

Ce nouveau type de solution au problème d'appariement des textes parallèles est caractérisé par sa robustesse. En effet, comme le dit Church dans la conclusion de son article :

> « *Char_align* has succeeded in meeting many of these goals because it works at the character level and does not depend on finding sentence and/or paragraph boundaries which are surprisingly elusive in realistic applications. »

Son indépendance vis à vis des unités linguistiques extrêmement difficiles à reconnaître correctement, permet de supporter, voire d'ignorer, les problèmes dus au bruit tels que l'omission d'un séparateur ou même l'absence d'une partie de texte dans un des textes d'entrée.

La robustesse de cette méthode a attiré plusieurs chercheurs qui cherchaient un équilibre entre robustesse et précision du système. Les méthodes proposées dans Langlais (1997), Simard & Plamondon (1998) et Kraif (2001) combinent alors une étape de mappage et une étape d'alignement des phrases recourant elle-même à plusieurs indices – longueurs, informations lexicales –, constituant ainsi la dernière génération de l'alignement « classique », les méthodes combinées.

Autres types de méthodes

Enfin, a été proposé un autre type d'algorithme (Fluhr et al., 2000 ; Semmar & Fluhr, 2007), capable de mieux supporter la contrainte des hypothèses utilisées par tous les algorithmes précédents comme mentionné dans la section 1.4.1.

Comme nous l'avons déjà expliqué brièvement, cette approche consiste à réaliser un alignement par la méthode de recherche d'information (ou d'interrogation documentaire multilingue), permettant ainsi de s'affranchir des limites dues aux hypothèses communes des méthodes précédentes. Cette approche très différente des autres consiste à trouver la phrase la plus similaire dans le texte en regard, transformé en base de données, à partir de la requête que constitue la phrase du texte de base.

Méthodes adaptées au traitement du japonais

Les chercheurs japonais proposent généralement des méthodes basées sur les techniques précurseurs adaptées à l'alignement du japonais par l'utilisation de dictionnaires.

Murao (1991) a conçu un système d'alignement s'appuyant sur un dictionnaire bilingue anglais-japonais. Sa méthode exploite les informations de correspondance lexicale comme Kay et Röscheisen, mais pour le calcul il a adopté une méthode de programmation dynamique utilisée dans les algorithmes proposés par Brown et Gale. Le système d'appariement proposé par Utsuro et al. (1994) est basé sur cette méthode de Murao. Ce système a été utilisé par Collier & Takahashi (1995) à l'occasion de la compilation d'un corpus bilingue au *Centre for Computational Linguistics* (CCL, Manchester), constitué d'articles d'un des grands quotidiens japonais Asahi.

Haruno, Yamazaki et Ishihara (Isahara & Haruno, 2000 ; Haruno & Yamazaki, 1996) ont réalisé une adaptation de la méthode de Kay à l'alignement de textes anglais-japonais en recourant également à des dictionnaires bilingues.

Enfin, Hwang & Nagao (1994) a proposé une méthode originale pour le coréen, consistant à traduire chaque phrase du texte de base afin de trouver la phrase correspondante du texte en regard par ressemblance avec cette traduction. Cette méthode permet de ne pas dépendre de la capacité des analyseurs morphologiques coréens, qui ne fournissent pas encore de résultats satisfaisants. De plus, elle met à profit la ressemblance lexicale et structurelle entre les langues japonaise et coréenne, permettant une traduction partielle relativement aisée.

1.5.2 Techniques d'alignement de mots

Alignement de mots et mappage

Comme nous l'avons déjà fait remarquer dans la section 1.4.2, l'alignement de mots est souvent réalisé dans le cadre de l'alignement de phrases – notamment dans la méthode de Kay & Röscheisen (1993) et les méthodes dérivées de cette

dernière –, encore que dans ce cas il peut n'être que partiel et produire en sortie des parties erronées.

Inversement, certains algorithmes d'alignement de mots partent de données déjà alignées au niveau des phrases. Dagan et al. (1993) proposent un algorithme d'alignement de mots utilisant, en raison de leur robustesse, non pas des textes alignés au niveau des phrases mais les résultats de char_align, alignés au niveau des caractères.

Fung & McKeown (1994) proposent un mappage avec les mots comme unités. En effet, la méthode de Church, char_align, qui met en correspondance les même caractères dans les deux textes, ne peut pas être appliquée au traitement des couples de langues écrites dans des alphabets différents. Ainsi, ils ont conçu K-vec, méthode réalisant un mappage non pas avec les caractères mais avec les mots comme unités, et spécialement adaptée au traitement des couples de langues utilisant différents ensembles de caractères tels qu'une langue européenne et une langue asiatique comme le japonais, le chinois ou le coréen. Toutefois, comme char_align, les résultats de cette méthode étant trop partiels pour un alignement de mots, Fung mentionne la possibilité de réalisation d'un alignement plus complet par combinaison avec l'algorithme de Dagan, Church et Gale, présenté dans le paragraphe précédent.

Cependant, l'alignement des mots est plus complexe que celui des phrases. Sans parler des mots grammaticaux pour lesquels une mise en relation est très difficile à effectuer, la correspondance de type 1-1 des mots en général est beaucoup moins évidente que lors de l'appariement de phrases. D'ailleurs, l'appariement 1-1 des unités constituant, par exemple, un mot composé ou une locution, a dans la plupart des cas peu de sens.

Ainsi, beaucoup de chercheurs s'intéressent, plutôt qu'à l'alignement des mots à l'aide des six modèles utilisés pour l'alignement des phrases (voir la définition d'une « perle » dans la section 1.2.3), à la reconnaissance d'unités supérieures ou égales aux mots, représentant des concepts plus faciles à mettre en correspondance, afin de réaliser un alignement au niveau de ces unités.

Alignement d'expressions

Lorsque l'on parle d'alignement d'« expressions »[5], cette opération se rapproche du domaine de l'extraction de lexiques bilingues (ou de l'extraction de terminologies bilingues).

Même si ces deux problèmes ne sont pas totalement identiques, nous ne les distinguerons pas dans cet exposé sur les techniques d'alignement, mais présenterons également des techniques visant l'extraction terminologique, car celles-ci sont suffisamment proches.

Il nous faut d'abord comprendre la différence entre ces deux domaines avant d'entrer dans la présentation des techniques existantes.

[5] Nous utilisons désormais le terme **expression** pour désigner l'ensemble des syntagmes constituant des unités supérieures ou égales aux mots et inférieures aux propositions.

Divergences entre alignement et extraction Kraif (2002) définit l'alignement des correspondances lexicales et l'extraction des lexiques bilingues comme suit :

> « [...] l'alignement lexical, concernant des segments variables en relation d'équivalence traductionnelle, et l'extraction de correspondances lexicales limitée à des couples de lexies équivalentes au niveau des codes linguistiques [...] »

Il fait apparaître par la suite leur différence en disant :

> « Extraire des correspondances lexicales valides au niveau des codes, à partir d'un corpus issu de la pratique concrète de la traduction, ne consiste donc pas à relier chaque mot de la cible avec le (ou les) mot(s) de la source qui entretiennent un rapport traductionnel avec lui, mais à *filtrer* les associations susceptibles de s'extraire de leur contexte. »

Il rend cette différence plus claire grâce aux deux phrases d'exemples suivantes.

fr. [...] *sur l'émission de billets de banque identifiables par les aveugles et par les personnes à vision réduite.*

ang. [...] *on the making of banknote for the benefit of the blind and partially sighted.*

« émission » et « *making* » peuvent, d'après lui, être alignés car il existe un lien de traduction, mais ils ne forment pas de correspondance lexicale, telle que l'on pourrait trouver dans un dictionnaire.

Techniques d'alignement d'expressions et d'extraction de terminologies bilingues L'alignement des expressions, tout comme l'extraction de terminologies bilingues, peut être décomposé en deux phases : une phase de reconnaissance des expressions à aligner dans chaque langue (ou d'acquisition terminologique monolingue) et une phase d'alignement bilingue.

Le système d'acquisition terminologique bilingue de van der Eijik (1993) est constitué – selon ce schéma – d'une étape d'acquisition monolingue et d'une autre d'alignement bilingue. Dans la première étape, les syntagmes nominaux de chaque texte sont extraits sur la base de patrons catégoriels. L'alignement de ces termes extraits est réalisé sur la base de statistiques de co-occurrences des termes dans des phrases alignées.

Termight, développé par Dagan & Church (1994), possède la même architecture. Les termes sont acquis par patrons catégoriels sur un texte étiqueté. L'alignement est réalisé à partir du résultat de l'alignement des mots. La traduction candidate s'étend du premier mot aligné au dernier.

La méthode de Gaussier (1998) se décompose aussi en deux phases, mais elle ne réalise dans la première étape l'extraction de candidats termes que dans la langue pour laquelle l'opération est considérée comme la plus facile, en l'occurrence l'anglais. L'extraction des termes anglais se fait à l'aide de patrons morpho-syntaxiques, et les termes dans le texte français sont repérés au moment de l'alignement.

Il est également possible de concevoir un aligneur d'expressions utilisant une méthode d'extraction terminologique monolingue pour la première phase de reconnaissance des candidats expressions. Il existe déjà un certain nombre de systèmes d'extraction terminologique monolingue.

Termino, application pionnière de l'acquisition automatique de termes, a été développée à l'Université du Québec à Montréal (David & Plante, 1990). Elle réalise, contrairement à l'ensemble des méthodes de l'époque, un traitement syntaxique non statistique du corpus, et est dotée d'une interface de validation permettant à l'utilisateur de sélectionner les résultats corrects parmi ceux qui sont proposés par le logiciel.

Ana, développé par Enguehard & Pantera (1995) à l'Énergie Atomique (CEA), extrait des candidats termes sans effectuer d'analyse linguistique. Ils sont reconnus au moyen d'égalités approximatives entre mots et d'une observation de répétitions de patrons.

Acabit, développé par Daille (1994, 1999) chez IBM, utilise comme données d'entrée un corpus étiqueté et désambiguïsé et combine des traitements linguistiques et des filtrages statistiques.

Lexter, développé par Bourigault (1996), présuppose également des corpus étiquetés et désambiguïsés. Il effectue une analyse syntaxique de surface dédiée au repérage et à l'analyse des syntagmes nominaux. Bourigault a également développé un analyseur syntaxique de corpus, Syntex (Bourigault, 2002), exploitant les résultats de Lexter. Il effectue l'analyse en dépendance de chaque phrase du texte, construisant ainsi un réseau de mots et syntagmes. L'article de Kübler & Frérot (2003) présente une application de Syntex à l'extraction des syntagmes verbaux à partir de textes parallèles anglais-français.

Xtract, développé par Smadja (1993), n'est pas spécifiquement dédié à la terminologie, mais est un outil d'extraction de collocations ne se limitant pas aux termes. L'extraction réalisée par le premier des trois modules composant ce système, repère les collocations à l'aide d'un filtrage statistique.

Fastr, développé par Jacquemin (1997, 1999), est un analyseur syntaxique robuste dédié à la reconnaissance dans les corpus de termes appartenant à une liste contrôlée fournie au système. Le principal objectif est l'identification efficace de ces termes apparaissant sous différentes formes. Il possède donc des métarègles permettant de repérer diverses variations. C'est à la base un outil, non pas d'acquisition de candidats termes, mais d'indexation automatique.

Alignement des mots et des expressions avec des textes japonais

Les systèmes présentés dans la section 1.5.1, ceux de Utsuro *et al* et de Haruno *et al*, alignent à la fois les phrases et partiellement les mots. L'alignement des mots est basé, dans les deux systèmes, sur leur distribution. Mais le problème de ces méthodes est qu'elles ne fournissent que des résultats grossiers et surtout partiellement erronés, mais considérés comme suffisants pour aligner des phrases.

Il existe également un grand nombre d'études sur l'alignement d'unités plus

grandes, les expressions, à partir de textes parallèles alignés au niveau des phrases. Elles visent généralement la reconnaissance des expressions et leur mise en correspondance et se divisent en deux types d'approches selon la méthode adoptée pour l'extraction des candidats : l'approche statistique et celle basée sur les techniques d'analyse syntaxique.

La principale approche statistique est celle utilisant l'extraction de *n*-grammes, ou *chunks*.

Le calcul de *n*-grammes avait comme inconvénient la nécessité d'une quantité importante de mémoire. Nagao & Mori (1994) proposent un algorithme efficace de calcul de *n*-grammes pour un nombre large arbitraire de *n*. Ikehara et al. (1996) proposent une amélioration de l'algorithme de Nagao et Mori en supprimant les chaînes redondantes, nombreuses dans les résultats de l'algorithme précédent. Ils adaptent également cet algorithme à la reconnaissance des collocations non continues.

Sur la base de ces études, Haruno et al. (1996) présentent l'extraction d'expressions par l'algorithme de Ikehara adapté au traitement non pas des caractères mais des mots. La méthode de Haruno et *al.* est également caractérisée par la reconnaissance des expressions non continues par calcul de l'information mutuelle. En effet, pour tous les couples de *n*-grammes extraits du même texte, l'information mutuelle[6] est calculée pour combiner les *chunks* ayant le meilleur résultat. Cette opération est répétée itérativement, constituant ainsi comme résultat une structure arborescente de la phrase. La mise en correspondance de ces séquences est réalisée par une méthode similaire à celle utilisée dans Matsumoto et al. (1993), à savoir la similarité des paires de mots dans les deux langues calculée à l'aide d'un thesaurus.

La méthode proposée par Kitamura & Matsumoto (1997) extrait les *n*-grammes en ne conservant comme candidats que ceux dépassant un seuil de fréquence prédéfini. Cette méthode est caractérisée par le fait qu'au moment de l'analyse morphologique, les mots grammaticaux sont supprimés. Mais le point le plus intéressant de cette étude réside dans le calcul de la similarité. Les auteurs proposent une amélioration du coefficient de Dice par prise en compte du poids représentant la fréquence des co-occurrences (précisé dans la section 2.1.7).

Les méthodes basées sur l'analyse syntaxique utilisent les arbres syntaxiques pour identifier les unités à aligner. Un certain nombre d'études sur l'alignement anglais-japonais de structures inférieures à la proposition ont été réalisées.

La méthode proposée par Yamamoto & Matsumoto (2001) combine la notion de dépendance avec la méthode des *n*-grammes. En effet, elle extrait les *n*-grammes, non pas à partir des phrases, séquences linéaires, mais à partir d'arbres syntaxiques de phrases. L'unité de base n'est pas le mot, mais un syntagme dit *bunsetsu*[7]. Elle commence d'abord par une analyse morphologique pour segmenter les phrases en syntagmes *bunsetsu*. L'opération se poursuit par une analyse

[6] Pour plus de précision, voir la section 2.1.7.

[7] Il s'agit de syntagmes constitués de plus d'un mot autonome et de plus d'un mot grammatical. Pour plus de détails, voir § 5.2.3. Pour les textes anglais, les auteurs considèrent deux types de

syntaxique statistique, qui donne comme résultat les relations de dépendance entre ces *bunsetsu*. L'extraction des candidats est ensuite réalisée, non pas par constitution de nouvelles unités sur la base du modèle adjacent – c'est-à-dire avec un, deux, jusqu'à n syntagmes voisins –, mais par constitution de nouvelles unités sur la base des résultats de l'analyse syntaxique. Par exemple, soit la phrase :

Je monte la tour Eiffel à pied.

Avec le modèle adjacent, on obtient comme candidats de longueur 2 : {(je, monte), (monte, la tour Eiffel), (la tour Eiffel, à pied)}, mais avec les résultats de l'analyse syntaxique indiquant que le syntagme « la tour Eiffel » qualifie le syntagme « monte », les candidats sont {(je, monte), (monte, la tour Eiffel), (monte, à pied)}. Les auteurs expliquent que l'utilisation des relations de dépendance est efficace car ces relations ont tendance à être conservées par la traduction, même pour des couples de langues ayant une structure très différente telles que l'anglais et le japonais.

D'autres méthodes s'éloignent plus de l'approche n-grammes et exploitent plus l'arbre syntaxique[8].

Les travaux de Matsumoto et al. (1993) proposent une méthode permettant de trouver des correspondances structurelles entre deux arbres de dépendance. Dans les méthodes (Kaji et al., 1992 ; Imamura, 2000 ; Watanabe et al., 2000), l'alignement des syntagmes est réalisé sur la base des mots mis en correspondance à l'aide d'un dictionnaire bilingue. Les mots alignés servent à ancrer les textes pour repérer les segments à extraire et la représentation arborescente permet de déterminer correctement les structures formées par ces mots ancres.

Enfin, il existe également des travaux de Fukui et al. (2001) utilisant, non pas des textes alignés au niveau de la phrase, mais des corpus constitués de brevets et alignés au niveau de zones spécifiques au format des brevets. Ces travaux de Fukui et *al.* sont dédiés à l'extraction de lexiques bilingues à partir de brevets japonais-anglais. À partir des zones alignées, l'extraction est réalisée à l'aide de modèles de mots composés appris préalablement à partir d'un dictionnaire de mots spécialisés. Pour la mise en correspondance des mots composés, la méthode de Kitamura et Matsumoto décrite précédemment est utilisée.

Alignement des expressions à partir de textes non parallèles

Il existe également un certain nombre de travaux sur l'alignement des expressions à partir de textes non parallèles, qui permettent d'obtenir des résultats satisfaisants plus facilement qu'avec l'alignement de phrases à partir de textes comparables. Ces études sont généralement basées sur l'hypothèse que les traductions des collocations sont également des collocations, même dans les textes non parallèles (Rapp, 1995 ; Tanaka & Iwasaki, 1996).

séquences comme *bunsetsu* : les syntagmes nominaux basiques – ne contenant pas d'autres syntagmes nominaux –, et les syntagmes nominaux basiques précédés par une préposition.

[8] Des études plus détaillées sur ces méthodes sont présentées dans le chapitre 12.

1.5.3 Techniques d'alignement de propositions

De même que l'alignement des expressions, l'alignement des propositions peut être décomposé en deux phases : une phase de reconnaissance des propositions dans chaque langue et la mise en correspondance de ces unités.

Pour la reconnaissance des propositions, un grand nombre de techniques ont été proposées. La détermination des propositions est souvent réalisée sur la base d'une technique de *shallow parsing* (analyse syntaxique de surface). Aujourd'hui, il existe beaucoup de travaux sur le *shallow parsing* et un état de l'art est présenté par Abney (1997). Des techniques sur le français sont également proposées dans, par exemple, Bourigault (1992) et Abeillé et al. (1998).

Mais, la plupart de ces travaux ont été menés dans le cadre du développement d'un analyseur syntaxique ou de la désambiguïsation syntaxique, et non pas en vue de l'alignement, exceptés les travaux réalisés par Papageorgiou (1997).

Il existe d'ailleurs peu de travaux sur l'alignement des propositions, en dehors de ceux de Piperidis et al. (2000). La mise en correspondance des propositions dans leur technique est réalisée par un calcul de probabilité sur l'association des propositions considérées, avec le type de traduction, leurs longueurs et les informations sur les mots contenus dans les propositions.

CHAPITRE 2

ÉTAT DE L'ART : MÉTHODES D'ALIGNEMENT DES PHRASES

Nous nous intéressons dans ce chapitre aux principales méthodes d'alignement de phrases. Nous allons tout d'abord présenter un des premiers algorithmes, basé sur les informations de correspondance lexicale (§ 2.1). Nous étudierons ensuite les techniques basées sur la corrélation des longueurs de phrases (§ 2.2) et leurs améliorations à l'aide d'informations lexicales (§ 2.3), pour passer à l'étude des alignements combinant différentes informations, dans la recherche d'un équilibre entre robustesse et résolution (§ 2.4). L'exposé se poursuivra par la présentation d'une méthode d'approche originale basée sur les techniques de recherche d'information (§ 2.5). Enfin, pour terminer cette étude, nous aborderons deux méthodes, proposées par des chercheurs japonais, ayant pour objectif l'alignement avec des textes japonais (§ 2.6).

2.1 Méthode basée sur les informations de correspondance lexicale

Cette méthode a été proposée pour la première fois par Kay & Röscheisen (1993). Elle est à l'origine de l'un des deux grands courants des algorithmes d'alignement, qui s'appuie sur l'information lexicale des phrases.

Les auteurs posent tout d'abord deux hypothèses : l'une sur les mots constituants les phrases en relation traductionnelle et l'autre sur la diagonalité de l'alignement.

Basée sur ces hypothèses, la procédure d'alignement de cette méthode est constituée d'un appariement grossier des mots, qui permet ensuite l'alignement

des phrases contenant les mots appariés.

L'étape d'alignement est composée de quatre opérations qui correspondent chacune à la construction d'une structure de données particulière : la table « *Word-Sentence Index* » (WSI), la table « *Alignable Sentence Table* » (AST), la table « *Word Alignment Table* » (WAT) et la table « *Sentence Alignment Table* » (SAT).

Nous allons maintenant étudier plus concrètement les deux hypothèses sur lesquelles cette méthode s'appuie et les quatre structures de données produites au cours de la procédure d'alignement avant d'analyser les avantages et les inconvénients de l'algorithme.

2.1.1 Deux hypothèses

Hypothèse 1 : correspondance des contenus

La première hypothèse est que les phrases correspondantes sont constituées d'éléments correspondants :
soient les textes parallèles,

texte A :			texte B :		
pqr	–	phrase A_1	$\sigma\tau\upsilon$	–	phrase B_1
stu	–	phrase A_2	$\phi\varphi\chi$	–	phrase B_2
υw	–	phrase A_3	$\psi\omega$	–	phrase B_3

Si les différents éléments ont les correspondances ci-dessous :

$p-\sigma$	$s-\phi$	$\upsilon-\psi$
$q-\tau$	$t-\varphi$	$w-\omega$
$r-\upsilon$	$u-\chi$	

la phrase A_1 « pqr » peut être considérée comme correspondante de la phrase B_1 « $\sigma\tau\upsilon$ », la phrase A_2 « stu », celle de la phrase B_2 « $\phi\varphi\chi$ », etc.

Hypothèse 2 : diagonalité de l'alignement

La seconde hypothèse est la correspondance des phrases selon la diagonale de la matrice ayant comme cardinal le nombre de phrases du texte A multiplié par le nombre de phrases du texte B.
La matrice et la diagonale pour les deux textes d'exemple A et B sont ainsi :

$$\begin{array}{cccc} B_3 & \cdot & \cdot & \bullet \\ B_2 & \cdot & \bullet & \cdot \\ B_1 & \bullet & \cdot & \cdot \\ & A_1 & A_2 & A_3 \end{array}$$

$$\text{Diagonale} = \{(A_1, B_1), (A_2, B_2), (A_3, B_3)\}$$

2.1.2 Table « *Word-Sentence Index* » (WSI)

Les occurrences d'un élément *a* n'ont pas forcément la même forme. Si l'une des langues considérées possède le concept de nombre et que *a* est un substantif, il peut apparaître avec ou sans la marque du pluriel. Ou encore s'il s'agit d'une langue flexionnelle, il est possible qu'il change de forme à chaque occurrence.

Afin de résoudre ce problème, l'algorithme commence par une étape préparatoire, pendant laquelle il cherche à réunir les éléments ayant le même contenu sémantique en leur attribuant une « forme normalisée ». Nous appelons ces formes ci-après « lemmes » (ou « formes de base »), encore que les formes obtenues avec la méthode de Kay n'aient souvent rien à voir avec les lemmes obtenus par des méthodes purement linguistiques.

Afin de concevoir un système capable de traiter n'importe quelle langue, les auteurs écartent la possibilité de recourir à des moyens extérieurs tels qu'un dictionnaire ou un analyseur morphologique, choisissant ainsi une méthode ne permettant d'obtenir qu'un résultat assez grossier mais considéré comme suffisant. En effet, le véritable objectif étant l'alignement des phrases, ils considèrent qu'une analyse morphologique très fine et précise n'est pas forcément nécessaire.

Ils posent comme hypothèse que les lemmes sont des sous-chaînes préfixales ou suffixales, donc qu'un mot est divisé en deux parties – dont l'une est le lemme, l'autre appartenant au paradigme de préfixe ou de suffixe. La division est considérée comme sûre si les deux parties apparaissent chacune dans d'autres mots.

Cette division est réalisée à l'aide d'une structure de données appelée *trie* (car élément d'un processus d'« *information re*trie*val* » (Knuth, 1997)). Elle permet de reconnaître les séquences initiales ou finales communes à plusieurs mots.
Soit la liste des mots :

$abcde$
$abcfg$
$abch$
$abcij$

trie nous donne comme information que la chaîne abc est une sous-chaîne initiale des quatre mots.

Les auteurs considèrent ensuite que les séquences communes à plusieurs mots marquent la frontière entre la forme de base et la sous-chaîne affixale. La chaîne abc étant la séquence commune, elle marque dans les quatre mots qui la contiennent une frontière comme suit :

$$\begin{aligned} abcde &\longrightarrow abc + de \\ abcfg &\longrightarrow abc + fg \\ abch &\longrightarrow abc + h \\ abcij &\longrightarrow abc + ij \end{aligned}$$

Après avoir détecté la frontière, on détermine le lemme. Le lemme est la sous-chaîne la plus longue des deux parties divisées par la frontière. Par exemple, la

détection dans le mot graphique « nouveaux » de la frontière entre « nouve » et « aux » entraînerait la caractérisation de « nouve » comme lemme.

Cette forme de base permet également de rassembler plusieurs formes effectives considérées comme des mots représentant le même contenu.

Le lemme rassemble plusieurs formes effectives sous une même forme de base. Prenons comme exemple le lemme « nouve ». Il pourrait regrouper les mots graphiques « nouveaux », « nouvelle », « nouvelles » comme ses formes effectives.

Ainsi, est obtenu le lemme de chaque mot qui constitue la table WSI.

2.1.3 Table « *Alignable Sentence Table* » (AST)

En s'appuyant sur les hypothèses décrites précédemment, la méthode réalise d'abord le calcul de la diagonale représentant les paires de phrases susceptibles d'être alignées et construit la table AST.

Soient les textes parallèles

texte A :			texte B :		
abc	–	phrase A_1	$\alpha\beta\gamma$	–	phrase B_1
ade	–	phrase A_2	$\alpha\delta\theta$	–	phrase B_2
cbf	–	phrase A_3	$\gamma\beta\lambda$	–	phrase B_3
abf	–	phrase A_4	$\alpha\beta\lambda$	–	phrase B_4

La diagonale de $A \times B$ est :

$$\text{Diagonale} = \{(A_1, B_1), (A_2, B_2), (A_3, B_3), (A_4, B_4)\}$$

Lors de l'alignement de textes réels, le calcul de la diagonale est plus compliqué que dans le présent exemple, car le nombre de phrases de chacun des deux textes est généralement différent, et surtout nous devons poser comme hypothèse qu'une phrase peut avoir plus d'une phrase correspondante.

Les deux extrémités des textes (la première phrase de chaque texte et la dernière phrase de chaque texte) sont deux paires dont la relation traductionnelle est quasiment sûre, paires que nous appellerons « ancres ». Ainsi, la première phrase d'un texte est associée avec celle de l'autre texte et sa dernière phrase avec celle de l'autre texte.

Pour les autres phrases situées entre ces deux ancres, la $j^{\text{ème}}$ phrase du texte A est associée avec plusieurs phrases du texte B aux positions proches de la diagonale. Plus la phrase considérée s'éloigne de l'ancre, plus le nombre de phrases avec lesquelles elle est associée est important.

Les paires de phrases susceptibles d'être alignées sont donc beaucoup plus nombreuses que le cas simple que présente l'exemple, ce qui entraîne plus de calcul dans les autres étapes également.

2.1.4 Table « *Word Alignment Table* » (WAT)

Les éléments d'une paire de phrases susceptibles d'être alignées sont ensuite comparés afin de calculer leur similarité de distribution.

Les distributions de chaque élément de l'exemple sont :

distribution de $a = \{A_1, A_2, A_4\}$
distribution de $b = \{A_1, A_3, A_4\}$
distribution de $c = \{A_1, A_3\}$
distribution de $d = \{A_2\}$
distribution de $e = \{A_2\}$
distribution de $f = \{A_3, A_4\}$
distribution de $\alpha = \{B_1, B_2, B_4\}$
distribution de $\beta = \{B_1, B_3, B_4\}$
distribution de $\gamma = \{B_1, B_3\}$
distribution de $\delta = \{B_2\}$
distribution de $\theta = \{B_2\}$
distribution de $\lambda = \{B_3, B_4\}$

Si un certain nombre de paires, constituées d'une occurrence de chaque élément à comparer, coïncident avec des paires de phrases susceptibles d'être alignées (i.e. la diagonale), ces deux éléments sont considérés comme éléments correspondants.

Par exemple, nous obtenons à partir des distributions de a et de α, les paires de phrases suivantes qui coïncident avec les paires de phrases susceptibles d'être alignées :

$$\{(A_1, B_1), (A_2, B_2), (A_4, B_4)\}$$

Nous considérons donc les éléments a et α comme éléments correspondants.

Autrement dit, plus le cardinal de l'intersection entre le produit cartésien des distributions de deux éléments et la diagonale est proche du nombre moyen de cardinaux de la distribution de ces deux éléments, plus la probabilité de correspondance de ces deux éléments est élevée.

Le produit cartésien des distributions de a et de α est :

$$\begin{aligned} R = \{ \quad & (A_1, B_1), (A_1, B_2), (A_1, B_4), (A_2, B_1), \\ & (A_2, B_2), (A_2, B_4), (A_4, B_1), (A_4, B_2), (A_4, B_4) \quad \} \end{aligned}$$

L'intersection entre la relation ci-dessus et la diagonale est :

$$R \cap \text{Diagonale} = \{(A_1, B_1), (A_2, B_2), (A_4, B_4)\}$$

La similarité des éléments est calculée à partir de ce cardinal de l'intersection à l'aide du coefficient de Dice (van Rijsbergen, 1979) :

$$\text{similarité} = \frac{2\left|R \cap \text{Diagonale}\right|}{|\text{distribution de } a| + |\text{distribution de } \alpha|} = 1$$

La similarité pouvant être comprise entre 0 et 1, une valeur 1 signifie que les éléments (ici a et α) sont considérés comme éléments correspondants.

Nous obtenons ainsi les paires d'éléments supposés correspondre :

(a,α) : distributions $a \times \alpha = \{(A_1,B_1),(A_2,B_2),(A_4,B_4)\}$
(b,β) : distributions $b \times \beta = \{(A_1,B_1),(A_3,B_3),(A_4,B_4)\}$
(c,γ) : distributions $c \times \gamma = \{(A_1,B_1),(A_3,B_3)\}$
(d,δ) : distributions $d \times \delta = \{(A_2,B_2)\}$
(e,θ) : distributions $e \times \theta = \{(A_2,B_2)\}$
(f,λ) : distributions $f \times \lambda = \{(A_3,B_3),(A_4,B_4)\}$

Ces paires d'éléments appariés constituent la table `WAT`.

Par ailleurs, cette méthode de calcul de la similarité est étudiée par beaucoup de chercheurs et un grand nombre d'améliorations ont été proposées. Nous présentons l'ensemble de ces travaux dans la section 2.1.7.

2.1.5 Table « *Sentence Alignment Table* » (SAT)

La procédure se poursuit par le calcul du nombre d'éléments correspondants que contient chaque paire de phrases susceptibles d'être alignées afin d'apparier les phrases et de construire la table `SAT`.

(A_1,B_1) contient (a,α), (b,β) et (c,γ)
(A_2,B_2) contient (a,α), (d,δ) et (e,θ)
(A_3,B_3) contient (b,β), (c,γ) et (f,λ)
(A_4,B_4) contient (a,α), (b,β) et (f,λ)

Si les correspondances sont justifiées par plusieurs éléments correspondants, alors les phrases sont considérées comme alignées. Ainsi, dans notre exemple, les phrases A_1 et B_1 sont alignées, de même que A_2 et B_2, A_3 et B_3 et A_4 et B_4.

Ce calcul de l'associativité des paires de phrases est semblable à celui de l'étape de création de la table `WAT` sauf qu'il exige une désambiguïsation des paires de phrases. Une paire de phrases est considérée comme ambiguë si une des occurrences d'un mot peut être associée avec plus d'une occurrence du mot avec laquelle elle est appariée.

Par exemple, supposons que la paire de mots (M_i^A, M_j^B) appartienne à la table `WAT`, et que l'ensemble des phrases $\{P_r,...,P_s\}$ soit les occurrences de M_i^A et $\{P_t,...,P_u\}$, celles de M_j^B.

Bien que (P_r,P_t) appartienne à la table `AST`, s'il y a une autre occurrence de M_i^A telle que (P_v,P_t) appartient à la table `AST` ou qu'il y a une autre occurrence de M_j^B telle que (P_r,P_w) appartient à la table `AST`, alors la paire (P_v,P_t) est considérée comme ambiguë et n'est pas prise en compte, de même que l'autre paire (P_r,P_w) ou (P_v,P_t) est ignorée.

Les paires de phrases de la table `SAT`, qui ont été associées par un certain nombre de paires de mots, sont considérées comme des « ancres ». Toutes les opérations consistent en fait à trouver les paires de phrases entre deux ancres, qui deviendront elles-mêmes des ancres aux itérations suivantes. Ce processus est ainsi recommencé de manière à trouver, entre deux ancres obtenues lors de précédentes itérations, de nouvelles paires de phrases qui seront considérées par

la suite comme de nouvelles ancres. Ces opérations sont répétées jusqu'à ce que toutes les phrases du texte A soient associés avec celles du texte B.

2.1.6 Algorithme général

Nous présentons dans cette section l'algorithme dans sa totalité.

Algorithme 1 Alignement des phrases de Kay et Röscheisen

♢ Données :
- Texte 1 et Texte 2 ;
- m : nombre de phrases du Texte 1 ;
- n : nombre de phrases du Texte 2 ;
- `WSI` : listes des mots (pour chaque texte) ;
- `AST` : liste des paires de phrases susceptibles d'être alignées ;
- `WAT` : liste des paires de mots alignés ;
- `SAT` : liste des phrases alignées ;

♢ Procédure :

1. Création de la table `WSI` : étape morphologique
   Réalisation pour chaque texte des étapes suivantes :
   a) extraction des mots graphiques composant le texte ;
   b) lemmatisation de chaque mot graphique
      i. détermination de la frontière décomposant la forme effective en deux parties à l'aide de *trie*.
      ii. s'il y a plusieurs frontières potentielles, alors :
      calcul pour chaque frontière potentielle de la valeur $kP(p)S(s)$, où $P(p)$ est le nombre d'occurrences dans le texte de la chaîne préfixale p, $S(s)$ celui de la chaîne suffixale s, et $k = longueur(p)$, la quantité k permettant de favoriser un découpage en chaîne préfixale longue.
      iii. détermination du lemme.
      iv. s'il existe un ou plusieurs mots graphiques contenant une chaîne semblable au lemme obtenu, alors :
      ce lemme est considéré comme étant également la forme de base de ces mots graphiques.
   c) repérage des phrases auxquelles chaque lemme appartient et calcul de leur fréquence.
2. Initialisation de la table `SAT` :
   a) création de la paire comprenant la première phrase du texte A et la première phrase du texte B ;

b) création de la paire comprenant la dernière phrase du texte A et la dernière phrase du texte B ;

c) stockage de ces deux paires dans la table SAT.

3. Calcul du seuil

4. Répéter f fois les opérations suivantes, f fixé préalablement

I. Création de la table AST contenant des paires de phrases dont l'alignement est envisageable :
pour toutes les paires de phrases de la table SAT ($\texttt{SAT}_i$ tel que $0 < i \leq n$)

a) stockage des $\texttt{SAT}_i$ et $\texttt{SAT}_{i+1}$ dans la table AST ;

b) pour tout autre phrase située entre $\texttt{SAT}_i$ et $\texttt{SAT}_{i+1}$,
la $j^{\text{ème}}$ phrase du texte A est associée avec $\sqrt{2j}$ phrases du texte B aux positions proches de la diagonale, c'est-à-dire la $j \cdot n/m^{\text{ème}}$ phrase du texte B pour la $j^{\text{ème}}$ phrase du texte A.

II. Création de la table WAT, liste des paires de mots alignés :

a) comparaison de tous les mots de la $i^{\text{ème}}$ phrase P_i du texte A avec tous les mots de la $j^{\text{ème}}$ phrase P_j du texte B telles que $(P_i^A, P_j^B) \in \texttt{AST}$:
soient $P_i^A = M_{i_1} ... M_{i_m}$ et $P_j^B = P_{j_1} ... P_{j_n}$, pour toutes les paires (M_{i_p}, M_{j_q}), calcul de la similarité de leurs distributions.

b) mise en ordre décroissant des paires de mots selon leur similarité et leur fréquence.

III. Création de la table SAT, liste des paires de phrases alignées :

a) calcul de l'associativité des paires de phrases :
pour toutes les paires de phrases de la table WAT ($\texttt{WAT}_i$ tel que $0 < i \leq n$) :
soit $\texttt{WAT}_i = (M_p^A, M_q^B)$, pour toutes les occurrences $\{P_r, ..., P_s\}$ et $\{P_t, ..., P_u\}$ de M_s^A et M_t^B, si $(P_v, P_w) \in \texttt{AST}$, $\neg\exists x\ t \leq x \leq u \wedge (P_v, P_x) \in \texttt{AST}$ et $\neg\exists y\ r \leq y \leq s \wedge (P_y, P_w) \in \texttt{AST}$,
alors incrémenter $\texttt{SAT}(P_v, P_w)$ de 1.

b) suppression des paires qui n'ont pas atteint le seuil :
si $\texttt{SAT}(P_v, P_w) < \texttt{seuil}$, alors $\texttt{SAT}(P_v, P_w) := 0$.

IV. Revenir à la création de la table AST (étape I).

2.1.7 Améliorations par des travaux postérieurs : différentes formules de calcul de similarité des distributions lexicales

La concept d'appariement des mots, basé sur la comparaison de leurs distributions, est repris non seulement par les travaux sur l'alignement des phrases mais aussi par des chercheurs travaillant sur l'alignement des mots. Ainsi, de nombreuses formules ont été proposées depuis pour le calcul de la similarité des distributions lexicales. Des évaluations et comparaisons de ces formules ont même été réalisées par Matsumoto et *al* dans Matsumoto & Utsuro (2000).

Nous présentons maintenant les principales méthodes de calcul de similarité.

Méthode de Kay

La méthode de calcul de similarité de Kay utilise, comme nous venons de le voir, le coefficient de Dice.

> Soient m_a et m_b les mots considérés, $\mathsf{freq}(x)$ la fréquence du mot x, et $\mathsf{freq}(x, y)$ la fréquence de co-occurrence des mots x et y apparaissant dans les mêmes perles.

$$\text{similarité} = \frac{2 \cdot \mathsf{freq}(m_a, m_b)}{\mathsf{freq}(m_a) + \mathsf{freq}(m_b)}$$

Méthode de Gale

Gale & Church (1991) présentent une autre méthode de calcul de similarité.

Soient :

N = nombre total de perles

$a = \mathsf{freq}(m_a, m_b)$

$b = \mathsf{freq}(m_a) - \mathsf{freq}(m_a, m_b)$

$c = \mathsf{freq}(m_b) - \mathsf{freq}(m_a, m_b)$

$d = N - a - b - c$.

$$\begin{aligned} \text{similarité} &= \frac{(ad - bc)^2}{(a+b)(a+c)(b+d)(c+d)} \\ &= \frac{(ad - bc)^2}{\mathsf{freq}(m_a)\mathsf{freq}(m_b)(N - \mathsf{freq}(m_a))(N - \mathsf{freq}(m_b))} \end{aligned}$$

Mais ces deux méthodes produisent, d'après Utsuro et al. (1994), des résultats assez proches.

Méthode de BACCS

Dans le système d'alignement adapté au japonais BACCS (*Bilingual Aligned Corpus Construction System*, voir la section 2.6.3), la similarité des mots m_a et m_b est calculée à l'aide d'une matrice appelée matrice de contingence (Fung & Church, 1994).

La matrice est constituée de quatre cases comme représenté dans le tableau 2.1 (voir page suivante). Les cases prennent les valeurs a, b, c et d, définies dans la description de la méthode de Gale présentée précédemment.

Si les mots m_a et m_b sont des traductions mutuelles, la valeur a sera élevée alors que b et c seront de petites valeurs. En revanche, si les mots m_a et m_b sont des mots n'ayant aucun rapport, a sera de faible valeur tandis que celles de b et

	m_a	
m_b	a	b
	c	d

TAB. 2.1 – Matrice de contingence

c seront élevées. Pour refléter plus précisément ces différentes conditions, l'information mutuelle est introduite :

$$\log\left(\frac{\mathsf{prob}(m_a, m_b)}{\mathsf{prob}(m_a)\mathsf{prob}(m_b)}\right)$$

Chacune de ces probabilités est calculée comme suit :

Soit $M = a + b + c + d$.

$$\mathsf{prob}(m_a) = \frac{a+c}{M}$$

$$\mathsf{prob}(m_b) = \frac{a+b}{M}$$

$$\mathsf{prob}(m_a, m_b) = \frac{a}{M}$$

L'information mutuelle étant moins fiable lorsque le nombre d'occurrences est restreint, on introduit le t-score afin d'évaluer sa fiabilité :

$$t \approx \frac{\mathsf{prob}(m_a, m_b) - \mathsf{prob}(m_a)\mathsf{prob}(m_b)}{\frac{1}{M}\mathsf{prob}(m_a, m_b)}$$

La relation traductionnelle de ces mots est donc évaluée à partir de ces deux valeurs, information mutuelle et t-score. L'information mutuelle insignifiante est alors filtrée par le t-score.

Amélioration du coefficient de Dice

La méthode proposée par Kitamura & Matsumoto (1997) présentée dans la section « Alignement des expressions avec les textes japonais » au § 1.5.2, utilise pour le calcul de la similarité des expressions, le coefficient de Dice avec une amélioration consistant en la prise en compte du poids représentant la fréquence des co-occurrences.

Les auteurs comparent d'abord les méthodes basées sur le coefficient de Dice et sur l'information mutuelle, et tirent comme conclusion que la meilleure performance est celle basée sur le coefficient de Dice. Cette meilleure efficacité du coefficient de Dice est également signalée par d'autres chercheurs tels que Omori et al. (1996) et Smadja et al. (1996). En effet, comme il a été affirmé dans Dagan et al. (1993), Haruno et al. (1996) et Omori et al. (1996), l'information mutuelle impose comme condition que les paires de candidats doivent avoir une fréquence suffisante pour que les résultats soient corrects.

Néanmoins, le calcul basé sur le coefficient de Dice a également comme défaut de ne pas pouvoir refléter le nombre de phrases où les candidats apparaissent. Par exemple, deux chaînes de mots apparaissant deux fois dans les même perles auraient la similarité maximum 1, tout comme deux chaînes apparaissant cent fois dans les même perles. Cependant, la possibilité de correspondance est, d'après leur étude, plus élevée pour les candidats de fréquence cent que pour ceux de fréquence deux.

Ainsi, Kitamura et Matsumoto introduisent le poids reflétant la fréquence des co-occurrences. La formule proposée est comme suit.

Soient e_a et e_b les expressions considérées et $p(\mathsf{freq}(e_a, e_b))$ le poids basé sur la fréquence des co-occurrences.

$$\mathsf{sim}(e_a, e_b) = p(\mathsf{freq}(e_a, e_b)) \cdot \frac{2 \cdot \mathsf{freq}(e_a, e_b)}{\mathsf{freq}(e_a) + \mathsf{freq}(e_b)}$$

2.1.8 Caractéristiques de ces méthodes : avantages et inconvénients

Point faible de l'appariement des mots basé sur la distribution

Le problème de l'appariement des mots basé sur la distribution est qu'il ne peut fournir que des résultats grossiers et surtout partiellement erronés, bien que considérés comme suffisants pour aligner des phrases.

Néanmoins, les mots de fréquence faible, notamment ceux de fréquence 1, ont parfois une influence non négligeable sur les résultats d'alignements de phrases. En effet, ce sont les mots les plus difficiles à aligner pour ces méthodes.

Par exemple, si les mots « ordinateur » et « comptoir » sont de fréquence 1 et apparaissent dans la même phrase française et que la phrase anglaise alignée avec elle contient un mot de fréquence 1, « *computer* », il est impossible de savoir lequel des deux correspond à ce mot anglais. Ce que Tsuji et al. (2000) appellent **situation de détermination impossible**.

Afin d'améliorer les techniques d'alignement des mots, ils ont étudié les caractéristiques des mots de fréquence faible dans les corpus dans le cadre de recherches sur l'extraction automatique de lexiques bilingues. Cette étude a montré que la suppression des mots figurant déjà dans les dictionnaires et la lemmatisation des mots anglais étaient insuffisantes pour améliorer les situations de détermination impossible, et que les méthodes statistiques basées sur les distributions lexicales étaient de manière générale insuffisantes, nécessitant donc l'emploi d'autres techniques, comme par exemple la création de règles de traduction au niveau des caractères ou encore un pré-alignement plus précis.

Lorsqu'un texte contient beaucoup de mots de fréquence faible, l'aligneur utilisant une méthode d'appariement des mots basée sur la distribution ne réussit à mettre en correspondance que peu de mots, d'où un alignement de phrases également très limité.

Indépendance par rapport aux langues à traiter ?

Cet algorithme est caractérisé par l'utilisation unique d'informations internes. D'après les auteurs, cette stratégie a comme avantage une indépendance vis-à-vis des langues à traiter. En effet, ils ont réussi à développer des étapes morphologique et de recherche des correspondances en général, sans recourir à un dictionnaire ou à une liste de connaissances linguistiques quelconque.

Cependant, cette prétention d'être applicable à toute langue fait l'impasse sur toute une catégorie de langues. En effet, la première opération élémentaire d'extraction des mots est basée sur l'existence de séparateurs graphiques. Or certaines langues, comme le japonais, ne possèdent pas de signes permettant de segmenter les phrases *a priori*.

Bien que l'introduction de certaines connaissances linguistiques soit bénéfique voire indispensable, l'idée générale des travaux de Kay est intéressante et performante, dans la mesure où les moyens extérieurs tels que des dictionnaires bi/multilingues ou des analyseurs morphologiques n'existent pas encore pour toutes les langues et surtout pour toutes les paires de langues.

Si certains considèrent cet algorithme comme une méthode probabiliste, il est possible de trouver assez facilement des justifications linguistiques. En effet, le sens principal d'un mot graphique est généralement porté par les radicaux, les suffixes représentant la fonction grammaticale – ou du moins des informations « secondaires » – telle que le nombre ou le genre (quoique les affixes, éléments un peu délicats, puissent modifier le sens principal). Si bien que leur méthode de reconnaissance des formes de base – trouver des chaînes préfixales (ou parfois suffixales) des formes effectives de manière à rassembler plusieurs formes effectives sous une même forme de base –, est tout à fait logique et évoque même les travaux des linguistes distributionalistes.

2.2 Méthodes d'alignement basées sur la corrélation des longueurs

Nous nous intéressons à présent au deuxième type de méthodes d'alignement, celles basées sur la corrélation des longueurs, qui a engendré beaucoup de techniques dérivées.

Ce sont des méthodes dérivées de l'étude de Brown et al. (1991) (méthode de Brown ci-après) ainsi que de celle de Gale & Church (1993) (méthode de Gale ci-après). Cette idée d'alignement selon les longueurs a tout d'abord été présentée par l'article de Brown et al. (1990) dans le cadre de travaux sur la traduction automatique. Les auteurs y présentent très brièvement la méthode utilisée pour aligner les phrases du Hansard, actes du parlement canadien, en vue de la spécification des paramètres du modèle de traduction.

Contrairement aux créateurs de la méthode précédente – méthode basée sur les informations de correspondance lexicale, cf. section 2.1 –, ils ne font aucune

hypothèse sur le contenu des phrases, mais ils utilisent comme point de départ le fait que les longueurs de phrases du texte original ont un rapport logique avec celles des phrases traduites. En d'autres termes, à des phrases longues correspondent des phrases longues et des phrases courtes sont traduites par des phrases courtes. En représentant la longueur d'une phrase par son nombre de mots ou de caractères, les auteurs construisent un modèle probabiliste et une mesure de dissimilarité entre les phrases à aligner.

Brown, Lai et Mercer utilisent des modèles de Markov cachés. Gale et Church proposent quant à eux une méthode consistant à trouver l'alignement optimal qui minimise la mesure de dissimilarité cumulée sur l'ensemble du texte par un calcul basé sur un algorithme classique de programmation dynamique.

Nous allons maintenant étudier les grandes lignes de fonctionnement, en tenant compte des différences entre ces deux algorithmes. Nous aborderons ensuite les travaux de Wu (1994), exemple d'adaptation à une langue non indo-européenne. Enfin l'étude se terminera par l'analyse des avantages et des inconvénients de ces méthodes.

2.2.1 Description de la méthode

La méthode de Brown, ainsi que celle de Gale, réalisent toutes deux le traitement en deux opérations : un pré-alignement grossier – c'est-à-dire un alignement au niveau section ou paragraphe – puis l'alignement des phrases à proprement parler.

Pré-alignement

On considère les indices de section tels que les titres conventionnels ou les signes typographiques comme points d'ancrage.

Méthode de Brown La méthode de Brown distingue d'abord ces points d'ancrage en deux types : petite ancre et grande ancre (*minor* et *major* en anglais). En général, les grandes ancres sont systématiquement traduites et les petites ancres sont parfois omises dans la/les traduction(s). L'alignement des points d'ancrage est ensuite réalisé en deux passages, le premier alignant les grandes ancres et le second les petites ancres. Au premier passage, on assigne à tous les couples possibles de grandes ancres des deux textes un coût compris entre 0 et 10 selon la similarité des deux séquences de caractères, la similarité maximum étant représentée par un coût nul. On réalise ensuite un alignement en considérant la minimisation de ce coût comme un problème standard de programmation dynamique. Le premier passage transforme les textes d'entrée en une séquence de sections situées entre deux grandes ancres alignées. Au second passage, on compte le nombre de petites ancres de chaque section afin d'éliminer celles pour lesquelles le nombre et l'ordre des petites ancres dans les deux textes d'entrée diffèrent. Cette opération rejette environ 10% des données de chaque texte d'entrée.

Méthode de Gale La méthode de Gale propose un alignement automatique de paragraphes balisés par des ancres, mais elle doit être suivie d'une vérification manuelle. Les auteurs mentionnent à la fin une possibilité d'amélioration par l'utilisation d'un algorithme plus élaboré pour l'alignement de paragraphes. L'amélioration proposée consiste en une distinction des ancres en deux types : dur (*hard* en anglais) et mou (*soft*). Les ancres dures doivent se trouver en nombre égal dans les deux textes d'entrée et elles ne peuvent pas être modifiées. En revanche, on peut déplacer les ancres molles si nécessaire tout en respectant la contrainte établie par les ancres dures. Ainsi, on pourra explorer la totalité des données sans être obligé d'en abandonner une partie si certaines ancres ne correspondent pas dans les deux textes d'entrée.

Alignement de phrases

Cette étape est dédiée à l'alignement des phrases contenues entre deux ancres. Brown réalise cette opération à l'aide d'un modèle de Markov caché et Gale par une méthode de programmation dynamique.

Méthode de Brown La méthode de Brown définit le texte T_l écrit dans la langue l comme une simple séquence de longueurs de phrases n_l (nombre de mots graphiques), balisée éventuellement par des marqueurs de paragraphe – retour chariot $¶_l$. Par exemple, un corpus parallèle composé d'une part d'un texte en français constitué de trois phrases contenant respectivement 19, 20 et 8 mots graphiques, se terminant par un retour chariot, et d'autre part d'un texte en anglais constitué de trois phrases contenant respectivement 17, 25 et 12 mots graphiques, se terminant par un retour chariot, est représenté comme suit :

$$T_f : 19_f 20_f 8_f ¶_f$$

$$T_a : 17_a 25_a 12_a ¶_a$$

On appelle « perle » l'ensemble des phrases et des marqueurs de paragraphe alignés.

Les auteurs posent comme hypothèse qu'une phrase dans une langue correspond à zéro, une ou deux phrases dans l'autre langue. Ainsi, huit types de perle sont possibles. Par exemple, dans le cas d'un alignement français-anglais, nous avons les huit possibilités de perle suivantes :

- perle-*fa* : une phrase française et une phrase anglaise ;
- perle-*f* : une phrase française et aucune phrase anglaise ;
- perle-*a* : aucune phrase française et une phrase anglaise ;
- perle-*ffa* : deux phrases françaises et une phrase anglaise ;
- perle-*faa* : une phrase française et deux phrases anglaises ;
- perle-$¶_f$: un marqueur de paragraphe français et aucun marqueur anglais ;
- perle-$¶_a$: aucun marqueur de paragraphe français et un marqueur anglais ;
- perle-$¶_f ¶_a$: un marqueur de paragraphe français et un marqueur anglais.

Un alignement est donc une séquence de certaines de ces huit perles représentant des phrases et des marqueurs de paragraphe. Si l'alignement correct des textes d'exemple est :

{ (1ère phrase française - 19_f, 1ère phrase anglaise - 17_a),
(2ème et 3ème phrases françaises - $20_f 8_f$, 2ème phrase anglaise - 25_a),
(3ème phrase anglaise - 12_a),
(marqueur de paragraphe - $¶_f$, marqueur de paragraphe - $¶_a$) }

il est représenté comme :

(perle-*fa*, perle-*ffa*, perle-*a*, perle-$¶_f ¶_a$).

La séquence de perles représentant l'alignement valide est générée par deux processus aléatoires : le premier est la génération des perles et le second le calcul de la probabilité de chacune des perles considérées en fonction des longueurs des phrases qu'elle contient.

Ces deux processus constituent un modèle de Markov caché.

Méthode de Gale La longueur des phrases est mesurée en terme de nombre de caractères, par contraste avec la méthode précédente qui la mesure par le nombre de mots graphiques. Les auteurs justifient leur choix par le fait qu'ils ont obtenu de moins bons résultats avec la version « mots » qu'avec la version « caractères ». Cette différence provient, d'après eux, du nombre plus élevé de caractères (dans leur étude, la longueur moyenne d'une phrase est de 117 caractères contre 17 mots seulement).

Pour l'hypothèse des combinaisons possibles de phrases à aligner, Gale ajoute aux cinq possibilités proposées par Brown, une autre combinaison constituée de deux phrases de chaque texte. Il s'agit de la situation où la première phrase du texte A et la première phrase du texte B ne sont pas des traductions mutuelles, ni les deuxièmes phrases des deux textes, mais où l'ensemble de la première et de la deuxième phrase du texte A constitue une traduction de l'ensemble de la première et de la deuxième phrase du texte B. Ainsi, Gale définit les six modèles de traduction suivants :

1. substitution (1-1) ;
2. suppression (1-0) ;
3. insertion (0-1) ;
4. contraction (2-1) ;
5. expansion (1-2) ;
6. fusion (2-2).

En s'appuyant sur cette hypothèse, les opérations se déroulent comme suit :

– Soit $D(i, j)$ le meilleur score entre les phrases $P_1, \ldots, P_i$ et leurs traductions $T_1, \ldots, T_i$, initialement $D(i, j) := 0$

– Pour chaque paire de paragraphes alignés,
considérer toutes les possibilités de couples constitués d'une phrase du texte de base, P_i ($1 \leq i \leq I$), et d'une phrase du texte en regard, T_j ($1 \leq j \leq J$) ;

1. calculer pour chaque couple (P_i, T_j) le coût de chacun des six modèles à l'aide de la fonction d.
   La fonction $d(x_1, y_1; x_2, y_2)$ est basée sur un modèle probabiliste qui produit à partir de leurs longueurs et de la probabilité du modèle de traduction qui les connecte, une approximation de la probabilité que les deux segments de chaque texte considéré soient des traductions mutuelles :
   a) $d(x_1, y_1; 0, 0)$ donne le coût de la substitution de x_1 avec y_1 ;
   b) $d(x_1, 0; 0, 0)$ donne le coût de la suppression de x_1 ;
   c) $d(0, y_1; 0, 0)$ donne le coût de l'insertion de y_1 ;
   d) $d(x_1, y_1; x_2, 0)$ donne le coût de la contraction de x_1 et x_2 en y_1 ;
   e) $d(x_1, y_1; 0, y_2)$ donne le coût de l'expansion de x_1 en y_1 et y_2 ;
   f) $d(x_1, y_1; x_2, y_2)$ donne le coût de la fusion de x_1 et x_2 correspondant à l'ensemble y_1 et y_2 ;
2. assigner à chaque couple (P_i, T_j) le meilleure score jusqu'au point (i, j) à l'aide de la fonction D.
   La fonction $D(i, j)$ calcule le minimum des six cas de modèle :

$$D(i,j) = \min \begin{cases} D(i-1,j-1) + d(x_1, y_1; 0, 0) \\ D(i-1,j) + d(x_1, 0; 0, 0) \\ D(i,j-1) + d(0, y_1; 0, 0) \\ D(i-2,j-1) + d(x_1, y_1; x_2, 0) \\ D(i-1,j-2) + d(x_1, y_1; 0, y_2) \\ D(i-2,j-2) + d(x_1, y_1; x_2, y_2) \end{cases}$$

3. déterminer la séquence de couples ayant le meilleur score, représentant un alignement valide.

2.2.2 Adaptation de l'algorithme à l'alignement avec les textes chinois

Wu (1994) présente dans la première partie de son article le résultat de l'adaptation des méthodes statistiques basées sur la corrélation des longueurs à l'alignement avec des textes chinois.

Il part de la remarque qu'en dépit des créateurs qui proclament le caractère indépendant de leur méthode vis-à-vis des langues considérées, la correspondance des longueurs semble plutôt relever de relations historiques entre les langues alignées. En effet, pour les langues d'une même famille, leur parenté explique la correspondance des longueurs aussi bien sur le plan lexical que syntaxique. Il a donc décidé de vérifier si la corrélation pouvait se généraliser à des langues non parentes comme l'anglais et le chinois.

Étant donné l'absence de séparateurs graphiques, Wu mesure la longueur des phrases en terme de nombre de caractères comme Gale. Il définit d'abord la notion de « nombre de caractères » pour les chaînes chinoises : la plupart des textes chinois contiennent non seulement des caractères chinois mais aussi des mots anglais – tels que noms propres et abréviations – en alphabet latin. Les caractères chinois sont comptés comme de longueur 2, et les caractères anglais et les signes de ponctuation comme de longueur 1. Il explique ensuite que cette règle correspond au nombre d'octets des textes stockés dans l'encodage hybride anglais-chinois connu sous le nom de *Big 5*. Mais l'encodage est un problème purement matériel, qui n'apporte aucune justification aux questions linguistiques de ce genre. Cette règle peut tout de même être justifiée linguistiquement, les caractères chinois formant un mot généralement avec seulement un ou deux caractères, tandis que les mots anglais sont constitués de beaucoup plus de caractères alphabétiques, ce qui entraînerait, si on les comptait de la même façon, une incohérence trop grande entre les longueurs des phrases avec des mots anglais et celles des phrases qui n'en contiennent aucun.

Bien que la relation des longueurs soit moins évidente pour les textes parallèles anglais-chinois que pour ceux des langues parentes, le résultat de la méthode purement statistique basée sur la corrélation des longueurs est très satisfaisant.

Le programme est néanmoins sensible aux paragraphes relativement longs contenant beaucoup de phrases notamment de longueur similaire, ce qui constitue la cause de la plupart des erreurs. Wu propose dans la seconde partie du même article une amélioration de la méthode qui résout ce problème via l'exploitation d'informations lexicales.

2.2.3 Caractéristiques de l'algorithme : avantages et inconvénients

Par opposition à la méthode de Kay que Brown et Gale critiquent tous les deux pour sa complexité de calcul, leurs algorithmes sont avant tout caractérisés par la simplicité donc la rapidité de calcul.

Néanmoins, comme ils le reconnaissent eux-mêmes, ils sont moins précis, n'alignant que partiellement les données. Comme nous l'avons vu, aussi bien la méthode de Brown que celle de Gale abandonnent une partie des données où une incohérence entre les deux textes d'entrée a été constatée. Ce défaut est néanmoins défendu par leurs créateurs qui considèrent que de nombreuses applications ne nécessitent qu'un alignement partiel.

Par ailleurs, comme l'indiquent les critiques de Simard et al. (1992) et Wu (1994), le résultat de cette méthode devient beaucoup plus mauvais dès que la situation de l'alignement se complique un peu. En effet, lorsque deux paragraphes parallèles considérés contiennent des nombres de phrases différents, on doit supposer que la traduction de certaines phrases a été omise ou qu'une/des nouvelle(s) phrase(s) ont été ajoutées par le traducteur, ou encore qu'il a réalisé une/des contraction(s) ou une/des expansion(s). Cependant, le programme échoue très facilement dans l'alignement de ce type de paragraphe, en mettant en

correspondance la totalité des phrases de cette partie sensible, car il est perturbé par quelques contractions ou expansions incorrectement réalisées.

Comme Simard et Wu, beaucoup de chercheurs attirés par cette méthode extrêmement simple, ont poursuivi leurs travaux en cherchant à améliorer cette faiblesse, souvent par l'introduction de l'exploitation d'informations lexicales. Nous allons examiner ces travaux dérivés dans la section suivante.

2.3 Méthodes avec amélioration par exploitation d'informations lexicales

Afin d'améliorer le plus grand défaut des méthodes économiques d'alignement basées sur la corrélation des longueurs, dont la sensibilité aux paragraphes relativement longs et contenant beaucoup de phrases est trop élevée, beaucoup de chercheurs exploitent des informations lexicales de différentes manières.

On trouve également des améliorations de la méthode de Kay, basée sur l'appariement lexical, par utilisation d'informations extérieures, en particulier des dictionnaires bilingues.

Nous nous intéressons dans cette section à ces améliorations. Nous allons d'abord étudier celles qui utilisent des éléments appelés « *cognates* » introduits pour la première fois par Simard et al. (1992). Nous passerons ensuite à la présentation de deux autres types d'améliorations : la méthode proposée par Wu (1994) qui utilise un ancrage supplémentaire à l'aide d'une liste de mots clés, et la méthode de Debili & Sammouda (1992) qui propose l'utilisation d'un dictionnaire bilingue pour le calcul de similarité des phrases dans un algorithme de type Kay. Enfin, l'étude se terminera par une discussion sur les avantages et les inconvénients de ces méthodes.

2.3.1 Amélioration introduisant la notion de « cognats »

Comme nous l'avons vu précédemment, tout en considérant la méthode basée sur la corrélation des longueurs comme une méthode simple et performante, Simard et al. (1992) font remarquer sa faiblesse qui apparaît dès que le problème devient un peu compliqué. Ils supposent alors que l'introduction de certaines connaissances linguistiques aiderait probablement la résolution de ce problème. Ils déduisent de leur intuition que la notion de « cognats » pourrait fournir une telle source de connaissances pour un coût minimal. Les cognats sont présentés comme des chaînes de caractères identiques, ou proches graphiquement, se trouvant dans les lexiques de langues ayant une relation historique plus ou moins étroite.

Leur amélioration consiste à calculer la « cognacité » (*cognateness* en anglais) des phrases en s'appuyant sur la conjecture : la relation de traduction entre deux phrases dans des langues différentes et leur cognacité sont corrélées, c'est-à-dire

qu'une paire de phrases qui sont des traductions mutuelles contient beaucoup plus de cognats qu'une paire aléatoire de phrases.

Nous allons maintenant étudier plus précisément ce qu'est un cognat et comment les détecter dans les corpus parallèles. Nous passerons ensuite à l'exposé des méthodes pour introduire la cognacité en vue d'améliorer l'alignement.

Cognats et transfuges

Le terme anglais *cognates* désigne d'après le glossaire (Bearth, 2003) les « mots apparentés ». On les appelle également « cognats » en français, terme qui constitue lui-même un exemple de cognat !

Cette notion est étudiée par exemple dans le cadre de la linguistique comparative ou la théorie de la traduction, en particulier sur les *false cognates* qu'on appelle en français mais aussi en anglais « faux amis ».

Dans l'article de Simard et al. (1992), on trouve la définition :

> « Informally speaking, cognates are pairs of tokens of different languages which share "obvious" phonological or orthographic and semantic properties, with the result that they are likely to be used as mutual translations. The pairs *generation/génération* and *error/erreur* constitute typical examples for English and French. »

Les auteurs ajoutent ensuite leur extension de cette définition en vue de l'alignement :

> « One might want to extend the notion so as to include such things as proper nouns (*Paris*; *London* and *Londres*), numerical expressions and even punctuation (question marks, parentheses, etc.). »

Ces extensions de cognats, invariants à la traduction, sont appelés **transfuges** par Langé & Gaussier (1995).

Détection des cognats dans Simard

Simard *et al.* décrivent un algorithme de reconnaissance des cognats comme suit :

Soit S_1 et S_2 un paire de phrases.

- création des listes T_1 et T_2 de mots t de chaque phrase ;
- comparaison des éléments des deux listes. Soient deux candidats t_1 et t_2 des listes de mots respectivement T_1 et T_2 ;
  - catégorisation des éléments des listes. t est un candidat pour une paire de cognat, s'il correspond à l'une des catégories suivantes :
    1. t est entièrement composé de lettres et de chiffres et contient au moins un chiffre ;
    2. t est exclusivement composé de lettres et contient au moins quatre lettres ;

3. t est un caractère de ponctuation simple.

– t_1 et t_2 sont cognats si et seulement si

1. les deux appartiennent à la catégorie 1 ou 3 et qu'ils sont complètement identiques ;
2. les deux appartiennent à la catégorie 2 et qu'ils ont leurs quatre premiers caractères identiques.

Ainsi, pour détecter une paire de cognats, Simard définit la sous-chaîne commune maximale comme une sous-chaîne initiale contenant au moins quatre lettres. Mais, il existe également d'autres méthodes de comparaison des chaînes.

Autres méthodes de détection des cognats

Borin (1998) qui a étudié l'efficacité de différents types de méthodes de comparaison de chaînes pour détecter les cognats, cite des méthodes comparant les sous-chaînes, outre initiales comme Simard, finales (Tiedemann, 1991) ou de position libre (Zhang & Kim, 1990).

Borin parle également des méthodes utilisant des connaissances linguistiques plus sophistiquées, en particulier celle de Covington (1996). La méthode consiste en un calcul des coûts d'alignement entre deux chaînes, qui représente la possibilité pour ces chaînes de former des cognats. Les coûts sont attribués selon des règles de nature phonologique. Le tableau 2.2 montre une partie de ces règles.

C (consonne) avec C identique	0
V (voyelle) avec V identique	5
V brève avec V longue, ou V avec S (semi-voyelle)	10
V avec V différente	30
C avec C différente	60

TAB. 2.2 – Règles d'attribution des coûts

D'après les études de Borin, cette méthode utilisant des connaissances linguistiques, ne produit pas, contrairement aux attentes de l'auteur, de résultats plus corrects que les méthodes par simple comparaison des caractères.

Parmi les méthodes simples, il en existe également qui calculent de manière plus complexe le coût pour définir une méthode plus précise de détection des cognats. Mettant en doute l'efficacité d'une simple comparaison de n-grammes, Kraif (1999, 2001) propose le calcul du rapport entre les longueurs des mots et celles des sous-chaînes maximales communes (SCM ci-après) à l'instar de la méthode de comparaison des chaînes de Debili et Sammouda (décrite dans la section 2.3.2), qui autorise les sauts. Par exemple, la longueur de la sous-chaîne des mots *docteur*/*dottore* (italien) selon les méthodes précédemment décrites est de 2, tandis qu'avec la méthode de SCM autorisant les ruptures, elle est de 2 + 1 + 1 = 4 (do + t + r).

Mais cette méthode ne tenant pas compte de la combinaison des sous-chaînes, elle risquait de générer beaucoup de bruit. Kraif a donc créé une version plus contraignante n'autorisant que les sous-chaînes quasiment parallèles, c'est-à-dire celles qui n'ont que des décrochements (insertion ou suppression) isolés entourés de caractères identiques et non pas des décrochements consécutifs. Ainsi, « pragmatique » qui est entièrement inclus dans « paradigmatique » n'est pas considéré comme une sous-chaîne, car « di » représente deux décrochements consécutifs[1].

Le rapport r entre les longueurs des mots et celles des SCM ainsi calculées, est obtenu comme suit :

soit M_1 et M_2 la paire de mots considérée,

$$r(M_1, M_2) = \frac{L(SCM)}{\max(L(M_1), L(M_2))}$$

où $L(m)$ est la longueur de la chaîne de caractères m.

Exploitation des cognats dans l'alignement

N'ayant pas obtenu de meilleurs résultats par une méthode d'alignement exploitant uniquement les cognats, Simard *et al.* introduisent la cognacité uniquement dans les situations où la méthode basée sur la corrélation des longueurs présente des problèmes pour aligner les phrases.

Leur méthode procède en deux passages : le premier passage est essentiellement identique à la méthode de Gale, excepté le fait que le résultat est constitué d'une liste des meilleurs alignements et non du meilleur alignement uniquement. Si le résultat du premier passage ne permet pas de choisir une solution unique, le programme réalise alors un second passage et utilise la fonction de calcul des scores basée sur les cognacités de phrases, qui remplace celle basée sur les longueurs de phrases. La cognacité γ de la paire de phrases P_1 et P_2 est définie comme :

$$\gamma = \frac{c}{\frac{(n+m)}{2}}$$

où n et m sont les nombres de mots des phrases P_1 et P_2 et c le nombre maximum de paires de cognats réalisables sans utiliser deux fois le même mot.

La notion de cognacité est exploitée également pour l'amélioration d'ancrage par Langlais (Langlais, 1997 ; Langlais & El-Bèze, 1997), Kraif (1999, 2001) ainsi que Simard & Plamondon (1998). Ces méthodes sont basées sur des algorithmes de dernière génération qui combinent de manière beaucoup plus stratégique les méthodes existantes qui utilisent des informations différentes telles que longueurs, cognats ou distribution lexicale. Nous étudierons ces méthodes dans la section 2.4.

[1] Le « a » est lui autorisé car étant une insertion isolée.

2.3.2 Méthodes proposées par Wu et par Debili et Sammouda

Méthode de Wu

L'application de la méthode de Gale à l'alignement de textes anglais-chinois montre la performance de la méthode même pour un couple de langues non parentes. Cependant, le programme échoue très facilement dans l'alignement des parties sensibles, comme par exemple les paragraphes contenant des nombres différents de phrases. Afin de résoudre ce problème, Wu utilise non seulement les longueurs mais aussi des critères lexicaux pour le calcul des probabilités d'alignement de deux phrases.

L'équation 2.1, proposée par Gale (cf. section 2.2.1), ne tient compte que des longueurs :

$$Pr(L_1 \rightleftharpoons L_2|L_1, L_2) \approx Pr(L_1 \rightleftharpoons L_2|l_1, l_2) \tag{2.1}$$

où $L_1 \rightleftharpoons L_2$ est une paire de phrases constituée d'une phrase de chaque texte d'entrée L_1 et L_2, et l_1 et l_2 sont leurs longueurs en terme de nombre de caractères.

Cette équation est remplacée par une nouvelle (2.2), prenant en compte les occurrences des mots d'une liste prédéfinie :

$$Pr(L_1 \rightleftharpoons L_2|L_1, L_2) \approx Pr(L_1 \rightleftharpoons L_2|l_1, l_2, v_1, w_1, \ldots, v_n, w_n) \tag{2.2}$$

où v_i et w_i sont les valeurs relatives à l'occurrence dans les phrases L_1 et L_2, respectivement du mot clé anglais$_i$ et du mot clé chinois$_i$ constituant une entrée de la liste prédéfinie.

Afin de ne pas trop nuire au caractère économique de la méthode, les mots clés de la liste doivent être restreints au strict minimum tout en possédant une efficacité suffisante pour réaliser correctement un alignement que la méthode basée sur la corrélation des longueurs seule ne permet pas d'obtenir. La liste a donc été créée en sélectionnant les mots qui vérifient les conditions suivantes :

1. les mots clés doivent être extrêmement fiables pour éviter de mauvaises associations sources de bruit supplémentaire ;
2. les mots clés doivent avoir une fréquence élevée pour réduire les calculs inutiles que provoqueraient des mots clés de fréquence nulle.

Méthode de Debili et Sammouda

Les grandes lignes de la méthode proposée dans Debili & Sammouda (1992) sont les suivantes.
Soient T_F et T_A les textes d'entrée,

– appariement des phrases :

1. comparaison de la phrase F_i avec toutes les phrases de la zone censée contenir la phrase A_n recherchée et inversement, comparaison de A_j avec toutes les phrases de la zone qui lui est associée[2] ;
   a) appariement des mots de F_i et A_j ;
      i. comparaison de chacun des mots de F_i avec tous les mots de A_j ;
         A. comparaison des mots f_s et a_t (précisée ci-dessous) ;
         B. inscription à la position (s,t) de la matrice Matmot de la note obtenue par comparaison ;
      ii. détermination de la meilleure note pour la ligne s dans Matmot ;
      iii. détermination de la meilleure note pour la colonne t dans Matmot ;
   b) calcul de la note reflétant la proximité des phrases comparées. La note est basée sur trois points : appariement des mots qui composent les phrases, longueurs des mots appariés et séquentialité de ces mots ;
2. si F_i et A_j sont mutuellement la meilleure traduction l'une de l'autre, alors leur appariement est retenu.

Les comparaisons de mots (étape 1(a)iA de l'algorithme) sont établies à l'aide d'un dictionnaire de transfert de mots simples. Pour calculer la note N des mots f_s et a_t, f_s est comparé à chacune des traductions $f_{k,t}$ de a_t obtenues par consultation du dictionnaire, et a_t est comparé à chacune des traductions $a_{k,s}$ de f_s.

Chaque comparaison de mots se traduit par une note. Elle est obtenue par la formule de comparaison de chaînes suivante :

$$N = \left[1 - \frac{L(c_1) - L(c_2)|}{L(c_1) + L(c_2)}\right] \cdot \sum_{i=1}^{n(t)} t_i^2 \tag{2.3}$$

où $L(c)$ est la taille en nombre de caractères de la chaîne c, et $n(t)$ le nombre de sous-chaînes maximales communes de longueur t.

Supposons que N_{f-a} soit la meilleure note obtenue dans le sens français-anglais et N_{a-f} dans le sens contraire. La note globale est alors obtenue en additionnant N_{f-a} et N_{a-f}.

Considérons un exemple.
Soient « ministère » et « *minister* » deux chaînes de caractères à comparer appartenant respectivement à la phrase française et à la phrase anglaise considérées. L'opération de comparaison se déroule comme suit :

[2] Ne sont considérés dans cette version de l'algorithme que des appariements 1-1.

1. Consultation du dictionnaire.
   L'entrée « ministère » contient comme traductions :
   – *agency*
   – *crown*
   – *department*
   – *ministry*
   – *office*
   et « *minister* » contient :
   – ministre
   – pasteur
   – secrétaire
2. Comparaison de la chaîne française « ministère » avec chacune des traductions de la chaîne anglaise « *minister* ». La meilleure note est obtenue avec le mot « ministre » :

$$N_{f-a} = \left[1 - \frac{|9-8|}{9+8}\right] \cdot \left(6^2 + 2^2\right) = 37{,}647$$

3. Comparaison de la chaîne anglaise « *minister* » avec chacune des traductions de la chaîne française « ministère ». La meilleure note est obtenue avec le mot « *ministry* » :

$$N_{a-f} = \left[1 - \frac{|8-8|}{8+8}\right] \cdot \left(6^2 + 1^2\right) = 37{,}000$$

4. Calcul de la note globale en additionnant les deux meilleures notes, soit :

$$N = N_{f-a} + N_{a-f} = 74{,}647$$

2.3.3 Avantages et inconvénients des méthodes

La notion de cognats améliore de manière simple et économique les méthodes statistiques qui n'utilisent aucune information lexicale.

Bien que l'article de Church et al. (1993) mentionne l'obtention de résultats intéressants lors de l'alignement des textes anglais-japonais, l'efficacité des méthodes basées sur les cognats est très limitée lors du traitement des langues non apparentées par rapport aux cas où il s'agit d'aligner des textes dans des langues appartenant à la même famille. La tentative de Wu semblait indiquer la direction à prendre pour donner un caractère plus universel à ces méthodes. Cependant, son souhait de constituer une liste optimale n'était pas compatible avec l'amélioration de la portabilité de l'algorithme. Ainsi, Wu n'a pas apporté d'amélioration plus universelle que celle réalisée par les cognats.

Quant à la proposition de Debili et Sammouda, leur méthode de vérification bi-directionnelle de traduction est très intéressante, ce qui n'empêche pas pour autant de se demander si une telle précision est vraiment nécessaire pour l'appariement grossier des mots dans le cadre d'un alignement de phrases.

L'amélioration des méthodes d'alignement statistiques s'est poursuivie ensuite dans la direction d'une recherche d'équilibre entre robustesse et résolution. Plusieurs chercheurs ont conclu que la solution résidait dans la combinaison beaucoup plus stratégique de méthodes existantes qui utilisent des informations différentes telles que longueurs, cognats ou distribution lexicale, formant ainsi une nouvelle classe d'algorithmes de dernière génération, algorithmes que nous allons aborder dans la section suivante.

2.4 Méthodes combinées

Nous nous intéressons dans cette section aux méthodes combinées. Elles sont le fruit de travaux récents qui combinent plusieurs techniques existantes, représentant elles-mêmes une sorte de panorama de l'état de l'art de l'alignement des phrases.

Nous allons commencer par la présentation de la méthode proposée par Langlais (Langlais, 1997 ; Langlais & El-Bèze, 1997) avant d'aborder celle proposée par Simard & Plamondon (1998). Après l'examen de la technique de Kraif (1999, 2001), l'étude se terminera par une discussion sur les avantages et les faiblesses de ces nouvelles méthodes.

2.4.1 La méthode proposée par Langlais

Langlais présente dans les articles (Langlais, 1997 ; Langlais & El-Bèze, 1997) le système JAPA développé au Laboratoire Informatique d'Avignon (LIA).

Caractéristiques du système

Le système est caractérisé par le fait que :
- il prend en entrée une paire de textes segmentés en phrases (segmentation non mise en doute) qu'il aligne au niveau des phrases ;
- un premier alignement au niveau des mots permet de délimiter un faisceau de recherche ;
- un algorithme de programmation dynamique recherche ensuite l'alignement optimal en considérant différents types de scores reflétant aussi bien des contraintes linguistiques et lexicales que des contraintes de surface *ad hoc* des appariements.

Première étape : réduction de l'espace de recherche

La réduction consiste en un alignement grossier au niveau des mots qui permet de produire un alignement au niveau des phrases. Ce dernier servira de base pour délimiter la zone de recherche dans l'étape suivante.

Le procédure d'alignement des mots se déroule comme suit :

1. création d'une matrice binaire M représentant le corpus bilingue à aligner. La $i^{\text{ème}}$ ligne de la matrice représente le $i^{\text{ème}}$ mot du texte d'entrée T_A et la $j^{\text{ème}}$ colonne, le $j^{\text{ème}}$ mot du texte T_B ;
2. affectation d'une valeur à chaque case. La case $M(i, j)$ prend la valeur 1 si le $i^{\text{ème}}$ mot du texte T_A et le $j^{\text{ème}}$ mot du texte T_B sont des mots de faible fréquence (en l'occurrence inférieure à 10) – afin d'éliminer le bruit que provoquent les mots grammaticaux – et en relation de traduction. Deux mots sont considérés comme étant en relation si :
   - ils forment un cognat. On considère deux mots comme cognats si :
     - ils contiennent chacun au moins un chiffre et qu'ils sont identiques ;
     - ils appartiennent à certains symboles de ponctuation qui sont utilisés quasiment de la même manière dans les différentes langues tels que « : » ou « ; » ;
     - constitués tous les deux exclusivement de lettres, ils partagent une même sous-chaîne préfixale de cinq lettres.
   - ils se trouvent être une des entrées du lexique de transfert.
3. calcul du meilleur score – pour l'alignement au niveau des mots – par une technique de programmation dynamique en privilégiant les chemins qui ne s'écartent pas trop de la diagonale ;
4. premier alignement de phrases à partir de l'alignement de mots ;
5. détermination d'un faisceau de recherche de largeur constante (à savoir égale à 8) centré autour du premier alignement de phrases.

Seconde étape : alignement des phrases

Le programme recourt ensuite à nouveau à un algorithme de programmation dynamique pour la recherche de l'alignement optimal en considérant des scores mettant à profit aussi bien des indices de surface que des indices linguistiques. Le score d'un appariement est le produit du score linguistique et du score de surface.

Informations de surface Le système utilise deux indices de surface : longueur de phrases et fréquence de chaque modèle de traduction. Les modèles de traduction considérés sont 1-1, 1-0 (ou 0-1), 1-2 (ou 2-1) et 2-2.

Le programme utilise pour calculer la probabilité d'appariement à l'aide de ces indices, le modèle proposé par Gale & Church (1993).

Informations linguistiques Pour exploiter les informations linguistiques, le système introduit des lexiques bilingues, la notion de cognats et enfin la notion empruntée des domaines de l'indexation et de la recherche d'information qu'est l'affinité lexicale (AL).

Le système recourt, comme nous l'avons déjà vu, à un lexique bilingue (bien que son utilisation ne soit pas obligatoire). Conscient de la disponibilité restreinte

de ce type de ressources, l'auteur justifie l'utilisation de lexiques bilingues par la possibilité de les obtenir par compilation automatique.

Les affinités lexicales désignent tout couple de mots (d'une même langue) partageant des relations à un niveau syntaxique et/ou sémantique. Elles sont extraites par analyse syntaxico-sémantique.

Le système utilise cette notion pour élaborer son lexique bilingue. Basé sur l'observation d'un chercheur (Martin et al. (1983) cités dans l'article) qui a montré que pour la langue anglaise 98% des relations lexicales mettaient en jeu des mots qui sont distants d'au plus 5 mots dans une même phrase, le système les détecte, en pratique, en examinant les co-occurrences dans une fenêtre d'une taille supposée suffisante, qui glisse sur chaque phrase du texte source.

Les AL extraites sont ensuite classées selon leur fréquence. Ce traitement s'appuient également sur des études antérieures (Maarek et al. (1991) cité dans l'article) qui ont montré qu'un mot était d'autant plus caractéristique d'un texte observé (T) qu'on le retrouvait fréquemment dans ce texte mais rarement dans un ensemble de textes (S) représentatif de la langue considérée. Les AL de score le plus élevé sont ensuite mises en correspondance par un test de vraisemblance. Les AL ainsi alignées sont enfin ajoutées dynamiquement au lexique bilingue utilisé pour l'alignement des phrases.

2.4.2 La méthode proposée par Simard et Plamondon

La méthode proposée par Simard & Plamondon (1998) consiste à combiner la robustesse des méthodes basées sur les informations des caractères – telles que « char_align » de Church (voir la section 1.5.1) – et la précision des méthodes basées sur des informations lexicales.

Cette idée est implantée comme une stratégie en deux étapes : la première réalise un mappage bi-textuel, travaillant sur la robustesse plutôt que sur la précision ; la seconde calcule l'alignement des phrases sur la zone de recherche construite à partir du mappage de l'étape précédente, utilisant cette fois une méthode qui favorise la précision plutôt que la robustesse ou l'efficacité.

Première étape : mappage bi-textuel

Cette étape est réalisée par un programme appelé Jacal (*Just Another Cognate Alignment program*), qui détecte, comme les programmes prédécesseurs tels que char_align, des séquences similaires de caractères afin de réaliser un mappage très fiable et indépendant des divisions logiques des textes telles que les sections, paragraphes ou phrases.

Plus concrètement, Jacal essaie de mettre en correspondance des éléments dits « cognats isolés » (*isolated cognates* en anglais).

Cognats isolés Soient A et B une paire de textes. Deux chaînes de caractères α et β forment une paire de cognats isolés, si elles sont toutes les deux à la fois des

cognats et des **chaînes isolées**. La notion de ressemblance est calculée comme la cognacité (voir la section 2.2), excepté par le fait qu'on compare deux séquences de la même langue.

Jacal considère deux chaînes comme cognats si leurs quatre premiers caractères sont identiques.

Une occurrence d'une chaîne est dite isolée s'il n'existe aucune chaîne de caractères ressemblante dans une certaine fenêtre autour de cette occurrence. Cette fenêtre d'isolation est mesurée en caractères, et est installée de manière à couvrir une fraction donnée du texte, à savoir 30%.

Les cognats isolés ainsi mis en correspondance sont généralement corrects, mais il y en a tout de même certains qui constituent de fausses correspondances. Pour éliminer les fausses paires, le programme supprime les points trop éloignés de la ligne supposée reliant les deux extrémités – début et fin – du corpus parallèle en utilisant une technique basée sur la régression linéaire.

Étape intermédiaire : segmentation et détermination de l'espace de recherche

Cette étape est dédiée à la construction, à partir du mappage obtenu, de l'espace de recherche pour l'alignement final.

Afin de déterminer l'espace de recherche, les auteurs considèrent l'alignement de phrases comme un cas particulier de mappage bi-textuel, celui dans lequel les points mappés doivent coïncider avec les limites de phrases.

En pratique, le système dessine un couloir le long des paires de points adjacents dans le mappage bi-textuel obtenu. La largeur du couloir est proportionnelle à la distance entre deux points connectés. Seules les limites de phrases se trouvant à l'intérieur du couloir sont alors considérées comme des points à traiter, constituant ainsi l'espace de recherche pour l'alignement final.

Seconde étape : alignement des phrases

L'alignement des phrases est implanté par le programme Salign, basé sur un des modèles statistiques de traduction lexicale proposés par Brown et al. (1993), appelé *Model 1*.

Ces modèles donnent une méthode pour calculer la probabilité conditionnelle $Pr(f|a)$, dite probabilité de la traduction (f, a), où f est une chaîne de caractères en français et a une chaîne de caractères en anglais. Cette probabilité $Pr(f|a)$ peut être interprétée comme la probabilité qu'un traducteur produirait, à partir d'un texte source a, la traduction f.

Brown et al. (1990) introduisent l'idée d'un lien entre une paire de chaînes, indiquant, pour chaque mot de la chaîne française, le mot dans la chaîne anglaise à partir duquel il a été traduit. Ces alignements entre les mots français et anglais sont appelés « connexions » (*connections* en anglais).

Avec *Model 1*, on choisit d'abord une longueur pour la chaîne française, en ne considérant que des longueurs raisonnables. Ensuite, pour chaque position

dans la chaîne française, on décide comment la connecter à la chaîne anglaise, et quel mot doit y être placé. Dans ce modèle, on suppose toutes les connexions pour chaque position française, mais l'ordre des mots dans a et f n'influe pas sur $Pr(f|a)$.

Un tel modèle peut être utilisé pour un alignement basé sur le calcul de scores. En effet, il peut réaliser l'estimation de la probabilité d'appartenance d'un ensemble arbitraire de mots dans une des langues, étant donné un autre ensemble dans l'autre langue. Ce qui est donc applicable à l'estimation de la similarité d'une phrase avec une autre.

2.4.3 La méthode proposée par Kraif

La méthode proposée par Kraif (1999, 2001) est également le fruit de recherches d'un équilibre entre robustesse et précision. C'est un algorithme basé sur les cognats et destiné à fournir d'abord un pré-alignement – une suite de points d'ancrage très sûrs – pour établir des îlots de confiance et réduire l'espace de recherche des algorithmes plus coûteux.

Afin d'obtenir les meilleurs résultats, l'auteur utilise, à l'instar des études récentes telles que celles décrites précédemment dans cette section, différents indices : longueurs, cognats, distributions lexicales. Suivant une heuristique très simple, le principe de précision d'abord, ces indices sont exploités par ordre de précision décroissante.

Le programme est constitué de trois étapes que nous étudions à présent.

Première étape : exploitation des transfuges

On exploite d'abord uniquement les chiffres et les symboles appelés transfuges (cf. § 2.3.1 ou § 3.4.6) qui sont des indices plus fiables que les cognats, produisant moins de bruit.

La mise en correspondance des transfuges est implantée par un processus itératif en deux temps comme suit :

Soient A et B deux sous-sections des textes d'entrée T_1 et T_2. Initialement $A = T_1$ et $B = T_2$.

1. détection des transfuges apparaissant le même nombre de fois dans les deux sections A et B. On apparie ces occurrences, notées par (i, j) où $i \in A$ et $j \in B$, pour obtenir un ensemble de points d'ancrage candidats ;
2. filtrage des points d'ancrage candidats selon les critères suivants, dont les trois premiers traduisent l'hypothèse de parallélisme, le dernier étant une condition supplémentaire pour maximiser la précision :
   - **diagonalité** : élimination des points situés à l'extérieur du couloir centré sur la diagonale de l'espace à aligner ;
   - **continuité** : suppression des points présentant une déviation forte par rapport aux points précédents ;

- **monotonie** : suppression des points entrant en conflit sur l'une de leurs coordonnées, ainsi que des points croisés (i, j) et (i', j') où $i > i'$ et $j < j'$;
- **surdétermination** : prise en compte uniquement des points générés par au moins deux transfuges différents.

Chaque point obtenu donne lieu à un découpage de la section alignée en sous-sections alignées. On réitère les étapes 1 et 2 sur chaque section de manière récursive, commençant par $A :=$ section s'étendant du début à (i_1, j_1), jusqu'à stabilité des îlots de confiance dégagés.

Deuxième étape : exploitation des cognats

On examine dans cette étape tous les couples de phrases alignables à l'intérieur des îlots de confiance obtenus, c'est-à-dire tous les points situés dans un couloir autour de la diagonale de chaque section.

La procédure se déroule comme suit :

Considérons les sous-sections A et B alignées par la première étape.

1. comparaison de tous les couples de phrases, p^A et p^B, situés dans un couloir de largeur constante (à savoir 10 phrases) ;
   - comptage de la fréquence f des cognats ;
   - inscription de la fréquence f des cognats dans la case (i, j) de la matrice des fréquences F ;
   - calcul à partir de la matrice des fréquences d'une nouvelle matrice C exprimant le lien statistique entre les lignes i et les colonnes j :

$$c_{ij} = \frac{(f_{ij} - f_i f_j)^2}{f_i f_j}$$

   - application d'une contrainte de réciprocité en retenant tous les points (i, j) tels que p_i^A atteint son maximum avec p_j^B, et p_j^B atteint son maximum avec p_i^A ;
2. filtrage de l'ensemble des points obtenus par les critères de continuité et de monotonie.

Troisième étape : alignement final

Un algorithme de programmation dynamique est appliqué pour l'appariement des phrases entre les points d'ancrage afin de produire un alignement complet.

La mesure de distance est basée sur la densité des cognats et la probabilité *a priori* des transitions (méthode proposée par Gale & Church (1993)).

2.4.4 Avantages et faiblesses

L'avantage le plus marqué de ces algorithmes est l'amélioration de la robustesse des systèmes. Comme le dit Kraif (2001) dans la conclusion :

> « [...] la méthode de préalignement est adaptée au développement d'heuristiques pour la détection d'omissions ou d'interversions de sections importantes, dans la mesure où la forte densité des points d'ancrage permet de faire apparaître clairement les ruptures dans le parcours du chemin. »

La nécessité de l'étape de pré-alignement nous apparaît donc maintenant clairement.

Il reste tout de même le problème lié à la notion de cognat qui limite son application à un ensemble de langues restreint.

Quant à la méthode de Simard *et al.*, l'existence de ressources importantes est présupposée. En effet, différents paramètres du modèle utilisé sont normalement estimés à partir de fréquences observées dans une grande collection de paires de segments (typiquement, de phrases) qui sont mutuellement traductions l'un de l'autre. Beaucoup de méthodes utilisant les informations lexicales ne présupposent pas de la disponibilité de ce genre d'information. Mais les auteurs défendent leur méthode par l'existence de grands nombres de textes parallèles déjà alignés pour les paires de langues telles que l'anglais et le français. Ils comparent le non-recours aux informations extérieures pour ces langues au fait de « *re-inventing the wheel every time* » (réinventer la roue à chaque fois).

Toutefois, comme ils le disent eux-mêmes, cette situation enviable concerne uniquement quelques langues : beaucoup de langues ne possèdent pas encore de corpus parallèles, même non alignés.

2.5 Méthodes d'alignement par la technique de recherche d'information

Fluhr *et al.* proposent un autre type d'algorithme (Fluhr et al., 2000 ; Semmar & Fluhr, 2007), qui, complètement différent des deux modèles classiques, supporte mieux la contrainte des hypothèses de bijectivité et de monotonie (voir la section 1.2.2).

Cette approche consiste à réaliser l'alignement par une méthode de recherche d'information multilingue (ou recherche documentaire multilingue), en particulier celle basée sur la « reformulation », dite enrichissement des requêtes.

Le choix de cette approche provient de l'observation comparative des problèmes de l'alignement des phrases d'une part, et de la recherche documentaire multilingue basée sur l'enrichissement des requêtes d'autre part. Ces opérations nécessitent toutes les deux un calcul de similarité entre deux textes dans différentes langues.

Avant d'entrer dans l'étude de cette méthode originale, nous allons tout d'abord nous intéresser au principe de recherche documentaire multilingue basée sur l'enrichissement des requêtes, pour terminer la présentation par une discussion sur les avantages et les faiblesses de la méthode.

2.5.1 Recherche d'information multilingue basée sur l'enrichissement des requêtes

La recherche d'information interlangue (*Cross-Language Information Retrieval* en anglais, CLIR) consiste à récupérer, à partir d'une requête formulée dans une langue donnée (généralement dans la langue maternelle de l'utilisateur, français par exemple), des documents écrits dans d'autres langues différentes de celle de la requête (anglais par exemple).

Enrichissement des requêtes

La recherche documentaire basée sur l'enrichissement des requêtes est une méthode qui réalise la recherche d'information par enrichissement des requêtes (*query expansion* en anglais), c'est-à-dire le remplacement de chacun des mots de la requête par d'autres mots exprimant le même concept. L'enrichissement des requêtes, s'applique, s'il s'agit de recherche documentaire monolingue, à la même langue en remplaçant un mot de la requête par des synonymes et/ou des hyponymes, etc., et il produit comme résultat, lorsqu'il est appliqué dans le cadre de recherches multilingues, toutes les traductions possibles dans une autre langue à l'aide par exemple d'un dictionnaire bilingue.

Principe de fonctionnement

Les auteurs présentent l'architecture du système SPIRIT (*Syntactic and Probabilistic Indexing and Retrieval of Information in Texts*) du projet EMIR (*European Multilingual Information Retrieval*).

Le système constitue une base de données à partir de documents auxquels il applique des analyses linguistique et statistique. Lorsqu'il reçoit une requête de l'utilisateur, il réalise une reformulation puis une comparaison des résultats de la reformulation avec les documents présents dans la base de données.

La procédure générale de recherche par enrichissement des requêtes utilisant un dictionnaire bilingue se déroule comme suit :

1. déduction de toutes les traductions possibles à partir des mots de la requête originale ;
2. élimination des traductions qui ne figurent pas dans la base de données ;
3. recherche de documents pertinents avec le module de comparaison. Plus l'intersection entre les concepts exprimés par le document et ceux exprimés par la requête est vaste, plus le document est considéré comme pertinent.

Le module de comparaison est capable d'évaluer rapidement toutes les intersections possibles entre les mots de la requête et les documents, et de calculer pour chaque document un poids représentant le degré de pertinence. Pour la recherche d'information, le poids dépend uniquement de l'intersection entre requête et document.

2.5.2 Alignement des phrases basé sur la méthode CLIR

L'alignement est constitué de deux étapes. La première consiste en l'obtention d'un alignement 1-1 de haute précision. Elle est réalisée sans tenir compte de l'ordre des phrases, ce qui permet d'aligner efficacement des phrases même lorsque leur ordre dans un texte n'est pas préservé dans l'autre texte. La seconde étape réalise l'extension de l'alignement 1-1 obtenu à celui incluant les correspondances 1-2 et 2-1 par fusion de la phrase précédente (ou suivante) non alignée avec celle qui la suit (ou la précède) et qui est déjà alignée.

Première étape : alignement 1-1

Un corpus est composé de deux ensembles de phrases ordonnées. Le système d'alignement réalise l'indexation de ces deux textes dans deux bases de données différentes à l'aide du système SPIRIT. La détection des liens entre les phrases dans la langue de base et celles dans la langue en regard est réalisée par recherche d'information multilingue.

Cette méthode n'est pas symétrique car l'une des langues est considérée comme langue de base et est utilisée comme langue de départ constituant les requêtes. Les auteurs posent comme hypothèse que le résultat dépendant fortement de la qualité du dictionnaire utilisé, il est fort probable que le choix de la langue de base influe également sur les résultats. D'après eux, le meilleur choix est sans doute la langue dont le dictionnaire bilingue possède la meilleure couverture.

Seconde étape : alignement 1-2 et 2-1

La seconde étape consiste, afin d'améliorer l'alignement 1-1 obtenu, à essayer de fusionner une phrase non alignée avec une phrase déjà alignée qui la précède ou qui la suit.

La procédure se déroule comme suit :

1. vérification pour chaque phrase non alignée P_i, si la phrase précédente ou suivante est déjà alignée ;
2. recherche d'information par croisement de langues avec comme requête la concaténation P_iP_{i+1} ou $P_{i-1}P_i$;
3. si on obtient le même résultat R_j que pour la recherche avec uniquement P_{i+1} ou P_{i-1} et que l'intersection entre P_iP_{i+1} (respectivement $P_{i-1}P_i$) et R_j est supérieure à celle entre P_{i+1} (resp. P_{i-1}) et R_j, alors
   R_j est aligné avec P_iP_{i+1} (resp. $P_{i-1}P_i$) et R_j.

2.5.3 Avantages et faiblesses

Le principal avantage de cette méthode est, comme il a été déjà présenté, l'absence d'hypothèse de parallélisme, qui permet de supporter l'absence de traduction de certaines parties ou l'insertion de nouveaux passages.

Malgré la robustesse, la méthode a également hérité d'une faiblesse de la méthode de recherche d'information par enrichissement des requêtes : la qualité du dictionnaire utilisé influe directement sur le résultat de l'alignement. Puisque l'alignement est un outil de compilation de dictionnaires, une forte dépendance aux dictionnaires n'est pas une caractéristique favorable.

De plus, transformer un texte – un ensemble « ordonné » de phrases – en une base de données – un ensemble « non-ordonné » – provoque certainement une perte. Comme Brown et Gale l'ont remarqué très tôt, les marqueurs typographiques tels que les retours chariots ou certains signes de ponctuation sont des éléments très intéressants pour l'alignement des phrases. Il est vraiment dommage de les supprimer et de ne pas profiter de ces éléments si porteurs d'information.

Par ailleurs, cette méthode robuste qui supporte bien le bruit et qui se montre particulièrement efficace pour l'alignement des corpus dits « *noisy-parallel corpora* », est également utilisée pour l'alignement des corpus parallèles de journaux au niveau des articles. L'étude de Collier et al. (1998) présente le résultat de la comparaison de deux méthodes d'alignement d'articles, l'une basée sur une méthode CLIR avec traduction automatique et l'autre également sur une méthode CLIR mais avec simple consultation de dictionnaires. D'après les auteurs, la méthode avec consultation de dictionnaires est plus efficace que celle avec traduction automatique dans le cadre de la comparaison des résultats à des niveaux de fort rappel, c'est-à-dire lorsque l'on souhaite obtenir des quantités importantes d'articles alignés. La méthode d'alignement des phrases des textes anglais-japonais proposée par Uchiyama et Isahara que nous présenterons dans la section 2.6.2 utilise une méthode CLIR avec consultation de dictionnaires pour aligner d'abord les articles de journaux, afin de réaliser ensuite leur alignement au niveau phrastique.

2.6 Méthodes adaptées pour l'alignement avec des textes japonais

Afin de terminer l'état de l'art des techniques d'alignement des phrases, nous nous intéressons à présent aux méthodes adaptées à l'alignement du japonais.

Nous aborderons tout d'abord la méthode proposée par Murao (1991), utilisée par Utsuro et al. (1994) ainsi que par Collier & Takahashi (1995).

Nous présenterons ensuite celle développée par Uchiyama & Isahara (2003) qui diffère de cette dernière par l'absence d'utilisation d'informations statistiques.

Puis nous nous intéresserons à la méthode utilisée pour le système BACCS (Isahara & Haruno, 2000 ; Haruno & Yamazaki, 1996).

Enfin, nous terminerons l'exposé par l'étude de Hwang qui a proposé une méthode originale pour l'alignement entre le japonais et le coréen (Hwang & Nagao, 1994).

2.6.1 La méthode proposée par Murao

Nous étudions maintenant la méthode de Murao (1991) basée sur les correspondances lexicales utilisant un dictionnaire bilingue anglais-japonais. Cette méthode est également utilisée par d'autres chercheurs. Le système d'appariement proposé par Utsuro et al. (1994) l'utilise pour l'étape d'appariement au niveau des phrases. Collier & Takahashi (1995) ont également utilisé un système basé sur la méthode de Murao à l'occasion de la compilation d'un corpus bilingue au *Centre for Computational Linguistics* (CCL, Manchester), constitué d'articles du *Asahi*, un des grands quotidiens japonais.

Algorithme général

Avant l'alignement, les mots sont extraits de chaque phrase éventuellement après une analyse morphologique. Les correspondances des mots extraits sont détectées à l'aide de dictionnaires bilingues ainsi que d'informations statistiques.

En utilisant ces informations de correspondance lexicale, on calcule le score de chaque perle. Pour la constitution d'une perle, cinq possibilités de combinaisons (1-1, 1-2, 1-3, 1-4 et 2-2) sont considérées. Le score h d'une perle p est calculé comme suit :

Considérons la perle p constituée de x phrases $s_{a-x+1}, \ldots, s_a$ dans le texte S et de y phrases $t_{b-y+1}, \ldots, t_b$ dans le texte T.
Soient $n_s(a,x)$ et $n_t(b,y)$ les nombres de mots contenus dans les phrases constituant une perle des textes S et T respectivement, et $n_{st}(p)$ le nombre de paires de mots correspondants dans la perle p. Alors, le score h de p est défini comme le ratio de $n_{st}(p)$ sur la somme de $n_s(a,x)$ et $n_t(b,y)$:

$$h(p) = \frac{n_{st}(p)}{n_s(a,x) + n_t(b,y)}$$

L'alignement est obtenu par calcul de la séquence des perles constituant le meilleur score. Cette opération est réalisée comme un problème classique de programmation dynamique.

2.6.2 La méthode proposée par Uchiyama et Isahara

Uchiyama & Isahara (2003) présentent une méthode de constitution de corpus parallèles à partir de deux corpus non entièrement parallèles : un corpus japonais constitué d'articles publiés entre septembre 1989 et décembre 2001 du journal *Yomiuri* d'une part, et un corpus anglais constitué d'articles du journal *Daily Yomiuri* de la même période d'autre part.

Les articles du journal *Daily Yomiuri* sont des traductions de certains articles du journal *Yomiuri*, représentant moins de 6% du nombre total d'articles du corpus japonais. Si bien que leurs travaux commencent par l'extraction des articles parallèles. Elle est réalisée avec une méthode de recherche d'informations par croisement de langues (CLIR). Les articles en japonais sont d'abord transformés

en un ensemble de mots anglais par consultation d'un dictionnaire. La recherche est ensuite réalisée avec un article anglais comme requête, de la même manière que la recherche d'informations classique.

Par ailleurs, du fait du nombre important de paires de phrases mal alignées à cause de la présence de bruit dans le corpus (omission de traduction, différences dues aux adaptations réalisées par le traducteur, etc.), ils proposent également une méthode d'évaluation de la fiabilité d'alignement afin de sélectionner les paires de phrases qui ont une forte chance d'être correctement alignées, et qui sont réellement utiles à une utilisation postérieure. La mesure tient compte non seulement de la similarité des phrases mais aussi de la probabilité de correspondance des articles.

Méthode d'alignement par programmation dynamique

L'alignement est réalisé par une méthode de programmation dynamique tout comme la méthode proposée par Murao décrite précédemment.

Mais, contrairement à cette dernière, la méthode d'Uchiyama et Isahara n'utilise pas d'informations sur la probabilité de correspondance des mots, obtenues par calcul statistique. Les auteurs considèrent que les résultats d'alignement utilisant uniquement des informations qu'ils peuvent obtenir avec un dictionnaire sont suffisamment satisfaisants.

La similarité est donc calculée à l'aide uniquement d'un dictionnaire bilingue constitué spécifiquement – à partir des dictionnaires EDR japonais-anglais et anglais-japonais –, sans avoir recours à des informations statistiques. L'ensemble des mots pleins est d'abord constitué pour chaque phrase, après analyse morphologique par ChaSen pour les textes japonais, et après *tagging* par *Brill's Tagger* puis lemmatisation à l'aide de librairies du WordNet pour les textes anglais.

Pour calculer la similarité des paires de phrases, un alignement des mots est d'abord réalisé comme suit :

1. la liste de toutes les paires (j,e) appartenant au dictionnaire est constituée ;
2. l'ambiguïté de chaque paire est calculée : c'est le nombre total de mots qui se trouvent dans le dictionnaire en tant que traduction du mot japonais considéré, donc l'ambiguïté d'une paire de mots (j,e) est le cardinal de l'ensemble $M = \{m|(j,m) \in \mathsf{Dictionnaire}\}$;
3. la liste des paires de mots est triée dans l'ordre croissant d'ambiguïté ;
4. les paires sont examinées une par une afin d'obtenir une liste des mots alignés L : pour une paire considérée (j_m, e_n), s'il existe déjà le mot japonais j_{m-i} tel que $(j_{m-i}, e_n) \in L$ ou le mot anglais e_{n-k} tel que $(j_m, e_{n-k}) \in L$, la paire (j_m, e_n) est rejetée, sinon la paire (j_m, e_n) est inscrite dans la liste L.

La similarité des phrases est calculée avec la formule suivante :
Soient J et E les ensembles de mots pleins contenus dans les phrases à comparer :

$$\mathsf{sim}(J,E) = \frac{\mathsf{co}(J \cap E) + 1}{|J| + |E| - 2\mathsf{co}(J \cap E) + 2}$$

où

- $|X|$, fréquence totale de l'ensemble des mots appartenant à X calculée par $\sum_{x \in X} f(x)$ où $f(x)$ est la fréquence de x dans X ;
- $J \cap E$ est l'ensemble des paires de mots (j, e) considérés comme traduction l'un de l'autre tels que $j \in J$ et $e \in E$;
- $\mathsf{co}(J \cap E)$, fréquence totale de cooccurrence des mots alignés calculée par $\sum_{(j,e) \in J \cap E} \mathsf{min}(f(j), f(e))$.

Avec les similarités ainsi calculées, l'alignement des phrases est réalisé par une méthode de programmation dynamique avec comme modèles de traduction possibles, les paires 1-n et n-1, où $1 \leq n \leq 6$.

2.6.3 La méthode du système BACCS

Haruno, Yamazaki et Ishihara présentent la méthode d'alignement utilisée pour l'environnement graphique d'alignement BACCS (*Bilingual Aligned Corpus Construction System*) (Isahara & Haruno, 2000 ; Haruno & Yamazaki, 1996). Cette méthode, basée sur celle de Kay, est également caractérisée par la combinaison de l'utilisation de dictionnaires bilingues et d'une méthode statistique.

La principale différence par rapport à la méthode de Murao se trouve dans le choix, non d'une méthode de programmation dynamique, mais d'une approche itérative – tout comme la méthode de Kay – pour calculer l'alignement des phrases à partir des informations de correspondance lexicale. Les auteurs justifient ce choix, que beaucoup de chercheurs ont abandonné à cause de la lourdeur de calcul nécessaire, par la précision du résultat qu'il peut offrir.

L'autre caractéristique réside dans la façon de calculer la similarité. Pour ce faire, le système utilise l'information mutuelle et le t-score. L'information mutuelle représente la similarité des distributions d'occurrences de mots. Le t-score représente la fiabilité de l'information mutuelle obtenue (voir la section 2.1.7).

Algorithme général

Le système reçoit comme données des textes parallèles japonais-anglais.

Il utilise deux principales structures de données : la matrice des phrases alignables appelée ASM (*Alignable Sentence Matrix*) et la matrice d'ancres dite AM (*Anchor Matrix*).

La matrice ASM représente l'ensemble des phrases susceptibles d'être alignées entre des ancres, et correspond donc à la table « *Alignable Sentence Table* » (AST) de l'algorithme de Kay (cf. section 2.1.3). La matrice AM représente, comme la table « *Sentence Alignment Table* » (SAT) de l'algorithme de Kay (cf. section 2.1.5), l'ensemble des ancres.

1. **Étape morphologique** :
   Le programme réalise tout d'abord une analyse morphologique des deux textes pour extraire uniquement les substantifs, les verbes, les qualificatifs, les adverbes et les mots inconnus. Cette opération éliminant les mots gram-

maticaux permet, d'après eux, d'empêcher une baisse de performance d'alignement due à la grande différence des structures représentées notamment par ces mots grammaticaux.

2. **Construction de la matrice ASM** :
   Initialement, les ancres sont constituées des deux extrémités (début et fin) des textes et éventuellement des limites des articles ou des chapitres.
3. **Mise en correspondance des mots** :
   Pour toutes les paires possibles de mots appartenant aux phrases de la matrice ASM, on calcule l'information mutuelle et le t-score. Les paires de mots ayant un score supérieur à un certain seuil sont considérées comme paires de mots correspondant statistiquement.
4. **Construction de la matrice AM** :
   Pour toutes les paires possibles de phrases de la matrice ASM, on calcule le nombre de paires de mots correspondants figurant dans les dictionnaires et de mots correspondants statistiquement. Soit *ANC* le nombre minimal de mots correspondants pour qu'une paire de phrases puisse être considérée comme appariée. Les paires de phrases contenant plus de mots correspondants que *ANC* sont considérées comme de nouvelles ancres.
5. **Mise à jour de la matrice ASM** :
   En utilisant les nouvelles ancres obtenues, on calcule à nouveau l'ensemble des phrases dont l'alignement est envisageable.

On répète les opérations 3, 4 et 5 en diminuant la valeur des paramètres – les seuils de score et *ANC* –, ce qui permet, selon les auteurs, d'obtenir des appariements selon l'ordre de précision d'abord et de fournir ainsi un meilleur résultat que celui obtenu par une méthode de programmation dynamique.

2.6.4 Méthode d'alignement japonais-coréen

Hwang propose une méthode originale pour le coréen (Hwang & Nagao, 1994), consistant à traduire les phrases du texte de base afin de trouver leur phrase correspondante dans le texte en regard par ressemblance avec cette traduction.

Il part d'abord de deux critiques sur les méthodes classiques. Premièrement, lorsqu'un être humain cherche à réaliser l'alignement manuel d'un texte parallèle, il traduit les phrases pour trouver leurs correspondants, au lieu de compter le nombre de caractères ou de mots. Deuxièmement, les méthodes basées sur le nombre de mots nécessitent une analyse morphologique, ce qui pose des problèmes dans le cas de l'alignement du coréen car il n'existe pour le moment aucun analyseur morphologique coréen capable de fournir un résultat satisfaisant.

Il a donc posé comme hypothèse que si l'on arrivait à traduire les phrases du texte de base en séquences ressemblant aux phrases du texte en regard, on pourrait aligner facilement les phrases automatiquement. De plus, la ressemblance non seulement structurelle mais aussi lexicale des langues japonaise et

coréenne, qui permet avec une traduction basique d'obtenir des phrases relativement proches de celles présentes dans le texte original, est un argument très encourageant.

Cependant, étant donné qu'il est impossible d'obtenir par traduction des phrases strictement identiques à celles du texte original, il introduit, pour la mise en correspondance des phrases traduites avec les phrases originales, le degré de similarité des séquences de caractères et la valeur statistique d'appariement des phrases japonaises et coréennes.

Dorénavant, pour faciliter la compréhension, tout au long de cette étude, nous utiliserons exclusivement le terme **phrases originales** pour les phrases présentes dans l'un des textes parallèles d'entrée et **phrases intermédiaires** pour les phrases obtenues par traduction des phrases originales au cours du traitement, phrases qui seront ensuite comparées avec les phrases originales du texte en regard afin de trouver les correspondances avec les phrases originales du texte de base.

La méthode est constituée de deux grande étapes : transformation des phrases japonaises en phrases coréennes et appariement des phrases par calcul de similarité.

Transformation des phrases japonaises en phrases coréennes

Dans cette étape, on réalise une analyse morphologique du texte japonais et on cherche la traduction de chaque mot extrait, à l'aide d'un dictionnaire bilingue japonais/coréen (voir Hwang et al. (1993)) pour transformer les phrases japonaises en phrases coréennes.

En cas d'échec de traduction d'une phrase, la transformation est réalisée à l'aide d'un tableau de transfert des caractères japonais/coréens et d'un tableau de transfert des caractères adjacents.

Transformation avec dictionnaire japonais-coréen L'auteur affirme que l'ordre des mots dans les phrases japonaises et coréennes étant très proches, la traduction mot à mot peut souvent produire une phrase ressemblant à celle du texte original et le simple remplacement mot à mot à l'aide d'un dictionnaire peut fournir des résultats satisfaisants. L'auteur définit tout de même quelques règles élémentaires permettant de traiter correctement les exceptions dues à la variation de forme des verbes ou liées à la dérivation, ou encore aux allomorphes des particules coréennes.

Transformation à l'aide du tableau de transfert des caractères adjacents japonais/coréens Cependant, les mots ne figurant pas dans le dictionnaire ne peuvent pas être traités de cette manière. Or, les mots japonais constitués uniquement d'idéogrammes, *kanji*, peuvent souvent être traduits en mots coréens par simple remplacement caractère par caractère. Toutefois, certains caractères se traduisent différemment selon le caractère qu'ils précèdent ou selon leur po-

sition dans le mot. Ainsi, pour le remplacement caractère par caractère, un autre type de tableau appelé *NH-Table* (*Nihongo to Hangul conversion Table*) est aussi utilisé.

NH-Table est créée à partir du dictionnaire bilingue en ne considérant que les entrées constituées exclusivement de *kanji*, dont le nombre de caractères est égal au nombre de caractères de leurs correspondants coréens. Ce tableau est une matrice dans laquelle la valeur NH(x, y) de la case (x, y) est le caractère correspondant en coréen au caractère japonais x lorsqu'il est adjacent au caractère y. Il existe deux types de *NH-Table* : le tableau de transfert des caractères adjacents droits et le tableau de transfert des caractères adjacents gauches. Dans le premier type de tableau, y est le caractère qui suit le caractère x, et dans le second y représente le caractère qui précède le caractère x dans le mot.

Transformation à l'aide du tableau de transfert des caractères japonais/coréens Les mots constitués en *kanji* que l'on n'a pas réussi à transformer avec les *NH-Table*, peuvent être remplacés caractère par caractère en mots coréens. On réalise donc, pour les mots entièrement en *kanji* qui ne sont traduits ni avec le dictionnaire ni avec les tableaux de transfert de caractères adjacents, une transformation à l'aide du tableau de transfert des caractères. Il contient environ quatre mille caractères japonais avec leur caractère correspondant en coréen. Les mots constitués uniquement avec le syllabaire *katakana* – utilisé pour les mots emprunts – sont également traités avec cette méthode.

Appariement des phrases par calcul de similarité

Calcul de la similarité des phrases intermédiaire et originale Pour la mise en correspondance des phrases originale et intermédiaire dans la même langue, est utilisée la similarité de phrases (*SP*) obtenue par le rapport des longueurs de phrases (*RL*) et par la similarité des séquences de caractères (*SC*).

> Soient J_i les phrases japonaises (où $i = 1,\ldots,n$), I_s les phrases intermédiaires (traductions de J_i, $s = 1,\ldots,m$), C_k les phrases coréennes ($k = 1,\ldots,u$) et RM le rapport moyen des longueurs de phrases japonaise et coréenne soit 9 : 10 (= 0,9).

$$
\begin{aligned}
SP(J_i, C_k) &= SC(I_s, C_k) \cdot RL(J_i, C_k) \\
SC(I_s, C_k) &= \text{similarité des séquences de caractères (précisée ci-dessous)} \\
RL(J_i, C_k) &= \frac{J_i}{(C_k \cdot RM)} \quad \text{si } (J_i < C_k \cdot RM) \\
&\quad\ \frac{(C_k \cdot RM)}{J_i} \quad \text{sinon}
\end{aligned}
$$

La similarité des séquences de caractères est calculée en tenant compte de l'ordre des caractères comme suit :

Soient n le nombre de caractères de I_s, m le nombre de caractères de C_k et $W = 4$ la valeur maximum du bonus attribué aux caractères se succédant à l'identique dans les deux séquences.

– Calcul de la similarité SC des chaînes de caractères I_s et C_k :

$$SC(I_s, C_k) = \mathsf{score}(n, m)$$

– Calcul du score $\mathsf{score}(i, j)$:

$$\mathsf{score}(i,j) = \begin{cases} 0 & \text{si } (i{=}0) \vee (j{=}0) \\ \max \begin{pmatrix} \mathsf{score}(i-1,j-1) + \min(sc(i,j), W), \\ \mathsf{score}(i-1,j), \\ \mathsf{score}(i,j-1) \end{pmatrix} & \text{si } (1 \leq i \leq n) \wedge (1 \leq j \leq m) \end{cases}$$

– Calcul de la similarité sc des caractères a_i et b_j :

$$sc(i,j) = \begin{cases} 0 & \text{si } (i = 0) \vee (j = 0) \\ sc(i-1, j-1) + \mathsf{comp}(i,j) & \text{si } (1 \leq i \leq n) \wedge (1 \leq j \leq m) \end{cases}$$

– Comparaison comp des caractères a_i et b_j :

$$\mathsf{comp}(i,j) = \begin{cases} 0 & \text{si } a_i \neq b_j \\ 1 & \text{si } a_i = b_j \end{cases}$$

Appariement des phrases Cinq modèles d'appariement sont définis : type 0 (1-1), type 1 (1-2), type 2 (1-3), type 3 (2-1) et type 4 (3-1). Pour chaque phrase originale de base, les similarités avec les phrases intermédiaires selon les cinq modèles sont calculées.

Lorsque la similarité des phrases du type 0 est la plus élevée, la détermination de l'appariement pour une phrase de base donnée prend également en compte la possibilité d'appariement de la phrase suivante, afin de pénaliser le type 0 ayant tendance à produire une similarité des séquences relativement élevée.

L'alignement est réalisé à partir des premières phrases. Une fois l'appariement des phrases considérées déterminé, l'appariement de la phrase suivante commence.

2.6.5 Avantages et faiblesses

N'ayant aucun élément de surface indiquant la correspondance de deux phrases tel que les cognats pour les langues européennes, les chercheurs japonais recourent à des informations extérieures, les dictionnaires. Conscients du problème de l'absence de certains mots dans le dictionnaire, ils exploitent également les informations lexicales obtenues par méthode statistique. Cette approche permet de réaliser des aligneurs adaptés à l'alignement des langues ayant des structures et des systèmes d'écriture très différents telles que le japonais et l'anglais.

Cependant, ces méthodes présupposent la disponibilité non seulement de dictionnaires électroniques – même si les auteurs soulignent que des dictionnaires non spécialisés et très basiques sont suffisants –, mais aussi d'un analyseur morphologique capable de produire des résultats satisfaisants. Or, les paires de langues vérifiant une telle condition sont encore restreintes. Pour une opération élémentaire telle que l'alignement, il est préférable de concevoir des algorithmes ne dépendant pas trop de moyens extérieurs.

Hwang essaie de résoudre les problèmes liés justement à l'absence de ces moyens extérieurs. Il a fait le choix de spécialiser entièrement ses travaux à une paire donnée, japonais-coréen, et a cherché à exploiter au maximum les particularités propres à cette paire de langues et favorables à l'alignement. Cette approche est intéressante – en dépit d'une absence totale de portabilité de l'algorithme – dans la mesure où elle indique une direction, complètement opposée au courant classique, pour la conception d'outils multilingues capables de traiter les paires de langues européenne/non-européenne et surtout les paires de deux langues non-européennes.

CHAPITRE 3

ÉLABORATION D'UN SYSTÈME D'ALIGNEMENT AUTOMATIQUE AU NIVEAU PHRASTIQUE : AlALeR

あられ　霰【arare】n.

1. Perle de glace. **2.** Petit biscuit de riz. **3.** inform. ***AlALeR*** *(système d'**Al**ignement **A**utonome, **Lé**ger et **R**obuste)* Aligneur adapté au traitement du japonais caractérisé par l'absence d'utilisation d'analyseur morphologique et de dictionnaire.

Ce chapitre est consacré à la présentation de notre système d'alignement des phrases, AlALeR, adapté au traitement du japonais, qui ne recourt à aucun moyen extérieur, ni dictionnaire ni analyseur morphologique, en mettant pleinement à profit certaines particularités du système d'écriture du japonais. Nous allons présenter tout d'abord les problèmes des systèmes existants et nos deux éléments de solution (§ 3.1) dont nous aborderons par la suite la mise en œuvre : amélioration de la segmentation sans analyseur morphologique (§ 3.2) et ancrage fiable par alignement des mots empruntés en katakana *(§ 3.3). L'exposé se poursuivra par la description du fonctionnement (§ 3.4) et de l'optimisation de la structure de données utilisée (§ 3.5). Enfin, la dernière partie du chapitre sera consacrée à l'évaluation du système (§ 3.6).*

3.1 Systèmes existants et nouveauté de notre système

3.1.1 Problèmes à résoudre

Les recherches sur la technique d'alignement ont débuté dans le cadre de travaux sur la traduction automatique. Si bien que les précurseurs ont cherché avant tout la simplicité de réalisation et de calcul, donnant ainsi naissance à des méthodes caractérisées par l'utilisation exclusive d'informations internes telles que la distribution lexicale (Kay & Röscheisen, 1993) ou la longueur des phrases (Brown et al., 1991 ; Gale & Church, 1993) (se référer au chapitre 2 pour plus de détails).

Les chercheurs occidentaux ont choisi, pour améliorer la technique, la poursuite de la voie initiée par ces précurseurs en introduisant de nouvelles notions telles que les cognats (Simard et al., 1992 ; Langlais, 1997 ; Kraif, 2001), qui ne font pas appel aux informations extérieures. Ils ont également développé la notion d'ancrage, déjà présente dans les travaux des précurseurs, pour obtenir une meilleure robustesse au bruit dû au formatage ou aux erreurs de traduction.

Néanmoins, du fait que le système d'écriture du japonais ne dispose pas de séparateur graphique indiquant les frontières entre les mots, les chercheurs japonais ont intégré très tôt des analyseurs morphologiques dans leurs systèmes d'alignement (Murao, 1991). De plus, le japonais est fortement différent des langues principalement traitées dans le TAL – telles que l'anglais, le français ou l'allemand – aussi bien sur le plan syntaxique que sur le plan lexical, ce qui n'a pas permis une simple application des méthodes utilisées pour ces langues au traitement des textes japonais. Aussi, les Japonais ont-ils également dû recourir à des dictionnaires bilingues et rechercher la performance plutôt que la simplicité (Haruno & Yamazaki, 1996).

Mais, est-il vraiment impossible de réaliser l'alignement de phrases de langues ayant une structure très différente sans l'aide de moyens extérieurs ? Sur le plan théorique, il nous a semblé, comme il a été déjà dit, que la méthode de Kay (distribution lexicale), si logique du point de vue linguistique, était tout à fait applicable au traitement du japonais. Mais nous avons également remarqué que l'introduction de certaines connaissances linguistiques était indispensable.

3.1.2 Nos solutions

Afin de concevoir un système autonome, les questions essentielles se résument en deux points : la possibilité de segmentation sans analyseur morphologique et la détermination d'ancrages sûrs comme ceux produits par l'appariement des cognats.

Nous avons élaboré une solution à ces deux grands problèmes de la façon suivante : la segmentation est réalisée, sans analyseur morphologique, par une analyse morphologique partielle basée sur une méthode traditionnelle qui profite d'une particularité du système d'écriture du japonais, possédant plusieurs types

de caractères différents ; l'absence ou l'insuffisance de cognats permettant de réaliser un préalignement d'ancrage fiable peut être compensée par l'exploitation des mots emprunts, entraînant l'obtention d'un meilleur alignement des mots sans recourir à un dictionnaire bilingue.

Nous présentons dans les sections suivantes ces solutions de manière plus détaillée.

3.2 Segmentation sans analyseur morphologique

3.2.1 Méthode classique de segmentation par type de caractère

Comme nous l'avons vu dans la section 2.1.8, l'extraction des mots de l'algorithme de Kay, basée sur l'existence de séparateurs graphiques, pose des problèmes pour les langues comme le japonais qui ne possèdent pas de signes permettant de segmenter les phrases *a priori*.

Si nous cherchions à segmenter entièrement la phrase, il nous faudrait un système d'analyse morphologique du japonais, dont l'objectif est justement de segmenter la phrase. Toutefois, il existe également une méthode classique d'analyse morphologique partielle permettant d'extraire, ne serait-ce que partiellement, les mots graphiques sans aucune connaissance extérieure, appelée segmentation par type de caractère[1]. En effet, il est possible de reconnaître la plupart des mots lexicaux en profitant d'une particularité du système d'écriture du japonais qui utilise trois types de caractères différents selon la nature des mots : *hiragana*, *katakana* et *kanji*[2].

- ***hiragana*** : premier syllabaire japonais souvent utilisé pour représenter la partie variable des mots variants et les mots grammaticaux ;
- ***kanji*** : idéogrammes utilisés pour représenter les mots lexicaux et les radicaux ayant un sens ;
- ***katakana*** : second syllabaire japonais employé pour la transcription des mots emprunts des langues étrangères (à l'exception du chinois).

Ainsi, comme le montre la figure 3.1 (voir page suivante), il est possible de reconnaître la plupart des mots lexicaux en extrayant les séquences de *kanji* ou de *katakana*. C'est d'ailleurs une des méthodes de segmentation utilisée pour l'analyse morphologique. Néanmoins, il existe de nombreuses exceptions telles que le cas où le changement de type de caractère se trouve à l'intérieur d'un mot. Il est donc impossible de couper totalement de manière correcte une phrase uniquement avec cette méthode.

Toutefois, étant donné, comme nous l'avons vu précédemment, que nous n'avons besoin justement que des mots lexicaux pour l'algorithme de distribution lexicale, et que l'extraction peut même ne pas être complète – puisque ce n'est pas

[1] Pour une présentation de cette méthode, se référer au « Chapitre II : Méthode de segmentation » de Nakamura-Delloye (2003a).

[2] Pour plus de détails sur le système d'écriture du japonais, voir le « Chapitre I : Notions de linguistique japonaise » dans Nakamura-Delloye (2003a).

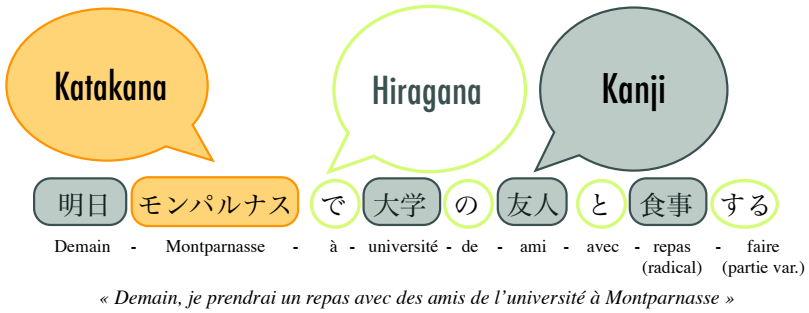


FIG. 3.1 – Phrase japonaise constituée de trois types de caractères

notre objectif principal –, cette méthode est sans aucun doute tout à fait suffisante pour notre système.

3.2.2 Amélioration proposée par Rayon

Rayon (2003) propose une amélioration de cette méthode de segmentation par type de caractère : par examen du contexte droit des séquences de *kanji*, l'auteur crée des règles permettant d'identifier à partir de leur contexte droit, la catégorie grammaticale de séquences de *kanji* et de rajouter, lorsqu'il s'agit d'un mot variable, la terminaison constituée de caractères *hiragana*. Son système réalise non seulement l'étiquetage des séquences de *kanji* extraites – éventuellement avec l'identification de leur terminaison –, mais aussi la lemmatisation des mots variables.

Mais, les deux problèmes principaux de ce type de segmentation n'ont pas été résolus. Premièrement, quand le changement de type de caractère se trouve à l'intérieur d'un mot, le système est incapable de l'identifier en tant qu'une unité et le segmente en autant de morceaux qu'il y a de changements de type de caractère. Par exemple, le mot ソ連 (*so-ren*, URSS) est segmenté entre ソ (*so*) en *katakana* et 連 (*ren*) en *kanji*.

Deuxièmement, quand la frontière entre les deux mots composants n'est pas marquée par un changement de type de caractère, le système reconnaît la séquence comme un seul mot sans réaliser la segmentation adéquate. Par exemple, la séquence 電気店街 (*den-ki-ten-gai*, quartier de l'électronique grand public) est reconnue comme un mot, alors qu'elle aurait dû être segmentée plutôt en trois éléments, 電気 (*den-ki*, électricité), 店 (*ten*, magasin) et 街 (*gai*, quartier). De même, un adverbe constitué entièrement en *kanji* sans être suivi d'une particule se retrouve inclus dans le mot qui le suit. La séquence 時々無性に (*toki-doki-mu-shô-ni*) est constituée de deux adverbes, 時々 (*toki-doki*, parfois) et 無性に (*mu-shô-ni*, extrêment), mais le premier étant entièrement en *kanji*, le système n'a pas pu reconnaître la frontière entre ces deux adverbes.

Pour l'appariement grossier des mots réalisé dans le cadre de l'alignement de

phrases, l'étiquetage et la lemmatisation sont des opérations non indispensables. En effet, la méthode d'alignement des mots basée sur la comparaison de leur distribution – proposée par Kay (voir la section 2.1) – est caractérisée par l'analyse morphologique partielle qui précède l'opération d'appariement, et réalise l'alignement des unités correspondant seulement aux parties porteuses de sens (radicaux) et ce sans faire de distinction des mots selon leur catégorie. En revanche, les questions de segmentation non résolues représentent un problème non négligeable car elles risquent d'entraîner la correspondance d'un mot graphique japonais avec deux ou plusieurs mots du texte français, ou l'inverse.

Nous devons donc trouver une autre solution qui convienne mieux à notre opération d'alignement.

3.2.3 Notre amélioration pour la segmentation des mots composés

Le second problème décrit précédemment portait sur les séquences de mots composés constituées de plusieurs substantifs juxtaposés les uns derrière les autres. Dans ce type de séquence, généralement entièrement en *kanji* ou en *katakana*, la frontière entre les deux mots composants n'est pas marquée par un changement de type de caractère. Mais, il nous paraît possible de traiter ces séquences avec la méthode utilisée pour l'étape morphologique dans l'algorithme de Kay & Röscheisen (1993), méthode que nous avons adoptée pour l'étape de lemmatisation des mots français dans notre système. Elle consiste, comme nous l'avons déjà vu, à trouver les sous-chaînes préfixales ou suffixales communes à plusieurs formes effectives des mots graphiques et à déterminer les radicaux, porteurs de sens. Il s'agit donc de la recherche des sous-chaînes préfixales communes à plusieurs formes effectives. La différence dans le cas du japonais est que les parties restantes ne sont pas des suffixes mais un ou même plusieurs autres mots portant eux-mêmes un sens propre. On obtient donc à partir d'un mot graphique ab, non pas sa forme de base a, mais deux formes de base a et b.

Unité minimum

Mais la division d'un mot constitué de plusieurs *kanji* en plus petites unités munies de sens, donnerait un nombre de morphèmes exactement égal au nombre de *kanji*, puisque posséder un sens est la nature même des idéogrammes.

Malheureusement, un *kanji* correspond rarement à un mot graphique des langues occidentales. Par exemple, le terme « politique » en japonais est constitué de deux *kanji* : « affaires de l'État » et « assumer » ; de même, tous les noms de domaine d'étude sont constitués de un ou plusieurs *kanji* désignant l'objet de l'étude suivi du *kanji* « étude » : « nombre » + « étude » = « mathématiques », « médical » + « étude » = « médecine », (« vivre » + « chose ») + « étude » = « vivant » + « étude » = « biologie ».

Il existe également des mots pour lesquels il est difficile de trouver un lien entre leur sens et celui de chacun des *kanji* les composant sans mener des re-

cherches étymologiques en chinois. Si bien que découper les mots en *kanji* est sans doute inefficace, voire nuisible à l'alignement.

Nous allons ici poser comme hypothèse que la succession de deux *kanji* forme un ensemble dont le sens est plus concret que celui de *kanji* pris séparément les uns des autres. Il est en effet plus aisé de trouver la correspondance entre une séquence de deux *kanji* et les mots graphiques des langues occidentales. Bien qu'elle n'ait pour l'instant aucune justification linguistique, cette hypothèse sera à la base de notre système, qui cherchera à trouver les séquences de deux *kanji*.

Mécanisme de segmentation

La méthode de Kay & Röscheisen (1993) repose, comme nous l'avons déjà vu, sur la structure de données *trie*. La figure 3.2 représente un exemple d'arbres vérifiant des chaînes préfixales et suffixales, créés à partir de sept entrées. Elle montre comment segmenter les mots japonais à l'aide de ces arbres.

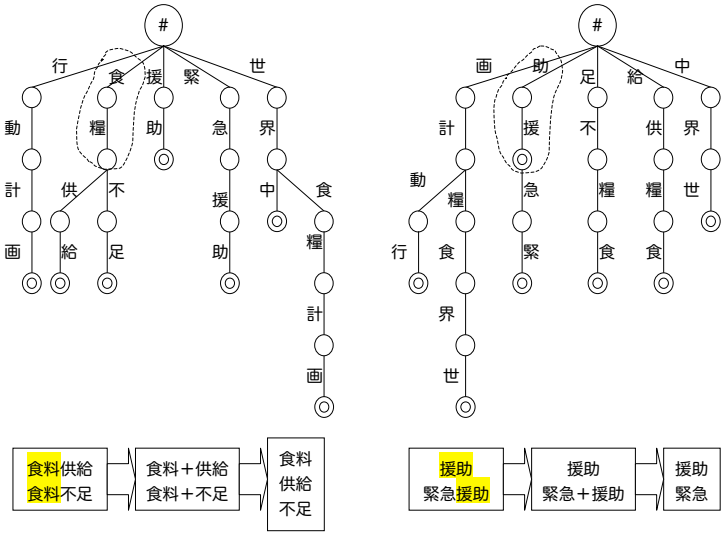


FIG. 3.2 – Arbres vérifiant des chaînes préfixales (fig. de gauche) et suffixales (fig. de droite)

L'arbre vérifiant des chaînes préfixales sert à trouver les chaînes préfixales communes à plusieurs mots et l'arbre vérifiant des chaînes suffixales, celles communes à plusieurs mots.

Étudions l'arbre vérifiant des chaînes préfixales de la figure 3.2. De la deuxième branche en partant de la gauche, étiquetée 食糧, et dérivant en deux branches étiquetées 供給 et 不足, on extrait la chaîne commune 食糧 (*shokuryô*, nourriture) et les deux chaînes suffixales, 供給 (*kyôkyû*, offre ; ravitaillement) et 不足 (*fusoku*, manque).

Les lemmes ainsi obtenus sont regroupés en une liste, appelée liste des lemmes. Nous réalisons ensuite, pour tous les mots lexicaux, la vérification, par

consultation de cette liste, qu'ils ne contiennent pas un autre lemme plus court. Si la sous-chaîne préfixale *a* du mot considéré *abc* appartient à la liste des lemmes, ce dernier est segmenté en *a* et *bc*, nous recommençons la vérification avec la partie restante *bc*.

En réalisant ainsi l'ensemble de ces opérations, nous pouvons réaliser la segmentation d'un mot graphique en plusieurs lemmes, lorsque la séquence en contient plus de deux. Nous évitons tout de même l'excès de segmentation à l'aide d'une règle reposant sur l'hypothèse que nous avons posée, à savoir que la succession de deux *kanji* forme un ensemble dont le sens est plus concret que celui de *kanji* pris séparément les uns des autres.

Les séquences de *katakana* sont segmentées de la même manière. Néanmoins, dans le cas des séquences de *katakana*, nous ne cherchons pas de sous-chaînes communes à plusieurs mots graphiques, mais des sous-chaînes semblables à un autre mot graphique ou à un lemme extrait d'un mot graphique. Ainsi, on empêche par exemple la segmentation de la séquence インストール (*insutôru*, installation) en deux parties, イン (*in*) et ストール (*sutôru*), du fait de l'absence de mots ou lemmes semblables, même si la première partie イン (*in*) est la sous-chaîne commune avec la séquence イントロダクション (*intorodakushon*, introduction).

L'algorithme, que nous avons développé, de segmentation des séquences constituées entièrement de *kanji* à l'aide de *trie* est présenté dans l'annexe A.1.

3.3 Ancrage fiable par alignement des mots en *katakana*

Le deuxième sujet important était de chercher un moyen de compenser l'absence ou l'insuffisance de cognats permettant de réaliser un préalignement d'ancrage fiable. Nous avons alors posé comme hypothèse que l'exploitation des mots emprunts entraînerait l'obtention d'un meilleur alignement des mots sans recourir à un dictionnaire bilingue.

Les mots en *katakana* sont des transcriptions des mots emprunts des langues étrangères (à l'exception du chinois). Il est donc largement possible de trouver le mot « original » français (ou le cognat du mot original) si on arrive à les retranscrire en alphabet latin.

Étant donné leur nombre limité – aussi bien pour les *katakana* que pour les lettres de l'alphabet latin –, les règles de retranscription sont sans doute définissables assez facilement. Cependant, à cause de la différence de système phonétique/phonologique entre le japonais et le français, la retranscription de la transcription risque d'être un peu voire assez différente du mot original.

Toutefois, nous pouvons également imaginer que la retranscription, quoique différente, reste une forme assez proche à la manière d'un cognat. Or, la mise en correspondance des cognats, mots à forme non totalement identique, est largement étudiée dans le cadre de travaux sur l'alignement entre les textes de langues apparentées. Il est donc tout à fait possible d'envisager la détermination de ces

pseudo-cognats par calcul de la similarité de forme.

3.3.1 Grammaire de retranscription et transducteur

Grammaire

Une retranscription peut être définie par trois éléments : la séquence d'un ou plusieurs caractères de sortie de l'étape précédente – i.e. une/des lettre(s) de l'alphabet latin –, le caractère d'entrée à traiter – i.e. un *katakana* – et la séquence d'un ou plusieurs caractères de sortie – i.e. une/des lettre(s) de l'alphabet latin –, et peut être représentée formellement comme suit :

$$(r,\ \alpha) \rightarrow t$$

où
r est la séquence d'un ou plusieurs caractères de sortie de l'étape précédente ;
α est le caractère d'entrée courant ;
t est la séquence d'un ou plusieurs caractères de sortie.

Le système de retranscription défini par un tel ensemble de règles peut être réalisé par un transducteur dont les symboles d'entrée sont des *katakana* et dont les symboles de sortie sont ceux constitués d'une ou plusieurs lettres de l'alphabet latin.

La grammaire de notre transducteur de retranscription, que nous avons définie spécifiquement, – détaillée dans l'annexe A.2 « Grammaire de retranscription des *katakana* » – est constituée d'un ensemble de règles de transition et d'un ensemble de règles de sortie :

- les règles de transition $t(E_i, \alpha; E_j)$ indiquent la transition provoquée par chaque symbole d'entrée α de l'état dit de départ E_i à l'état dit suivant E_j ;
- les règles de sortie indiquent le(s) symbole(s) de sortie lié(s) à chaque état.

Une règle de retranscription est donc décomposée en une règle de transition et deux (ou plusieurs[3]) règles de sortie.

Considérons la règle de retranscription :

$$(q,\ \alpha) \rightarrow r$$

À partir de cette règle, nous créons d'abord deux règles de sortie pour r et t telle que :

$$s(E_m\ ;\ q)$$
$$s(E_n\ ;\ r)$$

où
E_m est un état,

[3]Dans le cas où l'état possède plusieurs symboles de sortie (cf. paragraphe suivant « Pluralité des symboles de sortie liés à certains états »).

q est le symbole de sortie lié à cet état E_m,
E_n est un état, et
r est le symbole de sortie lié à cet état E_n.

Ensuite, nous définissons une règle de transition telle que :

$$\texttt{t(E}_m\texttt{, } \alpha \texttt{ ; E}_n\texttt{)}$$

où
E_m, état de départ, est l'état auquel est lié le symbole de sortie q,
α est le symbole d'entrée courant, et
E_n, état suivant, est l'état auquel est lié le symbole de sortie r

Cette règle de transition indique que le transducteur possède un chemin s'étendant de l'état E_m à l'état E_n, et étiqueté par le symbole d'entrée α.

Avec la plupart des *katakana*, on passe au(x) même(s) état(s) quel que soit l'état de départ. Les règles avec la variable X, du type $t(X,\alpha;a)$ qui apparaissent dans l'annexe A.2, signifient que, quel que soit l'état de départ, avec α comme symbole d'entrée, on passe à l'état a.

Particularités du transducteur

Le transducteur créé à partir de la grammaire présentée précédemment possède les trois particularités suivantes.

Pluralité des symboles de sortie liés à certains états Étant donné qu'un *katakana* peut être retranscrit en différents caractères, comme par exemple « カ » par « ka » ou « ca », à un état peuvent être liés plus d'un symbole de sortie. Le choix de ces candidats n'étant pas réalisable dans la plupart des cas – du moins avec seulement leur contexte immédiat –, ce système risque de provoquer une explosion combinatoire. Pour minimiser ce risque, nous avons limité le nombre maximum de symboles de sortie liés à un état à 2 (cf. exemple 3 page 470).

État vide Il existe également un état auquel n'est lié aucun symbole de sortie. Dans la grammaire de l'annexe A.2, cet état est représenté par le symbole 0, et sa règle de sortie `s(0 ; 0)` indique l'absence de symbole de sortie lié à cet état. En effet, les *katakana*, « ッ » indiquant la gémination et « ー » indiquant le prolongement de la voyelle, ne sont souvent pas marqués sur la forme du mot d'origine, d'où la nécessité d'un tel état (cf. exemple 4 page 471).

Néanmoins, certains *katakana* n'apparaissent qu'après certains caractères donnés. Les règles de transition de ce type de *katakana* énumèrent explicitement tous les états de départ possibles.

Transducteur non déterministe Par ailleurs, il existe également des *katakana* transcrits différemment selon le *katakana* qui les suit. C'est le cas de l'ensemble

des caractères qui peuvent représenter avec un caractère de petite taille une syllabe constituée avec une semi-voyelle. Par exemple, キ est transcrit par « k » s'il précède ャ, ュ ou ョ pour constituer le syllabe « kya », « kyu » ou « kyo » alors que s'il n'est pas suivi par l'un de ces trois caractères, il y a de fortes chances pour qu'il soit transcrit par « ki ».

Lorsqu'un de ces *katakana* constitue le symbole d'entrée, l'état suivant n'est pas sélectionnable de manière décisive. Notre transducteur est donc un automate d'états finis non déterministe. Afin de limiter au maximum la complexité de calcul, le choix n'est conservé qu'aux états qui suivent directement l'état à choix multiple. L'algorithme a été conçu de manière à rendre pertinent l'ordre des états suivants contenu dans une règle de grammaire : dès qu'on découvre que l'un des états suivants permet de continuer le calcul (la retranscription), on abandonne tous les autres candidats pour l'état suivant figurant après ce dernier dans la grammaire (cf. exemples 5 page 474 et 6 page 476).

Dans les annexes A.3 et A.4 sont présentés notre algorithme de retranscription par transducteur et des exemples permettant d'éclaircir cet algorithme ainsi que les explications précédemment décrites.

L'annexe A.5 montre un exemple de résultat de retranscription par le transducteur. Comme nous pouvons le constater, l'ajout de quelques règles supplémentaires peut tout à fait améliorer les résultats. Néanmoins, en limitant le nombre maximum de symboles de sortie liés à un état à 2, le nombre maximum de combinaisons est déjà en $O(lg^2)$ où lg est la longueur (le nombre de caractères, *katakana*) de la séquence entrée.

Par ailleurs, nous pouvons nous demander jusqu'à quel degré l'adaptation à une langue particulière, en l'occurrence le français, est indispensable. Par exemple, pour la comparaison avec les mots français, on a l'intuition que la règle de transcription du *katakana* ウ (*u*) en « ou » est essentielle. Mais, en réalité, étant donné que les mots en *katakana* sont principalement (pour des textes du domaine informatique, en particulier) des mots emprunts de l'anglais, cette règle, qui augmente le nombre de combinaisons, n'est pas forcément indispensable. Toutefois, pour les textes littéraires dont l'original est en français, elle est incontournable pour retranscrire correctement les noms propres, d'autant plus que dans ce type de texte les noms propres sont les éléments principaux permettant de réaliser un ancrage simple et fiable.

Notre grammaire est le fruit de la recherche d'un équilibre entre performance et optimisation de calcul, mais nous ne nions aucunement la possibilité de définir efficacement une grammaire par d'autres façons.

3.3.2 Calcul de similarité

Comme nous l'avons dit au début de cette section, à cause de la différence de système phonétique/phonologique entre le japonais et le français (ou d'autres langues auxquelles appartiennent les mots d'origine des mots en *katakana*), la re-

transcription de la transcription risque d'être un peu, voire assez, différente du mot original. De plus, étant donné que le nombre de règles est limité en raison d'une optimisation des calculs, il risque d'y avoir beaucoup de caractères manquants ou superflus.

Afin de supporter la divergence et de trouver de manière robuste l'équivalence entre les mots d'origine et leur retranscription, nous recourons aux méthodes de mise en correspondance des cognats largement étudiées notamment dans le cadre de travaux sur l'alignement entre les textes de langues apparentées (voir la section 2.3.1).

La méthode de calcul de similarité entre une séquence retranscrite et un mot français que nous avons adoptée est proche de celle de la sous-chaîne maximale parallèle utilisée dans Kraif (2001) pour la reconnaissance des cognats, que nous avons présentée dans la section 2.3.1. Notre formule, adaptée aux besoins particuliers de la retranscription des *katakana*, est définie comme suit :

Soient `chfr`, chaîne en français, et `chjp`, chaîne de retranscription d'un mot en *katakana*.
La similarité `sim` de `chfr` et `chjp` est :

$$\texttt{sim} = p(\text{SCM}) \cdot \frac{2\text{SCM}}{L_1 + L_2 - L_3} \cdot \frac{2\text{CCM}}{L_4 + L_5}$$

où

- L_1 = `longueur(chfr)`
- L_2 = `longueur(chjp)`
- L_3 = `nombre`('u' dans `chjp`)
- L_4 = `nombre`(consonnes dans `chfr`)
- L_5 = `nombre`(consonnes dans `chjp`)
- SCM = sous-chaînes maximales communes
- CCM = coût calculé à partir des consonnes communes de `chfr` et `chjp`
- p(SCM) = poids basé sur les sous-chaînes maximales communes

Notre formule diffère de celle de Kraif (2001) par le fait qu'elle tient compte non seulement de la sous-chaîne maximale mais aussi des consonnes communes. Le nombre de consonnes communes est pris en compte pour favoriser les deux chaînes ayant le plus de caractères consonantiques communs plutôt que celles dont les caractères vocaliques coïncident le plus.

Afin de calculer la longueur de CCM, on extrait d'abord toutes les lettres consonantiques des chaînes considérées. On ne considère aucune paire de deux chaînes dont les longueurs de séquences extraites sont trop différentes, c'est-à-dire dans notre méthode :

$$\frac{|L_4 - L_5|}{\texttt{max}(L_4, L_5)} \geq \frac{1}{2}$$

Pour les paires remplissant cette première condition, la longueur de CCM est ensuite calculée par la méthode de calcul des sous-chaînes maximales communes de

Kraif, à la différence qu'au lieu de rejeter les sous-chaînes représentant des décrochements consécutifs – c'est-à-dire des insertions ou des suppressions qui ne sont pas entourées de caractères identiques –, on donne une pénalité à chaque insertion dans la chaîne retranscrite d'une lettre consonantique[4] n'appartenant pas au mot original (potentiel) et à chaque suppression dans la chaîne retranscrite d'une lettre consonantique[5] n'appartenant pas au mot original (potentiel).

L'insertion en fin de chaîne retranscrite est également pénalisée (sauf « *y* » et « *w* ») tandis que la suppression ou la divergence entre les sous-chaînes préfixales ne sont pas pénalisées.

Ces règles traduisent le constat que les dernières lettres consonantiques supprimées correspondent souvent à des morphèmes grammaticaux (typiquement le « s » du pluriel) tandis que les dernières lettres superflues dans la retranscription ont une forte possibilité d'indiquer que ce n'est pas une retranscription du mot français considéré, par exemple entre :

mot fr : « sct » (= soc*i*été) et ;

retranscription jp : « sctm » (= sicetemu, une des retranscriptions de システム (système)).

Lorsqu'il y a une/des insertions et une/des suppressions en fin de chaîne – c'est-à-dire lorsque les deux chaînes ont une terminaison différente –, le choix n'est pas aussi évident. Il est possible que les lettres supprimées soient des morphèmes français et que les lettres insérées soient des morphèmes équivalents d'une autre langue (typiquement l'anglais) à laquelle appartient le mot d'origine du mot japonais retranscrit.

mot fr : « prtnrt » (= partenariat) et ;

retranscription jp : « prtnrshp » (= partonashipu, une des retranscriptions de パートナーシップ (*partnership*, ang.)).

Afin de bien prendre en compte cette possibilité, notre méthode ne donne pas de pénalité à ces cas, accordant de l'importance surtout à leur similarité.

Par ailleurs, le caractère « *u* » est ignoré lors du comptage de la longueur de la chaîne japonaise, car sa présence est souvent due à la « japonisation » – une consonne est toujours accompagnée d'une voyelle – des mots étrangers par insertion d'une voyelle entre deux consonnes adjacentes.

p(SCM) permet de favoriser les paires de chaînes ayant une sous-chaîne commune longue – plus cette sous-chaîne est longue, plus la paire est favorisée. Dans notre méthode, *p*(SCM) est défini comme log(SCM).

[4] Exceptées « *y* » et « *w* » utilisées comme des lettres vocaliques dans notre grammaire de retranscription (ex. « sisutemu » pour « système »).

[5] Exceptées « *y* », « *w* » et « *h* » représentées souvent par une lettre vocalique ou absentes dans la retranscription (ex. « babilonia » pour « babylonien » et « caludea » pour « chaldéen »).

Exemples

Le tableau 3.3 montre des exemples de résultat de calcul de similarité par notre formule.

1. `kananasukisu ---> kananaskis [1,000000]`
2. `contacuto ---> contact [0,788758]`
3. `puroguramu ---> programme [0,672237]`
4. `gurupu ---> groupe [0,535164]`
5. `baiotecunolozi ---> biotechnologies [0,510204]`
6. `partonarsipu ---> partenariat [0,505225]`
7. `sabusahara ---> subsaharienne [0,448158]`
8. `sisutemu ---> systèmes [0,399411]`

TAB. 3.3 – Similarités entre des retranscriptions et leur mot d'origine

Exemple 1 Considérons deux chaînes, un mot français et une retranscription d'un mot japonais :

« contact » et « contacuto »

1. L_1 = `longueur(chfr)` = 7 ;
2. L_2 = `longueur(chjp)` = 9 ;
3. L_3 = `nombre`('u' dans `chjp`) = 1 ;
4. L_4 = `nombre`(consonnes dans `chfr`) = 5 ;
5. L_5 = `nombre`(consonnes dans `chjp`) = 5 ;
6. SCM = sous-chaînes maximales communes = 7 ;
7. CCM = coût calculé à partir des consonnes communes = 5 ;
8. p(SCM) = poids basé sur les SCM = log(SCM) = log(7) ;

$$\texttt{sim} = \log(7) \cdot \frac{2\times 7}{7+9-1} \cdot \frac{2\times 5}{5+5} = 0,788758$$

Exemple 2 Considérons deux chaînes, un mot français et une retranscription d'un mot japonais :

« systèmes » et « sisutemu »

1. L_1 = `longueur(chfr)` = 8 ;
2. L_2 = `longueur(chjp)` = 8 ;
3. L_3 = `nombre`('u' dans `chjp`) = 2 ;
4. L_4 = `nombre`(consonnes dans `chfr`) = 6 ;

5. L_5 = `nombre`(consonnes dans `chjp`) = 4 ;
6. SCM = sous-chaînes maximales communes = 5 ;
7. CCM = coût calculé à partir des consonnes communes = 4 (consonnes communes) - 0 (pénalité : « y » et « s » en fin de chaîne ne sont pas pénalisées) = 4 ;
8. p(SCM) = poids basé sur les SCM = log(SCM) = log(5) ;

$$\texttt{sim} = \log(5) \cdot \frac{2 \times 5}{8+8-2} \cdot \frac{2 \times 4}{5+5} = 0,399411$$

3.3.3 Études connexes

L'alignement des mots en *katakana* par retranscription n'est pas une idée nouvelle. Différents articles tels que Tsuji (2002) proposent des méthodes d'alignement des mots *katakana* avec les termes équivalents en anglais. Il existe même des travaux de cette nature sur le couple de langues français-japonais (Tsuji et al., 2002). Ces travaux se caractérisent en ce qu'ils se fondent, pour construire les règles de translittération, sur les paires de mots japonais-français et japonais-anglais, extraites des dictionnaires. Les auteurs considèrent que cette utilisation des paires non seulement japonais-anglais mais aussi japonais-français permet d'obtenir des séquences retranscrites qui correspondent bien aux règles orthographiques du français, favorisant ainsi l'alignement des mots en *katakana* dont le mot d'origine est un nom propre français.

Nos travaux diffèrent de ces derniers, sans parler de l'utilisation du transducteur, par le fait que nous retranscrivons les mots katakana principalement suivant les règles Hepburn et que la « japonisation » des mots emprunts est prise en compte dans le calcul de la similarité des chaînes. De même, les règles orthographiques françaises, différentes de celles de l'anglais, n'ont pas été introduites spécifiquement : nous avons considéré que notre méthode de calcul de la similarité entre la chaîne retranscrite et les mots français, supporterait efficacement ces éventuelles divergences puisque, justement, elles sont basées sur différentes méthodes conçues pour déterminer les cognats. Notre choix était plutôt de ne pas multiplier le nombre de règles de retranscription pour éviter d'éventuels risques d'explosion combinatoire.

3.4 Fonctionnement du système

3.4.1 Schéma général du système

Le système reçoit comme données une paire de textes parallèles rédigés en français et en anglais, ou plus particulièrement d'un texte en français (ou en anglais) et d'un en japonais. Afin de s'affranchir des problèmes d'encodage, fré-

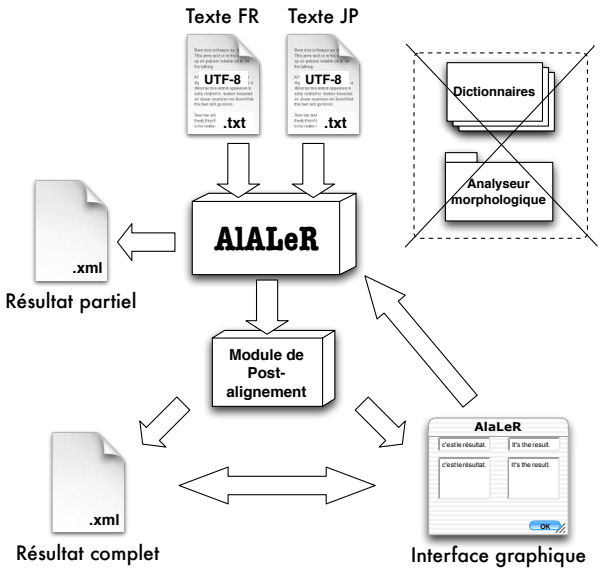


FIG. 3.4 – Schéma général du Système `AlALer`

quents lorsqu'il s'agit de traitements multilingues, `AlALeR` présuppose comme entrées des textes bruts au format texte, encodés en UTF-8[6].

Le système peut fournir comme résultat soit un alignement partiel très fiable des textes entrés, soit un alignement complet avec l'option « complet ». Lorsque cette option est choisie, le module de post-alignement réalise un appariement des phrases qui n'ont pas été alignées pendant le processus principal. L'appariement de ce module est réalisé selon la probabilité d'alignement de paires de phrases, calculée à partir de la corrélation de leur longueur.

Les résultats sont fournis, soit sous forme de fichier XML, soit par transfert vers l'interface graphique. Celle-ci permet non seulement de visualiser le résultat sous un format plus agréable à lire, mais aussi de faciliter la vérification et éventuellement la modification manuelle des résultats fournis par le système[7].

3.4.2 Procédure générale

La procédure générale est constituée de deux grandes étapes et d'une étape optionnelle produisant le résultat complet :

1. **Étape de construction de l'index du lexique**, au cours de laquelle les mots graphiques sont triés pour constituer quatre listes selon leur nature : transfuges, cognats, *katakana* et mots lexicaux.
2. **Procédure d'alignement**

[6] Les problèmes d'encodage sont abordés dans l'annexe A.7.
[7] Le système est implémenté en langage C++ et l'interface graphique avec les API Apple Carbon.

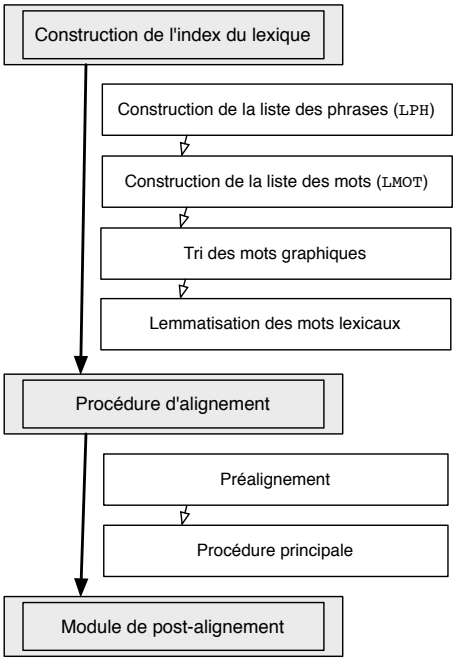


FIG. 3.5 – Ensemble de la procédure d'alignement

3. **Option complète : post-alignement et interface graphique**

Le schéma 3.5 représente l'ensemble de la procédure d'alignement du système `AlALeR`.

Nous allons maintenant présenter chacune des étapes. Étant donné que le système fonctionne un peu différemment selon les langues traitées, nous ne nous préoccupons ici que du cas d'un alignement de textes français et japonais, afin de mieux montrer la particularité de notre système.

3.4.3 Étape de construction de l'index du lexique

Cette étape est composée elle-même de quatre étapes :

1. Construction de la liste des phrases (`LPH`).
2. Construction de la liste des mots graphiques (`LMOT`).
3. Création de quatre listes à la suite du tri des mots graphiques :
   a) liste des transfuges (`LTRNS`) ;
   b) liste des cognats (`LCOG`) ;
   c) liste des mots en *katakana* (`LKTKN`) ;
   d) liste des mots lexicaux (`LEX`).
4. Création de l'index des mots lexicaux après leur lemmatisation (`ILX`).

3.4.4 Construction de l'index du lexique (1) Liste des phrases

Comme il a déjà été mentionné dans Simard & Plamondon (1998), la reconnaissance des phrases représente à elle seule, malgré l'impression de trivialité que l'on a généralement, une question à part entière. La segmentation en phrases de textes français ou anglais n'est pas évidente à cause du caractère polysémique du séparateur graphique principal de phrase, le point final.

Il est donc nécessaire de définir des règles assez détaillées permettant de segmenter correctement les séquences contenant des abréviations ou des sigles (« U.S.A »), des séquences symboliques (« abc@cdf.fr ») ou encore des nombres décimaux (1.5 en anglais).

Le point final japonais est beaucoup moins polysémique, facilitant ainsi la tâche de découpage.

À noter que dans cette étape, le système conserve chaque caractère de retour chariot deux fois : une fois à la fin de la phrase qui le précède, et la deuxième fois en tête de la phrase qui le suit. Considérés comme transfuges, ils se montrent très efficaces au moment du préalignement, en particulier lorsqu'il s'agit de textes littéraires qui ne contiennent généralement que peu voire aucun autre transfuge (voir aussi la section 3.4.9).

3.4.5 Construction de l'index du lexique (2) Extraction des mots graphiques

Lors de la deuxième étape, consacrée à la construction de la liste des mots graphiques, la liste pour le texte français est construite par extraction des séquences entourées de séparateurs graphiques des mots – préalablement définis.

Pour le texte japonais, nous réalisons une segmentation par type de caractère. Quoiqu'il soit impossible de segmenter totalement de manière correcte une phrase en mots uniquement avec cette méthode, il est possible de reconnaître la plupart des mots lexicaux en extrayant les séquences de *kanji* ou de *katakana*.

Les listes française et japonaise ainsi obtenues sont si différentes que leur simple comparaison sans y apporter aucune opération supplémentaire serait trop génératrice de bruit : la liste française contient des mots grammaticaux qui n'ont pas d'équivalents en japonais, comme les articles ou les pronoms relatifs ; la liste japonaise ne comporte presqu'aucun mot grammatical, écrit généralement en *hiragana* tel que les conjonctions, les prépositions, les auxiliaires, etc. Nous supprimons donc les mots grammaticaux de la liste `LMOT` du texte français à l'aide d'une liste de mots grammaticaux préalablement définie[8].

3.4.6 Construction de l'index du lexique (3) Tri des mots

Le tri est ensuite réalisé aussi bien pour la liste `LMOT` du texte français que pour celle obtenue à partir du texte japonais afin de construire quatre nouvelles listes :

[8] La liste des mots grammaticaux utilisée est présentée dans l'annexe A.8.

la liste des transfuges (LTRNS), la liste des cognats (LCOG), la liste des mots en *katakana* (LKTKN) et la liste des mots lexicaux (LEX).

Si nous classons les mots graphiques selon ces quatre catégories, c'est que les mots des trois premières ne nécessitent pas, contrairement à ceux de la dernière, de calcul de similarité de leur distribution pour être appariés. En effet, leur équivalence traductionnelle est calculable simplement par leur forme. Qui plus est, le résultat de ce calcul est beaucoup plus sûr que le résultat obtenu par la similarité des distributions. Cette calculabilité est assez évidente pour les deux premiers types lorsque l'on connaît leur définition.

Les cognats

Les « cognats », mots apparentés, sont des chaînes de caractères identiques ou proches graphiquement se trouvant dans les lexiques de langues ayant une relation historique plus ou moins étroite, telles que les paires anglais-français *generation/génération* et *error/erreur*. La notion de cognats améliore de manière simple et économique les méthodes statistiques qui n'utilisent aucune information lexicale, encore que son efficacité soit limitée aux langues appartenant à une même famille. Cependant, le japonais intégrant également dans son système d'écriture l'alphabet latin (ローマ字, *rôma-ji*), la possibilité d'obtention d'un résultat a été signalée très tôt dans Church et al. (1993).

Le système AlALeR ne considère comme cognats que les chaînes alphabétiques totalement identiques apparaissant dans les deux textes entrés. Le système constitue d'abord la liste LCOG du texte japonais en extrayant les mots écrits en alphabet latin. Ensuite, en se référant à la liste japonaise, il construit une liste française en recherchant les séquences identiques aux éléments de la liste japonaise.

Les paires de cognats ainsi reconnues constituent une liste, appelée table des « Cognats alignés » (COGAL).

Les transfuges

Les « transfuges » sont des chaînes invariantes à la traduction telles que les chiffres ou les symboles, inclus au début dans les cognats par les définitions traditionnelles du domaine de l'alignement, et regroupés plus tard par Langé & Gaussier (1995) pour constituer une nouvelle catégorie. Les listes de transfuges LTRANS sont constituées séparément dans les deux langues par simple extraction des séquences de symboles ou de chiffres.

Les paires constituées de deux mots appartenant aux listes LTRANS du japonais et du français, constituent ensuite la liste appelée table des « Transfuges alignés » (TRAL).

Les mots en *katakana*

La troisième liste contient les mots du texte japonais écrits en *katakana*. Le schéma 3.6 page ci-contre représente la procédure d'appariement d'un mot en

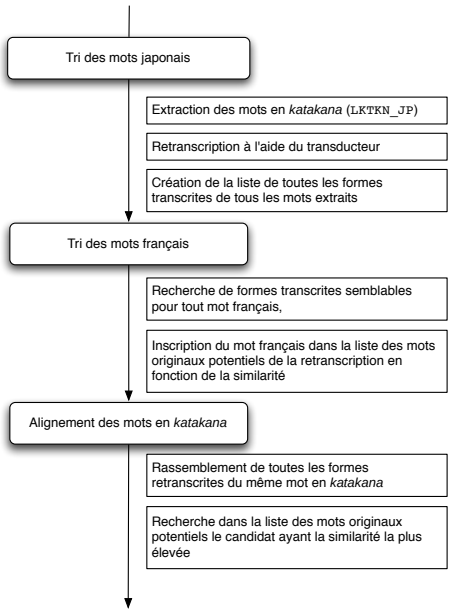


FIG. 3.6 – Procédure de retranscription et d'alignement des mots en *katakana*

katakana.

Extraits au cours du tri des mots japonais, ces transcriptions des mots emprunts sont retranscrites par le système à l'aide d'un transducteur, comme nous l'avons décrit dans la section 3.3.1, en une ou éventuellement plusieurs formes en alphabet latin. Puis, toutes les formes retranscrites des mots en *katakana* constituent la liste des « retrancriptions » (RETRANS).

Au cours du tri des mots français, pour tout mot français, on calcule la similarité entre le mot français considéré et chaque séquence de la liste RETRANS. Si la similarité avec une retranscription donnée atteint un seuil prédéfini, ce mot français est considéré comme le mot original de cette retranscription jusqu'à ce que l'on en rencontre un autre qui obtienne une similarité plus élevée.

À cette étape, on tient également compte de la similarité entre les mots français : si on trouve un mot français (par exemple « groupes ») ayant une similarité moins élevée mais pour lequel le mot considéré comme original (par exemple « groupe ») est une sous-chaîne préfixale (ou inversement, ce dernier est une sous-chaîne préfixale du mot original), on l'ajoute à la liste des mots originaux de la retranscription[9]. L'annexe A.6 montre un exemple de résultat de cette étape. On

[9]Nous n'avons pas encore à cette étape réalisé la lemmatisation des mots français : nous avons seulement la liste des mots graphiques. Nous aurions pu également réaliser la lemmatisation avant l'alignement des mots en *katakana*, mais afin d'éviter le regroupement des deux mots français ayant chacun un équivalent parmi les mots en *katakana* (par exemple, « programme » et « programmeur »), nous avons choisi cette procédure.

peut y constater la retranscription 32 « gurupu » qui possède deux mots originaux potentiels « groupe » et « groupes » avec la similarité 0,535164.

Après avoir terminé l'examen des mots français, on rassemble ensuite toutes les formes retranscrites du même mot japonais en *katakana* afin de trouver le mot français ayant la similarité la plus élevée d'une des retranscriptions. Arrivé à cette étape, on recalcule la similarité, mais cette fois la similarité de distribution afin d'exclure les correspondances hasardeuses.

Les paires composées d'un mot français et d'un mot japonais en *katakana* ainsi appariées constituent ensuite la liste appelée table des « *Katakana* alignés » (KTKNAL).

Les mots japonais en *katakana* qui n'ont pas trouvé d'équivalent une fois le tri des mots français terminé, sont stockés dans la liste des mots lexicaux pour leur laisser à nouveau une chance d'être finalement alignés par la similarité de distribution.

Le schéma 3.7 page suivante représente l'exemple d'appariement du mot en *katakana*, コンタクト (*kontakuto*).

Extrait au cours du tri des mots japonais, le mot コンタクト (*kontakuto*) est inscrit dans la liste LKTKN et retranscrit ensuite par le transducteur en quatre formes en alphabet latin qui sont stockées dans la liste des « retrancriptions » (RETRANS).

Au cours du tri des mots français (liste LMOT), on considère le mot français « contact ». La similarité avec une retranscription « contacuto » atteint le seuil prédéfini, le mot français « contact » est considéré comme le mot original de « contacuto », aucun autre candidat n'étant trouvé pendant le parcours intégral de la liste LMOT.

Ensuite, on rassemble toutes les formes retranscrites du mot コンタクト (*kontakuto*). N'ayant trouvé aucun mot original potentiel pour d'autres retranscriptions, le mot français « contact » est considéré comme le mot original de コンタクト (*kontakuto*). On vérifie leur correspondance en recalculant leur similarité de distribution et une fois qu'on constate une similarité de distribution satisfaisante, la paire « contact - コンタクト » est stockée dans la liste KTKNAL.

Les mots lexicaux

La dernière liste (LEX) contient des mots lexicaux.

La liste LEX japonaise est créée par extraction de tous les mots constitués de plus d'un *kanji*. Toutefois, les mots constitués d'un seul *kanji* ayant une fréquence importante (à savoir 12 pour notre système) sont également stockés dans cette liste.

Pour créer la liste LEX française à partir de la liste des mots LMOT, les mots grammaticaux sont tout d'abord supprimés de LMOT à l'aide de la liste des mots grammaticaux préalablement définie (voir l'annexe A.8). Les transfuges sont ensuite extraits afin de constituer la liste TRAL. Certains mots sont également extraits suite à la comparaison avec la liste LCOG du japonais et avec la liste des re-

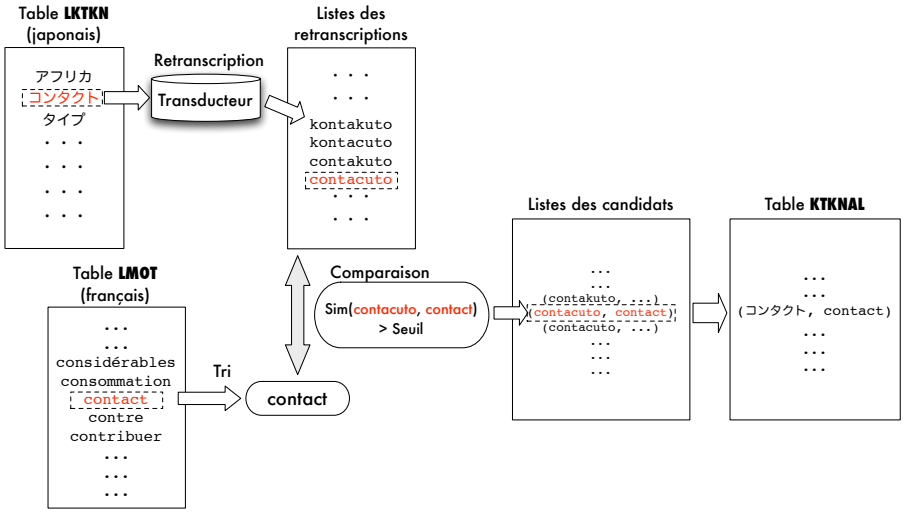


FIG. 3.7 – Appariement des mots en *katakana*

transcriptions des mots en *katakana* `RETRANS`, pour constituer respectivement les listes `COGAL` et `KTKNAL`. Le reste des mots constitue alors la liste `LEX`.

3.4.7 Construction de l'index du lexique (4) Lemmatisation des mots lexicaux

Lemmatisation des mots français

Nous avons eu recours à la méthode utilisée à l'étape morphologique dans Kay & Röscheisen (1993). Elle consiste à trouver les sous-chaînes préfixales ou suffixales communes à plusieurs formes effectives des mots graphiques et à trouver ensuite leurs radicaux, porteurs de sens. Ce traitement est implémenté efficacement grâce à l'utilisation de la structure de données appelée *trie* (cf. 2.1.2).

Lemmatisation des mots japonais

La lemmatisation des mots japonais consiste en la segmentation des séquences de mots composés constitués de plusieurs substantifs juxtaposés les uns derrières les autres.

La frontière entre les deux mots composant ce type de séquence – non marquée par un changement de type de caractère – est détectée par la recherche des sous-chaînes communes à plusieurs formes effectives.

Nous avons donc adopté, pour le japonais également, la méthode de Kay & Röscheisen (1993) reposant sur la structure de données *trie*, comme décrit dans la section 3.2.3.

Nous obtenons ainsi l'ensemble des données nécessaires à la mise en correspondance des mots permettant d'associer ensuite les phrases à aligner.

3.4.8 Procédure d'alignement

Notre système utilise une technique basée sur les informations des distributions lexicales, présentée par Kay & Röscheisen (1993). Cette méthode, reposant sur l'hypothèse que les phrases correspondantes comprennent des éléments correspondants, est constituée d'un appariement grossier des mots, qui permet ensuite l'alignement des phrases contenant les mots appariés.

Les deux textes à aligner sont représentés par une matrice dont les lignes correspondent à chacune des phrases du texte français et les colonnes à celles du texte japonais. Chaque case (i, j) est remplie au cours du traitement par des informations sur la probabilité d'alignement de deux phrases, la $i^{\text{ème}}$ phrase du texte français et la $j^{\text{ème}}$ phrase du texte japonais.

Cette étape est composée de deux sous-étapes :

1. **l'étape de préalignement**, au cours de laquelle un premier ancrage est réalisé pour limiter le nombre de possibilités d'alignement, à l'aide notamment des transfuges et des cognats ;

2. **la procédure principale**, au cours de laquelle les phrases sont alignées par un calcul de similarité de la distribution des mots qu'elles contiennent.
   Cette étape principale d'alignement, procédure itérative, est composée de trois opérations correspondant chacune à la construction d'une structure de données particulière :
   - Création de la table « Candidats des paires de phrases à aligner » (`CPR`).
   - Création de la table « Mots alignés » (`MAL`).
   - Création de la table « Résultat d'alignement » (`RAL`).

3.4.9 Procédure d'alignement (1) Préalignement

Le préalignement consiste à trouver des ancrages sûrs permettant de réduire la zone de recherche.

Considérons deux textes contenant chacun m et n phrases. Si la $k^{\text{ème}}$ phrase et la $l^{\text{ème}}$ phrase étaient alignées, l'espace de recherche serait non plus la zone $m \times n$ mais deux petites zones $k \times l$ et $(m-k) \times (n-l)$. Le résultat du préalignement influence non seulement le temps de calcul mais aussi le résultat final de l'alignement lui-même.

Le préalignement de notre système, inspiré de la méthode proposée dans Kraif (2001), est réalisé à l'aide des tables `TRAL`, `COGAL` et `KTKNAL` (présentées dans la section 3.4.6). Il se fait via deux parcours de ces tables.

Premier passage des listes

Lors du premier passage, seules les paires de mots des listes TRAL et COGAL ayant une fréquence 1 sont utilisées pour obtenir un alignement extrêmement sûr. Si plus d'une paire de phrases a été alignée, on élimine les points trop écartés de la diagonale.

Dans le cas où il n'y a aucun couple de fréquence 1, on réalise un alignement à l'aide des transfuges, notamment les retours chariots. Les retours chariots de fin de phrase permettent d'apparier les phrases en fin de paragraphe, et les retours chariots en tête de phrase indiquent les débuts de paragraphe, favorisant ainsi la mise en correspondance des premières phrases de paragraphe. Cette méthode, qui consiste à conserver deux fois un retour chariot, permet notamment d'aligner de manière sûre les titres entourés de deux retours chariots.

En créant une première table CPR de manière à définir une zone de recherche plus vaste, on cherche les co-occurrences de ces transfuges appartenant à cette table. Les points obtenus sont strictement désambiguïsés – c'est-à-dire deux points ayant un même élément, par exemple (1, 2) et (1, 3), ne sont pas retenus.

Second passage des listes

Au second passage, on travaille sur chaque zone décomposée par le résultat d'alignement du premier passage pour laquelle une nouvelle table CPR est définie. Si deux mots alignés appartenant à une des tables TRAL, COGAL ou KTKNAL ont la même fréquence dans une des zones, les paires de phrases contenant des co-occurrences de ces mots sont appariées. Après ce deuxième passage, les points isolés sont également éliminés.

Les tables de paires de mots alignés, TRAL, COGAL et KTKNAL, sont également utilisées au même titre que la table MAL lors de la constitution de la table RAL. Les paires de mots de fréquence différente, qui n'ont pas été prises en compte au cours du préalignement, sont exploitées à cette occasion.

3.4.10 Procédure d'alignement (2) Procédure principale

Table « Candidats des paires de phrases à aligner »

La table CPR est une matrice indiquant les paires de phrases susceptibles d'être alignées. Basée sur l'hypothèse de diagonalité de l'alignement, la zone constituée des cases correspondant aux paires candidates forme une ellipse avec pour axe principal la diagonale de la matrice, comme représenté figure 3.8 (voir page suivante), partie gauche. Mais, à l'issue du préalignement, la zone de travail est limitée, rendant la table CPR comme celle présentée dans la partie droite de la figure 3.8.

FIG. 3.8 – CPRs sans préalignement (à gauche) et avec (à droite)

Table « Mots alignés »

La table `MAL` contient l'ensemble des paires de mots supposés être traductions l'un de l'autre.

L'appariement des mots est réalisé selon la similarité de la distribution de chaque mot. Tous les mots appartenant à un même candidat paire de phrases sont comparés, et leur est attribuée une similarité basée sur leur distribution. De nombreuses formules ont été proposées jusqu'aujourd'hui pour le calcul de cette similarité de distribution lexicale. Notre méthode est inspirée de l'amélioration par Kitamura & Matsumoto (1997) du coefficient de Dice : en plus de la différence de fréquences, elle tient également compte de la fréquence elle-même, donnée contrôlée séparément dans les algorithmes antérieurs. La nouveauté apportée par notre formule est la prise en compte du nombre de phrases où les mots considérés apparaissent. Cette modification améliore les résultats lorsque deux paires ont une similarité identique, situation entraînant des conflits avec les méthodes précédentes.

Notre formule ainsi obtenue est définie comme suit :

Soient e_a et e_b les expressions considérées,

$$\mathsf{sim}(e_a, e_b) = p(f(e_a, e_b)) \cdot \frac{2 \cdot f(e_a, e_b)}{f(e_a) + f(e_b)} \cdot \frac{2 \cdot n(e_a, e_b)}{n(e_a) + n(e_b)}$$

où

- $f(X)$ = fréquence de la séquence X ;
- $n(X)$ = nombre de phrases où apparaît X ;
- $f/n(X,Y)$ = fréquence ou nombre de phrases des co-occurrences de X et Y ;
- $p(f(e_a,e_b))$ = poids basé sur la fréquence des co-occurrences.

Table « Résultat d'alignement »

La table `RAL` contient l'ensemble des paires de phrases supposées être traductions l'une de l'autre.

L'appariement des phrases utilise la table `MAL` obtenue précédemment, en plus des tables de paires de transfuges alignés (`TRAL`), de cognats alignés (`COGAL`) et de mots *katakana* alignés (`KTKNAL`), pour calculer combien de couples de mots de ces tables contient chaque paire de phrases appartenant à la `CPR`.

Si une paire de phrases comporte plus de paires de mots alignés que le seuil défini – en fonction de la taille du texte –, ces phrases sont considérées comme correspondantes traductionnelles. Ces nouvelles paires servant de nouvelles ancres, on crée une nouvelle `CPR` pour réaliser de manière itérative ces opérations d'alignement.

3.4.11 Module de post-alignement et interface graphique

Ce premier résultat, fiable mais partiel, peut être complété par une procédure de post-alignement plus robuste, pour être ensuite envoyé vers l'interface graphique.

Post-alignement basé sur la corrélation des longueurs

Le module de post-alignement extrait les sous-matrices constituées des phrases non alignées par le noyau `AlALeR` et calcule la probabilité d'alignement de toutes les paires possibles de phrases. Il réalise ensuite l'appariement de ces phrases avec une méthode de programmation dynamique de manière à mettre en relation toutes les phrases avec au moins une phrase de l'autre texte.

Les modèles de traduction pris en compte sont 1-1 (1 phrase japonaise et 1 phrase française en relation traductionnelle l'une de l'autre), 1-2, 2-1, 2-2, 1-3 et 3-1.

Pour toutes les possibilités de couples constitués d'une phrase du texte japonais J_i et d'une phrase du texte français F_j, on calcule le coût de chacun des six modèles à l'aide de la fonction c présentée ci-dessous. Il est calculé à partir des longueurs des phrases et du poids déterminé selon leur modèle de traduction. Le coût $c(i,j;n,m)$ du couple (i,j) avec comme modèle de traduction n-m est :

$$c(i,j;n,m) = \frac{2 \times |lg(i-(n-1),i) - lg(j-(m-1),j)|}{lg(i-(n-1)),i) + lg(j-(m-1),j)} \cdot poids(n,m)$$

où $lg(x,y)$ est la somme des longueurs des phrases de x à y, et $poids(n,m)$, le poids défini pour le modèle de traduction n-m.

Le score $S(i,j)$ est le meilleur score de la case $(0,0)$ jusqu'à la case (i,j). La fonction S calcule le minimum des six cas de modèle :

$$S(i,j) = \min \begin{cases} S(i-1,j-1) & + & c(i,j;1,1) \\ S(i-1,j-2) & + & c(i,j;1,2) \\ S(i-2,j-1) & + & c(i,j;2,1) \\ S(i-2,j-2) & + & c(i,j;2,2) \\ S(i-1,j-3) & + & c(i,j;1,3) \\ S(i-3,j-1) & + & c(i,j;3,1) \end{cases}$$

Les ancres correspondant aux phrases déjà alignées par le noyau du système limitant la zone concernée, on inscrit dans les cases correspondant à la zone non concernée la valeur négative -1 sans faire aucun calcul de score. La zone non concernée est définie comme suit :

- si (j_s, f_t) est une ancre, alors
  - tout (j_x, f_y) tel que $x < s$ et $y > t$ appartient à la zone non concernée ;
  - tout (j_x, f_y) tel que $x > s$ et $y < t$ appartient à la zone non concernée ;
- si (j_s, f_t) et (j_{s-1}, f_{t-1}) sont des ancres, alors
  - (j_{s-1}, f_t) appartient à la zone non concernée ;
  - (j_s, f_{t-1}) appartient à la zone non concernée.

Afin de forcer le passage par les ancres, le calcul du score diffère pour les cases situées à côté d'une ancre. Par exemple, si la case (i,j) est une ancre, pour toutes les cases autour telles que $(i+1,j)$, $(i+2,j)$, $(i+3,j)$, $(i,j+1)$, $(i,j+2)$, $(i,j+3)$, $(i+1,j+1)$, $(i+2,j+2)$ et $(i+3,j+3)$, on ne tient compte que de la possibilité permettant de passer par l'ancre. Ainsi, $S(i+1,j)$ vaut $S(i-1,j)+c(i,j;1,0)$[10].

Interface graphique

Une interface graphique a également été réalisée afin de faciliter la vérification ou une éventuelle modification manuelle des résultats.

Elle permet d'afficher les paires d'ensembles de phrases alignées (non seulement les paires 1-1 mais aussi les paires constituées de plus de deux phrases, comme représenté figures 3.9 page suivante et 3.10 page 120), une par une, ainsi que les phrases précédente et suivante dans chacun des deux textes. Il est ainsi plus facile de détecter des résultats erronés.

Elle permet également d'intervenir directement sur les résultats affichés à l'écran : la fonction interactive de modification permet à l'utilisateur de corriger d'éventuelles erreurs avec quelques gestes simples au fur et à mesure de la vérification, et d'enregistrer la version corrigée du résultat d'alignement dans un nouveau fichier au format XML.

[10] Cette méthode qui oblige le passage par les ancres génère parfois un alignement selon un modèle de traduction non pris en compte, du type 1-0 ou 1-4.

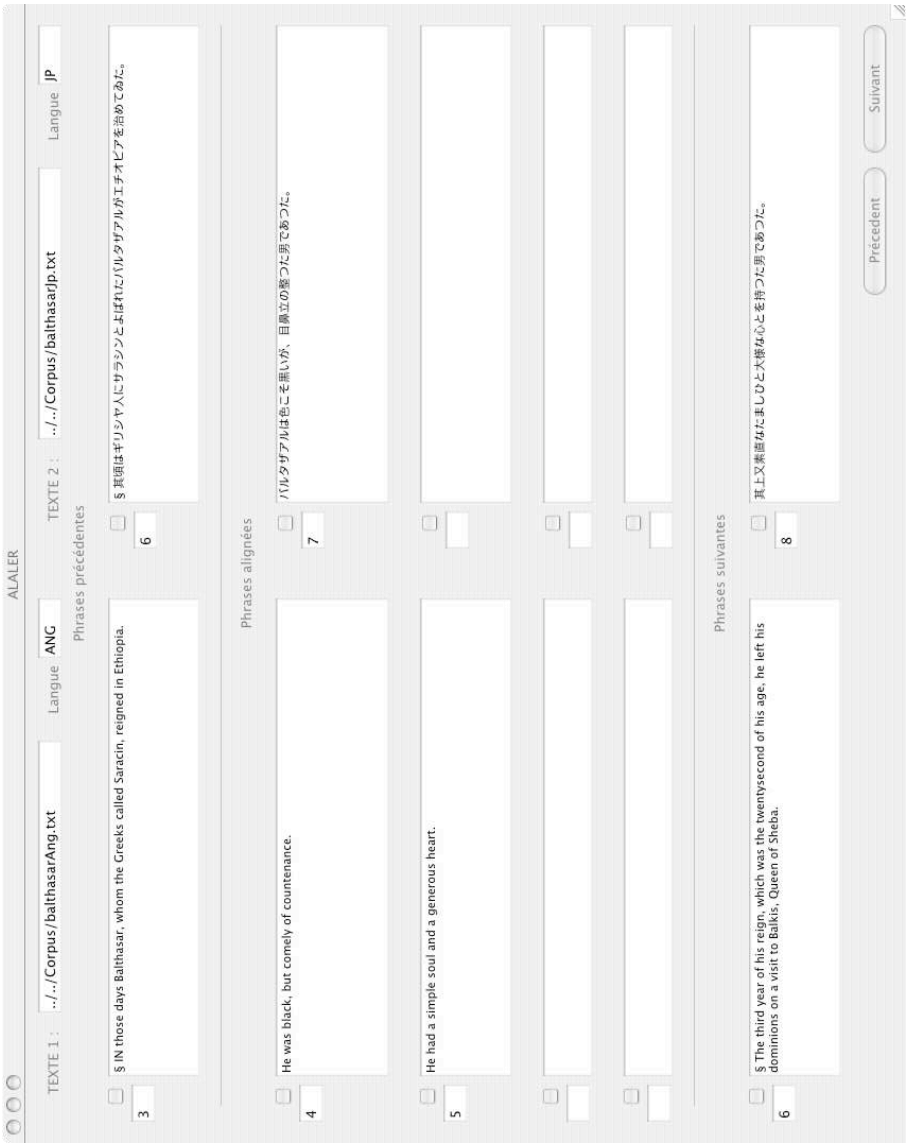


FIG. 3.9 – Interface avec affichage d'un résultat d'appariement de phrases 2-1

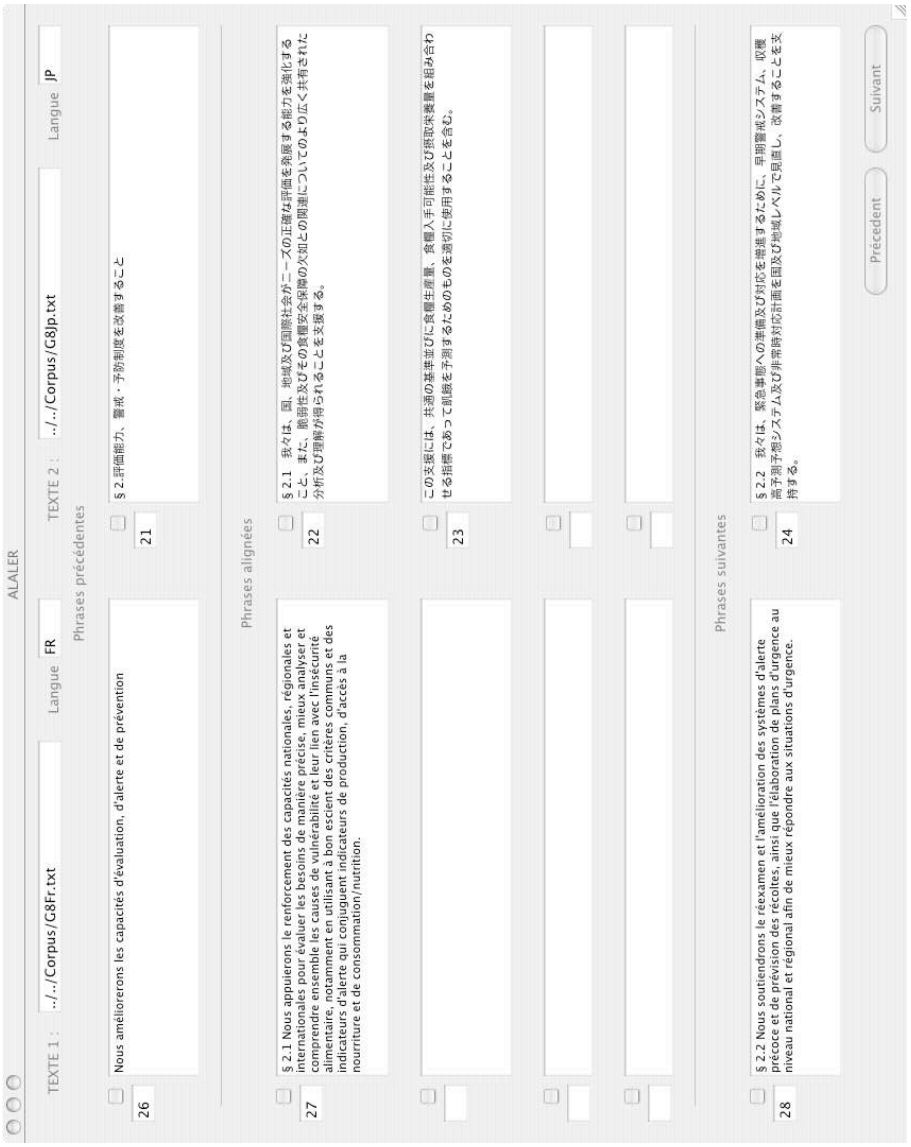


FIG. 3.10 – Interface avec affichage d'un résultat d'appariement de phrases 1-2

3.5 Structure de données optimisée pour les matrices éparses

Comme nous venons de le voir, notre système utilise une méthode d'alignement des phrases basée sur la similarité des distributions lexicales. Il est déjà connu que cette méthode d'alignement est très coûteuse en terme d'utilisation mémoire. Afin d'atténuer cet inconvénient, nous avons conçu une structure de données mettant pleinement à profit le fait que toutes les matrices utilisées dans la méthode sont des matrices éparses.

3.5.1 Matrice utilisée par la méthode

Comme nous l'avons déjà décrit dans la section 2.1 consacrée à l'exposé de la méthode proposée par Kay et Röscheisen, la méthode d'alignement basée sur la similarité des distributions lexicales suppose la diagonalité de l'alignement.

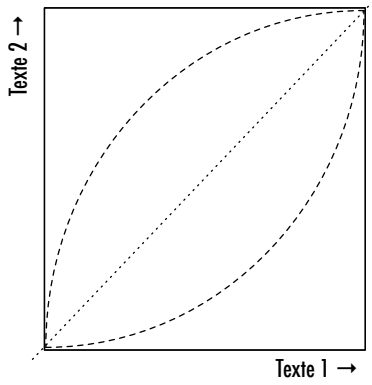


FIG. 3.11 – Matrice représentant la table des paires de phrases susceptibles d'être alignées

Ainsi, dans la matrice représentant la table des paires de phrases susceptibles d'être alignées (table CPR), la zone constituée des cases correspondant aux paires candidates forme, comme le montre la figure 3.11, une ellipse avec pour axe principal la diagonale de la matrice.

La comparaison des mots et leur appariement ainsi que l'alignement des phrases réalisé suite à ces opérations s'appuient tous sur l'hypothèse des paires candidates indiquées par cette table CPR. Si bien que toutes les matrices produites au cours des différents calculs ont toutes moins de cases remplies que cette matrice CPR. Cela signifie que toutes les matrices utilisées dans cette méthode sont des matrices éparses dont les cases non vides se concentrent autour de la diagonale.

3.5.2 Structures de données pour les matrices éparses

Pour faciliter les explications, nous nous limitons d'abord au cas des matrices binaires.

Une des possibilités de structure de données économique pour représenter une matrice éparse binaire est une liste L_M (ou un tableau) contenant n listes L_i, où n est le nombre d'éléments de l'axe x, chaque L_i contenant tous les indices j tels que $(i, j) = 1$. Ainsi, la matrice éparse $M(6 \times 6) = \{(1,1),(2,2),(3,3),(5,5),(6,6)\}$ est représentée comme :

$$
\begin{aligned}
L_M = \{ & \\
& L_1 = \{1\}, \\
& L_2 = \{2\}, \\
& L_3 = \{3\}, \\
& L_4 = \{\}, \\
& L_5 = \{5\}, \\
& L_6 = \{6\} \quad \}
\end{aligned}
$$

Cette structure a comme avantage la rapidité de recherche de tout élément qui vaut 1, mais l'accès à une case donnée revient au parcours de la liste concernée. Par exemple, on doit parcourir tous les éléments de la liste L_i pour savoir que $(i, j) = 0$.

Dans la procédure d'alignement de notre système, l'accès à une case donnée afin de consulter sa valeur est une tâche fréquente, si bien que cette structure n'est pas adaptée à notre réalisation.

Afin de faciliter l'accès direct à une case, nous avons profité du fait que nous connaissions la largeur maximum de la zone constituée des cases à valeur 1. En effet, pour les CPR créées au cours de la procédure principale d'alignement, la largeur maximum est définie comme $\sqrt{i \times 4}$ et pour les CPR créées lors du préalignement, elle est définie comme $\sqrt{i \times 10}$.

La structure conçue pour représenter ces matrices de largeur maximum connue est un tableau à deux dimensions T, la première dimension étant égale au nombre d'éléments de l'axe x et la seconde à la largeur maximum.

Comme le représente la figure 3.12 page ci-contre, chaque $T[i]$ a donc l éléments (l étant la largeur maximum) correspondant aux cases de la matrice initiale de $(i, j_{début})$ à (i, j_{fin}) où $j_{fin} - j_{début} = l$.

Pour obtenir $j_{début}$, on calcule d'abord $j_{diagonale}$ situé sur la diagonale et on y soustrait ensuite la moitié de la largeur :

$$
\begin{aligned}
j_{diagonale} &= i \times |y|/|x| \\
j_{début} &= j_{diagonale} - l/2
\end{aligned}
$$

Ainsi, l'accès à une case donnée est direct avec seulement un léger calcul supplémentaire pour l'obtention de l'indice j correspondant, indépendamment de la position et de la valeur de la case.

De plus, la première structure nécessite une autre couche pour représenter des valeurs autres que 1 et 0 pour les matrices non binaires, tandis que cette dernière –

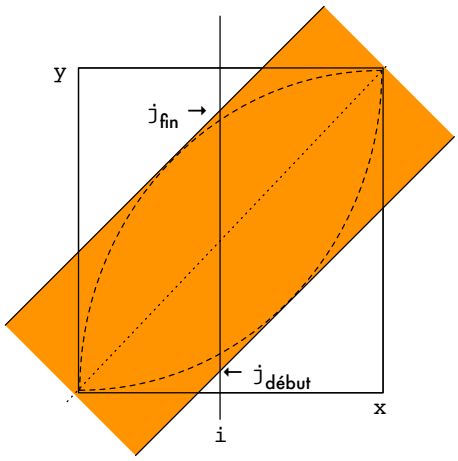


FIG. 3.12 – Matrice éparse de largeur fixe

tableau à deux dimensions représentant une matrice éparse de largeur fixe – peut représenter, telle qu'elle est, les matrices binaires ou non.

3.6 Évaluation des résultats obtenus

Nous avons testé les performances de notre système avec cinq textes parallèles français-japonais et deux anglais-japonais.

La procédure d'alignement du système `AlALeR` est constituée de deux parties, le noyau `AlALeR` et le module de post-alignement. Le noyau `AlALeR` est composé lui-même de deux opérations, le préalignement et la procédure principale. Pour chaque texte, nous avons analysé les résultats de ces trois étapes : le résultat du préalignement, le résultat partiel du noyau `AlALeR` et le résultat complet.

3.6.1 Environnement d'évaluation

- PowerMac G5, 2x2 GHz
- 512 Mo de RAM
- Mac OS X 10.4
- GCC 4.0

3.6.2 Caractéristiques des textes d'entrée

Nous avons utilisé cinq corpus[11] parallèles français-japonais (de 1 à 5) et deux corpus parallèles anglais-japonais (6 et 7) :

[11] Pour le contenu détaillé de chaque corpus, voir la Liste des corpus utilisés (page 547).

1. corpus Bio et BioJP (article sur un sujet scientifique du magazine du Ministère des Affaires Étrangères) ;
2. corpus FIV et FIVJP (article sur un sujet scientifique du magazine du Ministère des Affaires Étrangères) ;
3. corpus G8 et G8JP (texte du sommet G8) ;
4. corpus Unicode et UnicodeJP (page Internet « How to Unicode ») ;
5. corpus Zadig et ZadigJP (*Zadig*, roman de Voltaire) ;
6. corpus EU et EUJP (texte de l'Union européenne) ;
7. corpus Balth (*Balthasar*, roman de Anatole France) ;

Bio, Fiv et G8 sont des textes de petite taille de 1 500 mots et EU, Unicode et Balthasar sont des textes de taille moyenne de 4 à 5 000 mots. Zadig est un texte de taille supérieure à 25 000 mots.

Des informations plus détaillées sur chaque texte sont présentées dans le tableau 3.13. La ligne « Phr » montre le nombre de phrases contenues dans chaque texte et la ligne « M/C », celui de mots (pour les textes français et anglais) ou de caractères (pour les textes japonais).

	Bio		FIV		G8		EU		Unicode		Balth		Zadig	
Lang	Fr	Jp	Fr	Jp	Fr	Jp	Ang	Jp	Fr	Jp	Ang	Jp	Fr	Jp
Phr	69	75	54	52	53	47	252	238	274	268	321	423	1900	2198
M/C	1418	3615	1176	2597	1398	3077	3881	14308	4224	14155	4835	11491	26271	69475

TAB. 3.13 – Caractéristiques des textes

	Modèles de traduction												
	0-1	1-0	1-1	1-2	1-3	1-4+	2-1	2-2	2-3+	3-1	3-2	3-3+	4+ -1
Bio	0	0	55	7	1	0	3	0	0	0	0	0	0
FIV	0	0	43	3	0	0	2	0	0	0	0	0	1
G8	0	0	38	1	0	0	7	0	0	0	0	0	0
EU	0	4	208	5	1	0	17	0	0	0	0	0	0
Unicode	1	0	195	22	1	0	19	2	0	1	1	0	1
Balth	1	2	185	68	16	4	9	13	1	0	0	0	0
Zadig	7	6	1190	300	55	9	103	20	5	18	4	1	3

TAB. 3.14 – Modèles de traduction

Le tableau 3.14 présente la répartition par modèle de traduction de chaque paire de textes. La colonne 1-1 montre le nombre de paires en relation traductionnelle, constituées d'une phrase du premier texte (français ou anglais) et d'une du second texte (japonais), la colonne 1-2 le nombre de paires constituées d'une phrase du texte 1 et de deux phrases du texte 2, et ainsi de suite.

Nous pouvons constater avec la figure 3.15 page suivante que les textes littéraires ont une variation plus importante de leurs modèles de traduction que les autres textes.

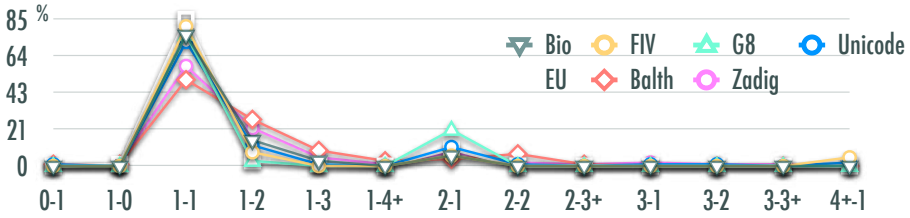


FIG. 3.15 – Répartition par modèle de traduction

Beaucoup d'études ont montré que les modèles complexes (c'est-à-dire ceux qui sont constitués de plusieurs phrases comme 1-3) perturbaient considérablement les systèmes d'alignement basés sur des méthodes probabilistes uniquement, au point de fausser tous les alignements effectués après l'analyse d'un modèle complexe.

3.6.3 Remarques générales

Le tableau 3.16 présente les résultats des trois étapes du système : le résultat de préalignement, le résultat partiel du noyau `AlALeR` et le résultat complet.

		Bio	FIV	G8	Unicode	EU	Balth	Zadig
Préalignement	Rappel	0,57	0,53	0,42	0,62	0,81	0,23	0,14
	Précision	0,98	0,93	1	0,96	0,98	0,99	0,91
Partiel	Rappel	0,81	0,66	0,95	0,87	0,91	0,49	0,66
	Précision	1	1	1	0,98	1	0,96	0,95
Complet	Rappel	0,99	0,94	1	0,96	0,96	0,89	0,86

TAB. 3.16 – Résultats d'alignement

Le très bon résultat de préalignement d'Unicode montre l'efficacité de l'alignement des cognats et des transfuges pour les textes informatiques.

Mais, ce n'est pas le cas pour les textes littéraires. Ce qui est efficace pour ces textes, c'est l'exploitation des retours chariots et des mots en *katakana*.

Le taux de rappel très bas de certains textes est dû au résultat limité du préalignement pour les textes littéraires, et à la présence importante de mots de fréquence faible pour FIV. C'est un point faible des méthodes basées sur la similarité de distribution.

Mais, dans notre système, un appariement final basé sur la corrélation des longueurs a bien compensé cet inconvénient. Cet ensemble de résultats nous permet de dire également que le système supporte assez bien les modèles complexes. Cette robustesse est due au résultat partiel extrêmement fiable.

3.6.4 Analyse des résultats de chaque étape

Chaque étape de traitement est source d'erreurs qui sont répercutées dans le résultat final.

Découpage en phrases

Les découpages erronés n'influent pas directement sur le résultat proprement dit, mais ils augmentent la difficulté d'alignement. De façon générale, l'alignement de phrases une pour une (1-1) est plus facile que une pour deux (1-2) ou deux pour une (2-1).

Lorsque le système reconnaît incorrectement les phrases et qu'il découpe un passage en deux phrases au lieu d'une ou inversement, la possibilité de correspondance croît et le risque d'erreur augmente considérablement.

C'était le cas par exemple avec le texte « Bio ». Le symbole indiquant une note de bas de page précédé directement par un séparateur de phrase, a empêché la segmentation correcte des phrases. Cette mauvaise segmentation a entraîné une perle de phrases du type 1-3, ce qui a multiplié la difficulté d'alignement.

Nous avons défini certaines règles détaillées permettant de traiter correctement des exceptions, mais les cas inattendus subsistent toujours comme nous en avons rencontrés dans le texte « Bio ».

Lemmatisation

La lemmatisation entraînant un regroupement des mots a une influence sur les associations des lemmes (de la table `MAL`) car elle modifie la fréquence et la distribution de ces lemmes, éléments décisifs de la mise en correspondance.

La plupart des lemmatisations erronées proviennent de l'absence de règles plus complexes telles que celles permettant de regrouper les mots « famine » et « faim » (dans G8) ou « gène » et « génétique » (dans FIV). Conséquence : la lemmatisation erronée empêche la mise en correspondance correcte des lemmes.

Ces problèmes pourraient être résolus, en grande partie, par la définition de règles plus détaillées. Mais, l'introduction de règles très complexes propre à une langue peut représenter un obstacle en cas d'adaptation à une nouvelle langue. De plus, l'analyse morphologique n'est pas notre objectif principal et l'influence de ce problème ne semble pas déterminante sur le résultat final. Nous n'avons donc pas cherché une amélioration de cette méthode de lemmatisation dans le cadre de cette thèse.

Quant au japonais, le résultat de la segmentation par recherche des sous-chaînes communes a été extrêmement satisfaisant. Quelques petites erreurs ont été constatées notamment dans le cas des conjonctions dont la première partie est écrite en idéogramme : ce type de conjonction est mal segmenté par la méthode de segmentation par type de caractère – l'idéogramme est rattaché au substantif précédant la conjonction –, ce que la recherche des sous-chaînes communes ne permet pas de corriger à moins que le substantif ne soit reconnu ailleurs.

Cependant, elles n'ont probablement pas d'influence sur le résultat de la mise en correspondance des mots, car une séquence non correctement lemmatisée est généralement une chaîne d'occurrence faible qui n'est de toute façon pas prise en compte lors de l'appariement des mots. Lorsque la séquence a une fréquence suffisamment élevée pour qu'elle soit prise en compte pour l'appariement, le lemme est généralement détecté correctement suite à la recherche des sous-chaînes communes.

Mise en correspondance des lemmes

Le calcul des phrases correspondantes étant basé sur le nombre de mots correspondants qu'elles contiennent, une mauvaise association des lemmes a une influence directe sur le résultat final.

Les mauvaises associations de lemmes proviennent premièrement, comme nous venons de le voir, des lemmatisations incorrectes.

Deuxièmement, elles sont influencées par le contenu de la table `CPR`, hypothèse des paires de phrases à aligner. En effet, la mise en correspondance est réalisée par comparaison des mots appartenant aux phrases supposées être alignées. Or si ces hypothèses sont elles-mêmes fausses, nous comparons des mots sans aucun rapport entre eux et nous obtenons des associations complètement fausses.

Le troisième type de problème est lié à la **polysémie** et à la **synonymie** et il est beaucoup plus difficile à résoudre. Dans un contexte monolingue, ces deux phénomènes illustrent « ce que l'on peut appeler la **non-biunivocité** des rapports entre le plan des formes et le plan des sens » (Fuchs, 1996). Dans un contexte bilingue, ils entraînent souvent un rapport non-biunivoque entre deux unités de langue différente. Or, l'algorithme d'alignement « grossier » des mots que nous employons ne prévoit que le rapport un à un (*one-to-one*) des unités, empêchant l'alignement d'une unité française avec une unité japonaise lorsque cette première a déjà été mise en correspondance avec une autre unité japonaise (ou vice-versa).

Par exemple, le mot japonais 食糧 (*shokuryô*) apparaît (dans « G8 ») aussi bien en tant que traduction de « alimentaire » que de « nourriture ». Cette traduction multiple peut provoquer deux types de conséquences : soit le mot est apparié avec la traduction dont la distribution est la plus proche, entraînant éventuellement une fausse mise en correspondance de l'autre traduction, soit les distributions sont si différentes qu'aucune association n'est réalisée. 食糧 (*shokuryô*) s'est retrouvé dans la première catégorie. Il a été apparié avec « alimentaire », et « nourriture » a été associé avec le mot « disponibilité »[12].

Dans Kitamura & Matsumoto (1997), est présentée une méthode d'appariement de ce type de mots polysémiques. Lorsqu'un mot du texte 1 est apparié avec un mot du texte 2 ayant une fréquence moins élevée, on continue à chercher

[12]Cela s'explique par le fait que 供給 (*kyôkyû*, disponibilité) est toujours employé avec 食糧 (*shokuryô*), dans les phrases où 食糧 (*shokuryô*) est traduit par « nourriture » (e.g. « la nourriture disponible », « la nourriture est disponible », etc.).

une autre correspondance de ce mot du texte 1 en soustrayant de sa fréquence le nombre d'occurrences déjà appariées avec la première traduction. Par exemple, 食糧 (*shokuryô*) de fréquence 31 est d'abord apparié avec « alimentaire » de fréquence 26. Ensuite, on cherche une autre correspondance de 食糧 (*shokuryô*) avec une fréquence de 5 (31 − 26) et on trouve « nourriture » de fréquence 4 dont la distribution est très proche. Toutefois, nous avons estimé qu'il n'était pas nécessaire de traiter aussi finement ce problème. Le calcul, sans doute assez coûteux, semble apporter une précision non indispensable pour notre système.

Un autre type de problème : les mots (ou expressions) composés qui ont comme correspondant dans l'autre langue une seule unité. Certains mots correspondent seulement à une partie d'expression composée ou même à un des morphèmes constituant un mot. Par exemple, le terme japonais 欠如 (*ketsujo*) est apparié avec « insécurité » alors que « insécurité » est traduit non seulement par un mot, mais par un ensemble de mots formant le syntagme nominal 安全保障の欠如 (*anzen hoshô no ketsujo*). Dans le résultat de cette évaluation, la partie non alignée 安全保障 (*anzen hoshô*) n'est alignée avec aucun mot français, mais elle aurait aussi bien pû entraîner une fausse association.

Dans le cas de mots composés, leur détection et leur alignement sont assez simples à réaliser, si chaque mot composant n'est utilisé que dans le même mot composé – c'est-à-dire, par exemple « *categories* » et « *job* » sont utilisés uniquement dans le mot composé « *job categories* » et jamais séparément. Nous avons tout simplement conservé toutes les paires ambiguës – c'est-à-dire celles ayant exactement la même similarité. Ainsi, nous avons réussi à obtenir l'appariement correct de plusieurs mots composés : 職種 (*shokushu*) avec « *job* » et « *categories* », 人的資源 (*jinteki shigen*) avec « *resource* » et « *human* ». Ce choix a entraîné, bien entendu, du bruit. Mais, malgré ce désavantage, cette méthode semble plus intéressante que l'abandon pur et simple de toutes les paires qu'on ne peut pas désambiguïser.

Lemmatisation et appariement des mots en katakana

Le tableau 3.17 page suivante présente le résultat d'extraction et d'alignement des mots en *katakana* : le nombre de mots extraits, le nombre de ceux qui sont appariés et le nombre d'appariements erronés.

Le rappel est la proportion des mots appariés parmi l'ensemble des mots extraits.

La précision est la proportion d'appariements corrects parmi les appariements effectivement réalisés.

La précision est satisfaisante alors que le taux de rappel n'est, à première vue, pas très élevé. Toutefois, lorsqu'on constate que ce sont principalement des noms propres et des néologismes qui ont un fort risque de ne pas figurer dans le dictionnaire, ce taux d'alignement correct de 40 à 50% représente un résultat intéressant.

	Bio	FIV	G8	Unicode	EU	Balth	Zadig
Mots extraits	50	43	21	163	62	34	152
Mots alignés	23	19	10	50	29	17	68
Erreurs	3	1	1	2	1	0	2
Rappel	0,46	0,44	0,48	0,31	0,47	0,5	0,43
Précision	0,87	0,95	0,9	0,96	0,97	1	0,97

TAB. 3.17 – Résultats d'alignement des mots en *katakana*

3.6.5 Comparaison des résultats avec et sans analyse morphologique

Nous avons également réalisé l'alignement de notre corpus en remplaçant notre fonction de segmentation et de lemmatisation du texte japonais – l'analyse du texte français/anglais étant toujours réalisée par notre fonction – par une analyse morphologique réalisée à l'aide d'un analyseur existant largement utilisé au Japon, `ChaSen`(Matsumoto et al., 2002).

Nous avons tout d'abord utilisé le résultat d'analyse morphologique de `ChaSen` sans aucun traitement des mots grammaticaux des deux textes d'entrée. Le résultat d'alignement des mots était extrêmement mauvais à cause du bruit dû aux mots grammaticaux. Les notions de mots grammaticaux entre le japonais et le français (ou l'anglais) sont, comme nous l'avions imaginé, trop différentes pour que ces mots puissent être alignés de manière automatique à l'aide uniquement de leur similarité de distribution. À titre d'exemple, la table `MAL` du corpus « G8 » avec `AlALeR` pur contient 37% de résultat erroné alors que celui avec `ChaSen` sans traitement des mots grammaticaux en contient 66%.

Nous avons ensuite testé avec suppression des mots grammaticaux. Nous n'avons conservé que les noms autonomes (de 1 à 19 et 40 selon le code de catégorie morpho-lexicale de `ChaSen`), les verbes autonomes (46 et 47), les qualificatifs autonomes (50 et 51). La table `MAL` du corpus « G8 » présente cette fois 47% d'appariements erronés. L'augmentation de 10% par rapport au résultat d'`AlALeR` pur est due à la difficulté d'appariement des verbes.

Mais la différence la plus intéressante entre le résultat d'`AlALeR` pur et celui obtenu avec l'utilisation de `ChaSen` réside dans la lemmatisation et par conséquent l'appariement des mots en *katakana*.

Le tableau 3.18 (voir page suivante) montre les résultats d'extraction et d'alignement des mots en *katakana* de ces deux configurations. La colonne de gauche de chaque texte est le résultat d'`AlALeR` et celle de droite, le résultat obtenu avec `ChaSen`. La dernière ligne est le produit du rappel et de la précision qui représente la proportion des mots correctement alignés parmi l'ensemble des mots extraits.

Pour tous les textes (sauf « Unicode »), `ChaSen` a extrait plus de mots en *katakana* que `AlALeR`. En effet, l'analyseur a sursegmenté plusieurs mots en *katakana* qui étaient absents du dictionnaire. Par exemple, カナナスキス (*kananasukisu*, kananaskis) est segmenté en trois mots, 仮名 (*kana*, syllabaires japonais), 茄子

	Bio		FIV		G8		Unicode		EU		Balth		Zadig	
ChaSen	-	√	-	√	-	√	-	√	-	√	-	√	-	√
Mots extraits	50	57	43	45	21	23	163	162	62	63	34	37	152	166
Mots alignés	23	24	19	14	10	8	50	44	29	30	17	17	68	62
Erreurs	3	4	1	0	1	1	2	2	1	1	0	1	2	2
Rap. × Pré.	0,4	0,35	0,42	0,31	0,43	0,31	0,30	0,26	0,45	0,46	0,5	0,43	0,43	0,36

TAB. 3.18 – Résultats d'alignement des mots en *katakana* II

(*nasu*, aubergine) et キス (*kisu*, « kiss » ang.), empêchant bien évidemment l'appariement avec le mot apparaissant dans le texte français « Kananaskis ». Tous les mots alignés par AlALeR mais non pas par ChaSen sont sursegmentés par ce dernier. En revanche, le cas de sursegmentation par AlALeR se limitait à 1. Pour les mots en *katakana* qui sont principalement des noms propres et des néologismes – mots souvent absents des dictionnaires –, la segmentation basée sur la comparaison entre les mots présents dans le même texte se montre plus efficace que la méthode s'appuyant sur la consultation d'un dictionnaire.

Cet ensemble de différence n'a cependant pas de grande influence sur le résultat final de l'alignement.

Le tableau 3.19 présente les résultats des trois étapes obtenus avec l'utilisation de ChaSen pour l'analyse morphologique des textes japonais.

		Bio	FIV	G8	Unicode	EU	Balth	Zadig
Préalignement	Rappel	0,61	0,51	0,54	0,58	0,70	0,13	0,14
	Précision	0,95	0,96	1	0,96	0,83	0,97	0,91
Partiel	Rappel	0,67	0,61	0,87	0,87	0,83	0,48	0,66
	Précision	1	0,97	1	0,99	0,91	0,97	0,95
Complet	Rappel	0,99	0,96	1	0,96	0,85	0,87	0,87

TAB. 3.19 – Résultats d'alignement avec analyse morphologique par ChaSen

La divergence minime de la plupart des résultats ne permet de parler d'aucune influence directe de l'utilisation de l'analyseur : en effet, les désambiguïsations permettant d'obtenir une haute fiabilité entraînent parfois la suppression des paires de phrases correctement alignées et le résultat d'alignement des phrases ne reflète pas forcément, de manière absolue, les résultats d'alignement des mots.

Seule la différence des résultats de « Unicode » est considérée comme significative avec une divergence supérieure à 10%. Mais l'influence du très bon résultat de l'alignement des mots en *katakana* n'est que partielle : c'est surtout dû au traitement des cognats et des transfuges pour lesquels ChaSen ne possède pas de règles permettant de les traiter efficacement. Cependant, la définition de ce type de règle semblant assez facile à réaliser, les résultats de « Unicode » ne permettent pas de prouver une meilleure performance de la méthode de segmentation de

notre système. Seulement, cet ensemble de résultats montre que notre méthode sans analyseur est au moins aussi efficace qu'une méthode utilisant un analyseur morphologique.

3.6.6 Réflexions sur l'utilisation mémoire et le temps de calcul

Comme nous l'avons déjà mentionné dans les remarques générales, l'impossibilité de mise en correspondance des mots de fréquence faible est un point faible des méthodes basées sur la similarité de distribution. Mais cet inconvénient a été bien compensé par un appariement final basé sur la corrélation des longueurs dans notre système.

Le point faible le plus conséquent de cet algorithme est l'utilisation importante de mémoire. Nous avons donc implémenté une structure de données qui profite du fait que toutes les matrices sont des matrices éparses.

Nous avons comparé l'utilisation mémoire et le temps de calcul d'implémentations utilisant chacune une structure de données différente.

		G8	Unicode	Balth	Zadig
Nb de mots		1 398	3 881	4 835	26 271
Mém. réelle (Mo)	Tableau	3	210	72	250
	STL	4	7	7	32
	Matrice éparse	4	11	8	45
Mém. virtuelle (Mo)	Tableau	31	235	76	1,14 Go
	STL	34	35	37	70
	Matrice éparse	33	39	37	80
Temps de calcul (sec.)	Tableau	2	8	16	1 196
	STL	9	53	226	260 151
	Matrice éparse	2	6	15	984

TAB. 3.20 – Utilisation mémoire et temps de calcul

Le tableau 3.20 montre la comparaison de l'utilisation des mémoires réelle et virtuelle ainsi que le temps de calcul selon la structure de données utilisée : tableau à deux dimensions, liste de paires du type STL (*Standard Template Library*[13]) et structure optimisée pour les matrices éparses.

Dans une implémentation avec des tableaux à deux dimensions, le système utilisait, comme le montre la figure 3.21 (voir page suivante), près de 300 Mo de mémoire réelle et 1 Go de mémoire virtuelle pour un extrait de Zadig de 18 000 mots et il a été impossible de réaliser un alignement de l'intégralité de Zadig.

Avec des listes de paires du type STL comme structure de données, l'utilisation mémoire était considérablement réduite, mais le temps de calcul a augmenté d'un facteur vingt.

Néanmoins, la structure de données optimisée pour les matrices éparses que nous avons conçue spécifiquement a permis finalement la réalisation d'un ali-

[13] Il s'agit d'une librairie standard du langage C++.

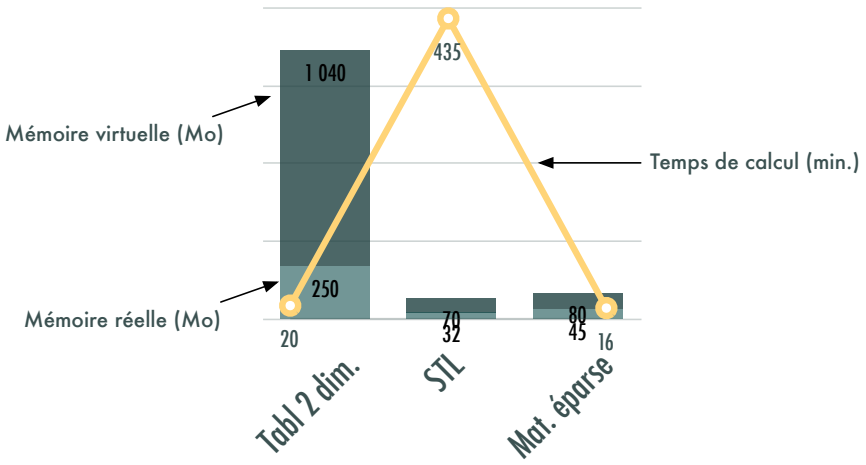


FIG. 3.21 – Alignement d'un extrait de Zadig de 18 000 mots

gnement plus rapide qu'avec les tableaux à deux dimensions, avec une utilisation mémoire beaucoup plus réduite. Mais le temps de calcul reste quand même important, à savoir 1 heure pour Zadig.

3.7 Conclusion

Les résultats d'alignement fournis par notre système AlALeR ont montré la possibilité de conception d'un aligneur traitant les textes japonais qui ne recourt à aucun dictionnaire ni analyseur morphologique. Ce résultat est d'abord dû à la stratégie d'appariement des mots japonais en *katakana*. Ceux-ci étant très nombreux dans les textes traduits, la retranscription des mots japonais en *katakana* pour trouver leur mot d'origine a été d'autant plus efficace qu'ils sont souvent absents des dictionnaires. En effet, ce sont très souvent des néologismes ou des noms propres. Cette stratégie s'est montrée extrêmement robuste, ce que nous n'aurions pas pu constater si nous avions dépendu d'un dictionnaire.

Nous avons également testé le système avec quelques traductions de brevets techniques et nous avons obtenu de très bons résultats grâce à la présence très importante de transfuges. Néanmoins, les phrases de ce type de document sont si longues que l'alignement au niveau phrastique ressemble plutôt à un alignement de paragraphes. Comme Simard le fait remarquer dans Simard (2003), l'alignement à un niveau sous-phrastique est plus bénéfique que celui réalisé au niveau phrastique, notamment en vue de la constitution de mémoires de traduction. Nous aborderons dans les parties qui suivent les travaux dédiés à la réalisation d'un système d'alignement automatique des propositions, qui permettra

très certainement de fournir une base de données plus intéressante, aussi bien pour la conception de mémoires de traduction que pour les études de linguistique contrastive.

Deuxième partie

La notion de proposition : études linguistiques

PLAN DE LA PARTIE

コトバの問題は形式一点張りで片づくものではないことは言うまでもないが、しかしこれまでの日本文法はあまりにも日本語自身の形式を軽視してきた。当然の結果として何物をも発見していない。
（三上章『現代語法新説』1955）

Il est évident que les questions de langue ne peuvent pas être résolues en tenant compte uniquement d'indices formels, mais les grammaires japonaises, jusqu'ici, méconnaissaient trop les formes. Par conséquent, nous n'avons bien entendu rien découvert.
(MIKAMI, Akira. *Gendaigohô-shinsetsu*, 1955)

La présente partie est consacrée aux études linguistiques réalisées dans le but de saisir la notion de proposition, que nous souhaitons aligner dans des textes écrits en français et en japonais.

Le premier chapitre est dédié à l'examen de la notion de proposition en français (**ch. 4**). Travail à caractère appliqué, notre thèse ne propose pas de réflexions approfondies sur des problèmes fondamentaux de linguistique française, mais elle s'appuie sur des théories proposées par des linguistes, fondées sur de nombreuses années de recherche. Nous nous appuyons plus particulièrement sur les travaux de Le Goffic (1993a) du fait de l'importance qu'il a accordée à la syntaxe et surtout aux indices formels, importance rendant ses recherches très utiles aux travaux informatiques du TAL liés principalement au champ syntaxique, tels que les nôtres.

Les études linguistiques sur le japonais comportent trois chapitres portant sur : les notions préliminaires (**ch. 5**), la phrase japonaise (**ch. 6**), et la phrase complexe (**ch. 7**).

Ces travaux sur le japonais se caractérisent également par l'importance accordée à la forme, due à leur nature appliquée. En effet, comme le signale le passage de Mikami cité ci-dessus, les études linguistiques du japonais mettent généralement, encore aujourd'hui, l'accent sur le sens, au mépris, souvent, des formes. Mais une réalisation informatique nécessite une théorie systématique et cohérente qui profite le plus possible des indices formels, seuls éléments que la machine peut manipuler correctement.

Avant d'entrer dans les études, sont présentées quelques conventions sur la notation des exemples japonais.

CONVENTIONS SUR LA NOTATION DES EXEMPLES JAPONAIS

Représentation des exemples

これ　は　一例　です。 – (A)
(*kore - wa - ichirei - desu*) – (B)
(ce - [thème] - un exemple - [copule]) – (C)
« C'est un exemple » (corpus XXX) – (D)

Les exemples de phrases japonaises sont constitués de leur représentation en écriture japonaise (indexée A ci-dessus), de leur romanisation (en italique entre parenthèses, B), de leur traduction mot à mot (entre parenthèses, C) et de leur traduction complète (entre guillemets, D), suivie éventuellement de la source (entre parenthèses, D).

Segmentation des phrases japonaises

Pour faciliter la compréhension, les phrases japonaises sont segmentées de façon très grossière, ne correspondant pas toujours à une segmentation en mots – sauf dans le cas où une segmentation détaillée est jugée nécessaire pour une compréhension correcte.

Romanisation des phrases japonaises

Les exemples de phrases japonaises sont transcrits en alphabet latin selon le système Hepburn. Une seule exception : nous transcrivons le caractère を, non par « *o* » défini dans ce système conformément à sa prononciation, mais par « *wo* » afin de refléter la distinction entre お et を (prononcés tous les deux « *o* ») et du fait d'une lisibilité jugée meilleure.

Traduction mot à mot

La traduction d'un mot est le sens jugé le plus adéquat dans le contexte de la phrase où il apparaît.

Pour la traduction des mots grammaticaux, le sens grammatical est présenté entre crochets. Ainsi, dans l'exemple précédent, le mot *wa* traduit par [thème]

n'est pas le mot japonais équivalent du mot « thème » en français, mais cela signifie qu'il est indicateur de thème. De même, *desu* ne désigne pas le mot « copule », mais il s'agit de la copule japonaise.

Traductions d'extraits

Les traductions françaises de passages d'ouvrages japonais sont de nous, sauf lorsque spécifié explicitement.

Source des exemples

Lorsqu'un exemple cité est extrait d'un ouvrage, la source est marquée à la fin de l'exemple. Les informations précises sur les corpus sont regroupées et présentées à la fin de notre thèse, avant la bibliographie (cf. Liste des corpus utilisés lors des études linguistiques p. 550 et suivantes).

Noms propres japonais

En respectant l'usage japonais, nous donnons, pour tous les noms propres des linguistes japonais cités, en premier le nom de famille (en majuscules) et en second le prénom (en minuscules) :

NAKAMURA$_{\text{(NOM DE FAMILLE)}}$ Yayoi$_{\text{(PRÉNOM)}}$

Lorsque nous citons seulement le nom de famille, nous l'écrivons en minuscules, comme nous le faisons dans le reste de la thèse : ex. Nakamura.

CHAPITRE 4

ÉTUDE DE LA PROPOSITION EN FRANÇAIS

Nous nous intéressons dans ce chapitre à la notion de proposition en français. Avant d'entrer dans l'exposé de nos travaux, nous allons tout d'abord passer en revue très brièvement quelques notions préliminaires (§ 4.1) afin de définir l'ensemble de la terminologie de base que nous utiliserons tout au long de la présente thèse. Nous présenterons également le contexte des études (§ 4.2) afin de montrer les contraintes particulières à nos travaux.

La discussion sera ensuite ouverte avec la définition de la proposition (§ 4.3), suivie des études sur les sous-classes des propositions et les éléments externes (§ 4.4). Nous aborderons ensuite la classe la plus importante : les propositions subordonnées. Nous examinerons d'abord les travaux existants sur cette sous-classe de propositions (§ 4.5) avant de présenter nos propres typologies des subordonnées (§ 4.6) et des connecteurs (§ 4.7), conçues afin de résoudre la problématique des approches critiquées. Enfin, la dernière partie sera consacrée à la discussion des problèmes plus généraux de la détection des propositions (§ 4.8).

4.1 Notions préliminaires : éléments de la phrase française

La phrase française[1] est constituée d'un sujet, d'un verbe et éventuellement d'un ou plusieurs compléments. Il existe deux types de compléments : l'un assurant une **fonction primaire** et l'autre, une **fonction secondaire**. Les fonctions primaires se situent au niveau de la phrase, tandis que les fonctions secondaires se situent au niveau des constituants de la phrase.

[1] Nous empruntons essentiellement l'analyse syntaxique de la phrase française à Le Goffic (1993a).

Les compléments du plan primaire se divisent eux-mêmes en deux classes : **compléments essentiels** et **compléments accessoires**. On distingue encore les compléments accessoires **intra-prédicatifs** des compléments accessoires **extra-prédicatifs**. Les premiers sont rattachés au verbe et le spécifient sous un rapport, alors que les seconds ne font pas partie du prédicat.

Différents types de compléments extra-prédicatifs existent : thème, invocation directe du destinataire, circonstants qui portent sur la phrase dans son ensemble, éléments qui organisent le discours, etc. Leur extériorité est marquée par le détachement – une pause à l'oral et une ponctuation à l'écrit –, mais la place des circonstants a également un rapport étroit avec leur portée. Le début de phrase est le lieu privilégié pour les compléments de phrase, extra-prédicatifs. Les circonstants en début de phrase sont « *a priori* en rapport avec le reste de la phrase dans son ensemble et presque sur un pied d'égalité » (Le Goffic, 1993a, p. 460).

Le tableau 4.1 reproduit de Le Goffic (*Ibid.*, p. 13) présente les schémas de la phrase française[2] avec l'ensemble des éléments constituants du niveau primaire.

Phrase française

Sujet	Prédicat			Éléments extra-préd.
Sujet	Verbe	Compléments essentiels	Compl. access. intra-préd.	Compl. access. extra-préd.

TAB. 4.1 – Structure de la phrase française

4.2 Contexte de l'étude : détection des propositions en vue de l'alignement

Le but des études linguistiques présentées dans ce chapitre est de définir une grammaire pour la détection des propositions permettant, non seulement de reconnaître les frontières de ces unités, mais aussi d'analyser leurs relations syntaxiques. En effet, bien que la détection des propositions vise généralement une simple reconnaissance de leurs frontières, nous envisageons également leur mise en relation car nous avons posé comme hypothèse qu'elle serait utile pour l'alignement de ces unités, du fait de la différence importante de structure des phrases française et japonaise.

Nous utilisons comme entrée du système les résultats de moyens extérieurs, le *tagger* et le *chunker* développés à Paris 7. Un *tagger* attribue aux *tokens* des étiquettes de catégorie morpho-syntaxique et un *chunker* réalise à partir d'un résultat de *tagger* un *chunking*, c'est-à-dire le regroupement d'un certain nombre de

[2]Ce tableau ne représente pas l'ordre effectif de la réalisation linéaire des phrases. Ainsi, les compléments accessoires intra- et extra-prédicatifs, ayant de nombreuses possibilités de positionnement, sont par exemple regroupés en tête de phrase.

tokens de manière à constituer des syntagmes, dits *chunks*. Cinq types de *chunks* sont définis : adverbiaux, adjectifs, nominaux, prépositionnels et verbaux. Les terminaux de notre grammaire seront donc ces cinq catégories de *chunks* et éventuellement les catégories attribuées à des *tokens* par le *tagger*, qui n'ont pas été traités par le *chunker*.

Notre défi est donc de définir une grammaire, non pas très précise avec calcul d'informations diverses qui donnerait différentes possibilités d'analyse pour une seule phrase, mais une grammaire très simple avec des informations disponibles restreintes, mais efficace et opérationnelle.

4.3 Qu'est-ce qu'une proposition ?

Comme nous l'avons précisé dans l'introduction, le choix de la proposition comme unité principale de traitement a nécessité que nous nous mettions tout d'abord à la recherche d'une définition de la proposition. Mais, ce que nous avons découvert en commençant cette recherche est assez problématique : la notion de proposition est employée dans différents domaines, et sa définition peut varier même à l'intérieur d'un même domaine.

Nous allons donc maintenant étudier différentes définitions de la proposition, afin de trouver la plus adaptée à nos travaux.

4.3.1 Sens logique

De l'Antiquité à la fin du XVIII^{ème} siècle, la proposition – plus que la phrase ou l'énoncé – était la catégorie principalement utilisée dans les travaux sur le langage (Léon, 2003).

Dans la tradition aristotélicienne, la proposition désignait l'unité permettant dans le langage d'exprimer des jugements – deuxième des trois activités de l'esprit, la première étant la conception exprimée par des « termes généraux » et la troisième, le raisonnement exprimé par des suites de propositions (Gochet & Gribomont, 1990).

Aristote définit le concept de proposition comme suit :

> « [...] tout discours n'est pas une proposition, mais seulement le discours dans lequel réside le vrai ou le faux, ce qui n'arrive pas dans tous les cas : ainsi la prière est un discours, mais elle n'est ni vraie ni fausse » (Aristote, *De l'Interprétation*).

La proposition de cette théorie est constituée de trois éléments : deux « termes généraux » reliés par le verbe copule « est » ou « n'est pas ».

4.3.2 Du sens logique au sens linguistique

Cette notion logique prend un caractère grammatical à partir du XVII^{ème} siècle[3].

Les Messieurs de Port-Royal élaborèrent la proposition comme notion grammaticale, permettant ainsi le développement de la syntaxe à une époque où les études des grammairiens se concentraient au niveau du syntagme et ne se préoccupaient pas de la proposition. « L'étude du sens et des relations logiques prévaut sur celle des formes. À la base de toute construction grammaticale, on trouve la proposition, constituée du sujet, du prédicat et de la copule qui sera la pierre de touche de la syntaxe à partir de Port-Royal », explique Léon (2003).

Cependant, ils n'ont pas réussi à dégager complètement la proposition de son contexte logique, et ce sont des Encyclopédistes, Du Marsais et Beauzée, qui ont réalisé le passage de la notion de proposition de la logique à la grammaire.

Dans l'article « construction » de L'Encyclopédie, Du Marsais distingue proposition logique et proposition grammaticale :

> « Quand on considère une proposition grammaticalement, on n'a égard qu'aux rapports réciproques qui sont entre les mots ; au lieu que dans la proposition logique on n'a égard qu'au sens total qui résulte de l'assemblage des mots. »

Il sépare également la proposition du jugement, en la définissant comme un assemblage de mots qui a un sens défini et exprime un jugement, par opposition au jugement, défini comme l'acte même de penser quelque chose à propos d'une chose.

Il analyse chaque proposition logiquement en un sujet et un attribut, mais il distingue également les propositions en principales et incidentes (i.e. relatives et complétives).

Dans l'article « proposition » de L'Encyclopédie publiée en 1765, rédigé par Beauzée, la copule a été supprimée et la proposition est devenue bipartite, accordant ainsi beaucoup plus d'importance au verbe.

C'est ainsi que s'est réalisé le passage de la proposition logique à la proposition grammaticale.

4.3.3 Sens psycholinguistique

Les psychologues et les psycholinguistes parlent également de proposition sémantique ou simplement de proposition.

Gineste (2003) présente la définition de la proposition sous un sens psychologique, empruntée de Le Ny (1987) :

> « La proposition, "est définie, d'un point de vue logique, comme la plus petite unité de discours à laquelle puisse s'appliquer une valeur

[3] Ces études s'appuient essentiellement sur les travaux de Léon (2003) présentant un panorama historique des trois notions phrase, proposition et énoncé.

> de vérité, vrai ou faux. D'un point de vue psychologique, cette définition se transforme en : la plus petite sémantique intégrée susceptible d'être traitée et mémorisée." »

Selon cette théorie, les connaissances, que ce soient celles du monde ou celles linguistiques – lexicales ou syntaxiques –, sont représentées mentalement sous le format de base qu'est la proposition, constituée d'un prédicat et d'un argument. À la réception d'une phrase, par exemple « un rossignol chante »[4], le lecteur ou auditeur compose une unité sémantique représentée également par une proposition, en l'occurrence CHANTE(rossignol).

La proposition est donc en ce sens « l'unité de base de la structuration des connaissances dans la mémoire et de leur élaboration » sans laquelle « un système de représentations sémantiques ne pourrait pas s'ériger » (Gineste, *ibid.*).

4.3.4 Proposition dans la linguistique contemporaine

La proposition entrée dans les notions grammaticales comme nous venons de le voir précédemment, reste cependant toujours une question élémentaire pour les linguistes qui, d'après Tesnière (1988), essayent avec cette notion « de faire de la lumière sur la notion de phrase ». Ce qui a entraîné diverses définitions dans le milieu de la linguistique actuel. Tesnière (*ibid.*) qualifie cette tentative de « malheureuse » en citant O. Bloch : « les auteurs ne sont même pas d'accord sur ce qu'il faut entendre par le terme de proposition. »

Sens syntaxique, sémantique et tendance au refus de la notion

Selon l'article « proposition » du dictionnaire de linguistique compilé par Dubois et al. (1994), il existe deux types de sens : sémantique et syntaxique.

Selon la définition sémantique, « il y a proposition toutes les fois qu'il y a énonciation d'un jugement », mais cette définition constitue une sorte de retour vers le sens original logique.

Dans le sens syntaxique, c'est « une unité syntaxique élémentaire constituée d'un sujet et d'un prédicat ». Mais à l'intérieur même des définitions syntaxiques, il est possible de distinguer deux sortes de définitions : l'une reconnaissant les propositions aussi bien à un mode personnel qu'impersonnel (cf. propositions infinitives et participiales) telle que celle de Bescherelle (Hatier, 1990) ; l'autre n'admettant celles à un mode impersonnel que sous certaines conditions plus ou moins strictes (Riegel et al., 1994 ; Wagner & Pinchon, 1991 ; Grevisse, 1993 ; Le Goffic, 1993a).

Par ailleurs, comme Tesnière (1988) qui rejette le recours à la notion de proposition en préférant utiliser l'unité qu'il appelle nœud, unité syntaxique intermédiaire inférieure à la phrase et supérieure au mot, certains linguistes syntacticiens (Blanche-Benveniste et al., 1990 ; Gardes-Tamine, 2003) rejettent cette no-

[4] Exemple tiré de Gineste (*ibid.*).

tion même de proposition trop empreinte de son origine logique et proposent une solution alternative.

4.3.5 Notre choix pour l'alignement automatique

Conditions sur le choix d'une définition

Le but de la réalisation de notre système d'alignement est la constitution automatique de bases de données regroupant des textes parallèles écrits dans deux langues différentes. Dans ces bases de données figure l'indication de la correspondance des éléments de chacun des deux textes, permettant ainsi l'utilisation de ces éléments correspondants alignés comme des exemples de traduction ou des données d'analyse linguistique comparative.

Dans ce cadre, deux conditions s'imposent dans le choix de l'unité à aligner – ou le choix de sa définition. Premièrement, les unités à aligner doivent être détectables de manière automatique, c'est-à-dire qu'elles doivent posséder une indication physique (ou graphique) de leur délimitation. Deuxièmement, il doit y avoir une équivalence entre ces unités dans les deux langues à traiter, sachant que plus cette équivalence est grande, meilleurs sont les résultats. Nous allons donc maintenant passer en revue chacune des définitions que nous venons de voir, en considérant ces deux critères afin de choisir la meilleure solution pour notre opération d'alignement et ses applications.

Examens pour chaque type de définition

Pour utiliser une des définitions logique, psychologique ou sémantique, dans le cadre du traitement automatique, nous sommes à nouveau confrontés à la définition formelle des autres unités. Qu'est-ce qu'un jugement ? Sous quelle forme, s'il en existe une, est-il réalisé sur le texte effectivement produit ? À quoi correspond sur le plan formel l'unité que Le Ny appelle « la plus petite sémantique intégrée susceptible d'être traitée et mémorisée » ? Existe-il des moyens graphiques ou des catégories de mots permettant de repérer ces unités ? Il serait sans doute idéal, s'il était possible de simuler le fonctionnement de notre cerveau, d'utiliser les unités correspondant à celles utilisées lors du traitement mental, mais il se trouve que les avancées des recherches dans ce domaine ne nous le permettent pas encore.

Pour cette définition de Le Ny, nous pouvons faire un rapprochement avec le nœud de Tesnière, et peut-être aussi l'unité de syntaxe de Blanche-Benveniste. Quand on considère l'analyse prédicative réalisée par Le Ny (1979), on constate une certaine correspondance entre les propositions présentées par Le Ny et les nœuds de différents niveaux dans l'arbre de dépendance de Tesnière. Cette notion est sans doute intéressante à étudier de façon plus poussée, mais pour l'alignement, nous préférons une unité plus large. En effet, comme le signale Halliday (1962), « plus on s'approche de la phrase, plus la probabilité d'équivalence devient grande ».

Le recours à la définition « formelle » logique (« la proposition est constituée d'un sujet et d'un prédicat ») nous fait finalement retomber sur une unité assez proche de la définition au sens syntaxique.

Il résulte de ce que nous venons de voir que le choix le plus judicieux est de retenir comme définition de la proposition celle au sens syntaxique. Nous adopterons plus particulièrement celle de Le Goffic qui la définit par le repérage d'un sujet et d'un prédicat, et qui définit une classe syntaxique complètement distincte pour les groupes infinitival et participial. En effet, nous ne pouvons retenir ces syntagmes à verbe infinitif ou participial, car nous devrions alors faire face à un autre problème difficile : celui de la délimitation entre les formes verbales participiales et les adjectifs, sans laquelle la proposition s'élargirait et désignerait des syntagmes correspondant à l'ensemble nodal de Tesnière.

4.4 Sous-classes des propositions et éléments externes

Nous avons défini dans la section précédente la proposition. Mais, celle-ci est généralement encore catégorisée en sous-classes. Quelles sont les différentes propositions que nous devons reconnaître ? Existe-il des unités autres que les propositions, qui leur sont extérieures sur le plan syntaxique et dont la détection serait favorable, afin de les en séparer ?

Nous allons tout d'abord étudier dans cette section différents classements des propositions existants (§ 4.4.1) pour en déterminer les types à détecter (§ 4.4.2). Nous aborderons ensuite les unités extérieures aux propositions (§ 4.4.3) que sont les éléments extra-prédicatifs, que nous séparerons des propositions détectées.

4.4.1 Différentes typologies proposées : un état de l'art

Dans cette section, nous étudions différentes typologies de la proposition.

La plupart des grammaires (Grevisse, 1969 ; Chevalier et al., 1964 ; Wagner & Pinchon, 1991) définissent – plus ou moins – quatre types de propositions :

1. juxtaposée ;
2. coordonnée ;
3. subordonnée ;
4. incidente (ou incise).

Gardes-Tamine (1998) y ajoute corrélative pour « il pleuvait si fort que Jean ne sortit pas » généralement incluse dans la subordination.

Le Goffic (1993a) – qui appelle les propositions parfois sous-phrases – définit un peu différemment les classes et divise d'abord les propositions en deux grands types : avec ou sans connecteur. Le second type est ensuite lui-même classé en trois types : incises et incidentes, constructions paratactiques[5], ainsi que

[5]Le Goffic utilise le terme parataxe pour désigner les stades intermédiaires entre la subordination et l'indépendance syntaxique de deux phrases.

constructions dégradées au niveau du verbe. Il parle non plus de proposition coordonnée, mais simplement de la possibilité de coordination de deux phrases indépendantes (*Ibid.*, p. 501) :

> « Deux phrases indépendantes peuvent être coordonnées, c'est-à-dire reliées tout en restant sur un pied d'égalité. »

Problème lié à la notion de proposition principale

Certains comme Grevisse (1969), Chevalier et al. (1964) ou encore Riegel et al. (1994) définissent la proposition principale, mais beaucoup de linguistes (Wagner & Pinchon, 1991 ; Wilmet, 1997 ; Gardes-Tamine, 1998 ; Le Goffic, 1993a ; Delaveau, 2001) rejettent cette notion de « proposition principale » du fait de l'inexactitude de l'analyse linéaire appelée traditionnellement « analyse logique ». Le Goffic (*Ibid.*, p. 43) explique :

> « Le fait de parler de "proposition subordonnée" suppose un verbe principal mais n'entraîne pas l'existence d'une "proposition principale", qui se trouverait réduite au seul verbe dans *Que Paul ait gagné montre qu'il était le plus fort*. Les propositions sont emboîtées hiérarchiquement, et non juxtaposées. »

4.4.2 Notre définition des propositions

La finalité de nos travaux étant une application en traitement automatique, nous avons choisi de nous appuyer sur des critères uniquement formels. À cet effet, nous nous basons sur le type de connecteur et la position d'apparition dans la phrase, pour distinguer quatre types de sous-structures de phrases munies elles-mêmes d'un sujet et d'un prédicat. Ces quatre types correspondent chacun à une catégorie de proposition que nous étudions dans cette section : **racine**, **subordonnée**, **coordonnée** et **détachée-insérée**.

Proposition racine

Toute phrase comprenant au moins une structure phrastique possède une construction phrastique racine qui ne dépend syntaxiquement d'aucun élément de la phrase. Nous appelons désormais cette construction phrastique proposition racine.

Malgré les critiques tout à fait logiques de certains qui dénoncent l'inexactitude de la notion de proposition principale, nous voulons d'autant plus défendre l'existence de l'unité au premier niveau de la phrase que nous envisageons l'alignement de ces unités. Afin de compenser le défaut de la conception classique strictement linéaire, nous considérons que la proposition principale est constituée non seulement de la partie restante de la phrase après extraction des subordonnées, mais aussi d'une sorte de trace des subordonnées extraites.

Nous introduisons alors dans notre représentation des propositions principales, des symboles indiquant l'élément manquant, qui servent en fait à représenter les propositions subordonnées enchâssées extraites.

Ainsi, de la phrase « Quand je suis arrivé, il était déjà rentré », nous extrayons et représentons les propositions comme suit :

Phrase : *Quand je suis arrivé, il était déjà rentré*
Racine : *[A] il était déjà rentré*
Subordonnée : *Quand je suis arrivé,* (indexée A)

Définition 1 (Proposition racine)
Dans une phrase contenant au moins une autre construction phrastique enchâssée à l'aide d'un connecteur ou d'une virgule, la structure phrastique racine – qui ne dépend syntaxiquement d'aucun élément – dans laquelle cette/ces sous-structures sont extraites et représentées par des symboles, est appelée proposition racine. On appelle également proposition racine la proposition indépendante constituant toute seule une phrase simple.

Cependant, cette définition peut entraîner une proposition racine constituée seulement d'un verbe et d'un ou plusieurs symboles indiquant l'élément manquant, représentant plutôt une matrice qu'une proposition. Par exemple, l'analyse de la phrase d'exemple précédemment décrite citée par Le Goffic, entraîne la représentation des propositions constituantes comme suit :

Phrase : *Que Paul ait gagné montre qu'il était le plus fort*
Racine : *[A] montre [B]*
Subordonnée 1 : *Que Paul ait gagné* (indexée A)
Subordonnée 2 : *qu'il était le plus fort* (indexée B)

Appeler proposition une structure telle que « [A] montre [B] » peut être contestable. Néanmoins, nous préférons garder cette appellation, car cet emboîtement des éléments n'est pas un phénomène propre à la racine : une subordonnée peut être une structure de ce type. Nous distinguons les propositions non pas selon leur structure, mais selon leur niveau dans la phrase. Nous désignons par le terme de « proposition racine » la structure racine qui ne dépend syntaxiquement d'aucun élément quel que soit le type de ses constituants, et nous appelons subordonnée la construction phrastique qui dépend syntaxiquement d'un élément, et ce indépendamment du type de ses constituants.

Ce choix, peut-être défavorable du point de vue linguistique, est pris, encore une fois, par considération notamment de notre objectif final qu'est l'alignement.

Proposition subordonnée

Nous définissons une proposition subordonnée comme suit :

Définition 2 (Proposition subordonnée)
La phrase peut contenir d'autres phrases : une structure de phrase non autonome, intégrée à l'aide d'un connecteur de subordination dans une structure de phrase supérieure, est une proposition subordonnée. Sa position dans la phrase est bien définie selon son type.

Les connecteurs de subordination, qui posent également des problèmes, seront abordés dans la section 4.2.

Certaines subordonnées, que Le Goffic considère comme subordonnées sans connecteur, se rattachent à ce type :

– Qu'il pleuve ou qu'il vente, Paul sort tous les jours.

Proposition coordonnée

Avec notre définition de la phrase, nous devons considérer toute séquence entourée de deux séparateurs graphiques comme une seule phrase. Nous analysons donc comme propositions coordonnées deux constructions équivalentes à des phrases, reliées non par deux séparateurs graphiques de phrase, mais par une conjonction de coordination[6] ou par une virgule.

Nous définissons les propositions coordonnées comme suit :

Définition 3 (Proposition coordonnée)
Dans la phrase graphique constituée de plus de deux unités équivalentes à des phrases, l'unité, éventuellement indépendante et autonome, faisant partie de la phrase et reliée par une conjonction de coordination ou une virgule à la proposition qui la précède, est appelée proposition coordonnée.

Cette définition basée uniquement sur un critère formel entraîne également l'inclusion de propositions non coordonnées mais subordonnées, et regroupe des propositions classées dans des catégories différentes dans les travaux linguistiques présentés dans l'état de l'art.

Les phrases ci-dessous sont considérées avec notre définition comme coordonnées (nous indiquons la catégorisation selon Le Goffic entre parenthèses à titre d'exemple) :

– Mon père est professeur et ma mère travaille dans une banque.
– Mon père est professeur, ma mère travaille dans une banque.
– J'accepte, dit-il. *(incise)*
– Vous m'auriez appelé, je serais venu tout de suite. *(subordonnée paratactique)*
– Plus il gagne de l'argent, plus il en veut. *(subordonnée paratactique)*
– Paul a beau crier, on ne l'écoute pas. *(subordonnée paratactique)*
– À peine était-il arrivé, il prenait les choses en main. *(subordonnée paratactique)*

[6] Selon le lexique utilisé par le *tagger* que nous employons, les conjonctions de coordination sont : *et, ni, ou, mais, donc, car, or, soit, c'est-à-dire, voire, sinon, comme, tantôt, y compris, puis.*

Nous avons cependant gardé le terme « coordonnée » pour éviter au maximum un néologisme, position cependant discutable.

Par ailleurs, la représentation hiérarchique de la coordination pose également des problèmes, comme Fuchs & Victorri (1993b) le signalent, car « cette relation n'est, par définition, pas hiérarchique puisque les éléments coordonnés sont mis "sur le même plan" ». Pour des raisons non pas linguistiques mais purement pratiques, nous la représentons comme si le deuxième élément (respectivement, tous les éléments postérieurs) était subordonné au premier (resp. celui qui les précède) en ne marquant la divergence par rapport aux éléments effectivement subordonnés que par l'étiquette « Coordonnée » attribuée aux éléments postérieurs.

Ainsi, la représentation sera comme suit :

Phrase : *Mon père est professeur et ma mère travaille dans une banque.*
Racine : *Mon père est professeur [A]*
Coordonnée : *et ma mère travaille dans une banque.* (indexée A)

Proposition détachée-insérée

Enfin, nous définissons les propositions détachées-insérées comme suit :

Définition 4 (Proposition détachée-insérée)
Nous appelons proposition détachée-insérée une construction phrastique sans connecteur entourée et détachée par deux symboles de ponctuation de même type – virgules, parenthèses ou tirets – et insérée dans une autre phrase. Elle est caractérisée en ce qu'elle peut apparaître en différents endroits de la phrase.

Ce sont des propositions appelées usuellement incises et incidentes.

Il a (on s'en doute) accepté.

À noter que les guillemets n'appartiennent pas aux symboles de ponctuation détachant les propositions. En effet, ils ont vraisemblablement un rôle différent des autres : les symboles tels que les parenthèses ou les virgules enchâssent et insèrent dans une phrase des éléments plus ou moins périphériques, alors que les guillemets servent à souligner des constituants souvent primaires de la phrase. Aussi, les guillemets ne constituent-ils pas des propositions détachées-inserées, mais ils peuvent éventuellement accompagner d'autres types de propositions non détachées.

Par ailleurs, nous ne considérons comme propositions détachées-insérées que les structures constituées d'un sujet et d'un prédicat dont le verbe est bien présent. Ne sont pas traitées, du moins dans le cadre de cette thèse, d'autres structures dégradées au niveau du verbe, notamment les constructions détachées de Combettes (1998) – que nous aborderons dans la section 4.4.3 –, du fait de la distinction difficile entre les éléments réellement extérieurs à la proposition et ceux intérieurs tels que les éléments coordonnés pouvant parfois paraître entourés de virgules.

4.4.3 Éléments extra-prédicatifs

Nous avons vu dans la section 4.1 qu'il existait des éléments extérieurs à l'opposition sujet-prédicat.

Il nous paraît plus cohérent de séparer du reste de la phrase ces constituants, portant sur la phrase dans son ensemble sans appartenir au prédicat, lors de la détection des propositions, quelle que soit leur structure interne.

Les compléments accessoires extra-prédicatifs sont typiquement des constructions détachées, situées en particulier en début de phrase qui est pour l'énonciateur « une zone de liberté relative, avant d'être pris dans le réseau serré des relations syntaxiques de son énoncé » (Le Goffic, 1993a). Cette position est, comme nous l'avons déjà vu dans la section 4.1, la position privilégiée pour les circonstants de phrase, extra-prédicatifs.

Le Goffic énumère les éléments extra-prédicatifs apparaissant en début de phrase comme suit :

1. Éléments invariables
   a) renvoyant à la situation d'énonciation :
      À mon avis ...
      Comme je vous l'avais énoncé, ...
   b) organisant le discours :
      Mais, donc, par conséquent (articulation temporelle ou logique)
      Du point de vue de ...,
   c) portant sur l'énoncé comme un tout :
      Heureusement
      Apparemment
   d) fournissant un cadre circonstanciel ou logique :
      L'autre jour, ...
      Cette affaire étant réglée, ...
      Quel que soit x, ...
   e) précisant l'objet du discours :
      En ce qui concerne ..., Quant à ...
2. Éléments nominaux ou adjectivaux
   a) vocatifs :
      Paul, es-tu prêt ?
   b) actants thématisés :
      Cette affaire, je la connais bien.
   c) adjectivations détachées :
      Furieux, il ...

Les frontières entre ces éléments sont parfois floues : les chercheurs travaillant sur la notion de cadre de discours (Charolles, 1997, 2003 ; Prévost, 2003) signalent la distinction difficile, voire impossible, entre introducteurs de cadre et syntagmes thématisés.

Par ailleurs, la catégorisation peut différer selon les critères adoptés. Par exemple, dans les travaux de Combettes (1998), certains de ces éléments – adjectifs, participes, constructions absolues, infinitifs prépositionnels, adverbes et circonstants prépositionnels – catégorisés dans des classes différentes sont regroupés sous le nom de construction détachée (CD) du fait notamment de leur nature commune de prédication seconde.

Bien que la catégorisation de ces éléments puisse être différente selon le point du vue adopté, leur extériorité est, semble-t-il, largement reconnue. Nous extrayons donc de la proposition ces syntagmes détachés en tête afin de leur accorder un statut équivalent à une proposition.

Ainsi, certaines des constructions que Le Goffic appelle propositions « dégradées au niveau du verbe » seront détectées en tant qu'éléments extra-prédicatifs, sans que nous définissions spécifiquement – du moins dans le cadre de la présente thèse – les propositions participiales ou nominales.
Exemples :

– La nuit tombant, ils rentrèrent.
– Les choses étant ce qu'elles sont, voilà ce que je propose.
– Il errait, l'air furieux.

Ce sont des constructions qui ne comportent pas de verbe fini et aucun constituant interne ne permet de reconnaître *a priori* leur prédicat – et donc leur statut de proposition. Nous les détecterons donc en reconnaissant leur extériorité par rapport au réseau syntaxique du reste de la phrase, par des règles traitant l'ensemble des éléments extra-prédicatifs.

Certaines subordonnées appelées paratactiques par Le Goffic sont également détectées de manière similaire :

– Si malin qu'il soit, ...

Cette phrase sera analysée par les règles traitant le syntagme adjectif ou adverbial suivi d'une proposition corrélative en fonction secondaire (à l'instar de l'analyse faite par Le Goffic pour les locutions conjonctives comprenant un « que » corrélatif telles que « si bien que »). Le syntagme ainsi constitué sera séparé de la proposition racine par la règle traitant les éléments extra-prédicatifs.

À noter que nous ne séparons pas – du moins dans le cadre de la présente thèse – les éléments extra-prédicatifs apparaissant à une autre position que la position initiale du fait de leur extériorité beaucoup moins nette, surtout pour les éléments situés en fin de phrase. Nous examinerons les conséquences de cet ensemble de choix avec les résultats de l'alignement automatique des propositions, mettant en œuvre non seulement l'étude sur les éléments extra-prédicatifs, mais l'ensemble des études linguistiques que nous présentons dans ce chapitre.

4.4.4 Récapitulatif

Récapitulons maintenant les unités à détecter que nous avons définies. Nous avons défini quatre type de propositions selon les deux critères formels, connecteur et position. Nous y avons également ajouté un type particulier ayant une

structure non phrastique, les éléments extra-prédicatifs.

- **propositions** :
    1. **racine** ;
    2. **subordonnée** : introduite par un connecteur de subordination à un endroit déterminé ;
    3. **coordonnée** : introduite par un connecteur de coordination ou une virgule après tout complément du verbe principal ;
    4. **détachée-insérée** : proposition sans connecteur entourée de deux séparateurs de même type et insérée en différents endroits :
- **éléments extra-prédicatifs** : détachés par une virgule en tête de phrase, ex. introducteur de cadre, thème.

4.5 Étude des travaux existants sur les subordonnées

Nous nous intéressons maintenant à la classe de proposition la plus importante, proposition subordonnée, afin de déterminer la typologie la plus adéquate pour notre opération.

Nous analysons d'abord différents types de typologies existantes des subordonnées afin d'étudier leur problématique pour notre finalité. Nous avons classé les typologies proposées en quatre types que nous allons étudier un par un : typologies classiques (§ 4.5.1), classement de Le Goffic (§ 4.5.2), typologies selon la catégorie du mot équivalent (§ 4.5.3), typologies selon la fonction dans la racine (§ 4.5.4).

Après cet état de l'art et les examens critiques, nous proposerons nos éléments de solution (§ 4.5.5) pour la définition d'une typologie adaptée au développement d'un détecteur des propositions.

4.5.1 Typologies classiques des subordonnées

Les classements les plus usuels sont réalisés selon une méthode combinée : les subordonnées sont d'abord classées selon la nature du connecteur, puis certains types sont divisés eux-mêmes en sous-catégories selon la fonction que joue la subordonnée dans la phrase.

Dans les éditions postérieures à la 11ème édition de *Le bon usage* telles que Grevisse (1993), refondue par Goosse, les propositions sont divisées en trois catégories selon la nature du connecteur.

1. **propositions relatives** : commençant par un pronom relatif (qui, que, quoi, dont, où, lequel, quiconque) ou par un syntagme contenant le pronom relatif ou parfois par un nom accompagné d'un déterminant relatif

a) relatives sans antécédent

b) relatives avec antécédent

2. **propositions conjonctives** : commençant par une conjonction ou une locution conjonctive de subordination
   a) propositions conjonctives essentielles
   b) propositions corrélatives
   c) propositions adverbiales
3. **propositions d'interrogation et d'exclamation indirectes** : rattachées à la phrase par aucun mot particulier, à l'exception de l'interrogation globale qui est rattachée à la phrase par la conjonction de subordination « si ».

Beaucoup de grammaires (Hatier, 1990 ; Gardes-Tamine, 1998 ; Wagner & Pinchon, 1991) proposent une typologie comparable (cf. tableau 4.2).

Bescherelle	relatives		complétives	circonstancielles	interrogations indirectes
Goosse	relatives		conjonctives essent./corrél.	conjonctives adverbiales	interr./exclam. indirectes
Gardes-Tamine	relatives substantives	relatives adjectives	conjonctives pures	conjonctives circonstancielles	interrogatives indirectes
Wagner et Pinchon	relatives		conjonctives	circonstancielles	interrogations indirectes

TAB. 4.2 – Correspondance des classes de subordonnées

Problèmes liés à la difficulté d'étiquetage des connecteurs de subordination

Sans parler de la question théorique liée à la notion discutable de « relative sans antécédent », le plus grand problème de ces classements classiques pour notre traitement automatique est qu'ils présupposent l'analyse correcte des connecteurs. Or, l'étiquetage des connecteurs est, comme nous allons le voir, extrêmement difficile, parfois impossible, surtout à l'étape de *tagging* sans une analyse syntaxique plus large.

Nous avons utilisé jusqu'ici le terme connecteur de subordination sans le définir exactement. Avant de montrer la difficulté de leur étiquetage, nous essayons de déterminer ces connecteurs du français, éléments qui nous paraissent souvent maladroitement définis.

Wagner & Pinchon (1991) distinguent quatre types de « mots dont le rôle consiste à marquer le caractère dépendant de la proposition qu'ils ouvrent » :

1. des conjonctions (**que, comme, quand, si**) et des locutions conjonctives construites au moyen de **que** (**afin que, alors que, de peur que, du moment, que, lorsque, pour que,** etc.), de **où** (**du moment où, là où**) ;
2. des adverbes interrogatifs **quand ? comment ? où ? pourquoi ?** et des pronoms interrogatifs ;

3. des pronoms relatifs représentants ;
4. des adverbes de quantité simples (**tant, tellement**).

Beaucoup de grammaires telles que Riegel et al. (1994) proposent une définition des connecteurs de subordination plus ou moins semblable à celle-ci et cette catégorisation correspond également *grosso modo* à celle de l'étiqueteur que nous utilisons.

Cette définition pose un problème crucial pour les travaux à caractère appliqué : la difficulté d'étiquetage.

Certains de ces mots sont très ambigus et un étiquetage erroné provoquerait des erreurs dans l'opération postérieure. Dans le tableau 4.3, nous avons représenté les différentes étiquettes que peuvent recevoir les connecteurs selon la catégorisation adoptée pour le corpus de Paris 7 (Abeillé & Clement, 2003). Nous pouvons y constater la forte ambiguïté de ces connecteurs. La détermination de la pertinence de ces distinctions pour notre opération est d'autant plus utile que leur étiquetage correct est loin d'être simple. Le choix d'une étiquette adéquate nécessite souvent une analyse syntaxique.

	pronom		det.	adverbe			conjonction		autres
	rel.	inter.		inter.	excl.	autre	sub.	crd.	
dont	√								
qui	√	√							
que (qu')	√	√			√	√	√		
quoi	√	√							
lequel*	√	√							
où	√			√					
quel*			√	√					
comment				√					
pourquoi				√					
combien				√	√				
quand				√			√		
comme					√		√	√	prép.
si						√	√		note de musique ou affirmation
s'							√		clitique

* ainsi que toutes leurs formes fléchies.

TAB. 4.3 – Ambiguïtés des connecteurs

Il existe d'abord des erreurs liées à la distinction très difficile de deux étiquettes possibles, telle que celle entre relatif et conjonction de subordination[7] :

> C'est la ville de notre enfance, ce sont des paysages **que**$_{*[C\text{-}S]}$ nous traversions.

ou entre conjonction de subordination et adverbe interrogatif :

[7] Les étiquettes attribuées par le *tagger* sont marquées entre crochets en indice. Les étoiles devant indiquent que l'étiquette attribuée est erronée.

quand$*_{[ADV\text{-}int]}$ rien ne va , rien ne va !

Cependant, ces erreurs ne posent pas de problème lorsqu'il s'agit uniquement de la détection des frontières de propositions. Les erreurs plus graves qui risquent d'empêcher la reconnaissance même des frontières de propositions sont des confusions d'étiquettes entre celles susceptibles d'introduire une proposition et celles qui ne le sont pas. Dans la phrase suivante, les deux « comme » sont considérés comme conjonction de subordination alors que « *Comme* est P [préposition] quand il introduit une comparative réduite (sans verbe) » (Abeillé & Clement, 2003) :

> Qu'il soit total **comme**$*_{[C\text{-}S]}$ à Kilinochchi, ou partiel **comme**$*_{[C\text{-}S]}$ à Jaffna, le pouvoir des tigres est expéditif.

Nous constatons également, comme dans les phrases suivantes, que le « que » de conjonction de subordination introduisant une complétive est parfois considéré comme adverbe simple – qui n'est pas un introducteur de proposition – ou l'inverse :

> Comme cette dernière, plusieurs sociologues relèvent **que**$*_{[ADV]}$ jamais la nostalgie pour les années 1970-1990 n'a été aussi forte [...].

> Si M. George W. Bush n'est **que**$*_{[C\text{-}S]}$ le dernier de la lignée, c'est également l'un des plus performants dans ce registre de l'homme politique simultanément inféodé aux priorités des milieux d'affaires et capable de s'exprimer avec la voix des damnés de la terre.

Les mêmes erreurs peuvent se produire avec « si » :

> Ils risquent même de constituer une menace sérieuse **si**$*_{[ADV]}$ jamais la situation politique se détériore de nouveau.

En outre, il existe un autre problème : incohérence entre des mots appartenant à une même catégorie. En effet, deux mots ayant la même étiquette peuvent avoir des comportements syntaxiques différents. Par exemple, « dont » et « que » sont tous les deux des pronoms relatifs. Or, « dont » peut introduire non seulement une proposition (ex. 1, 2), mais aussi un syntagme (ex. 3)[8], alors qu'un « que » relatif n'introduit qu'une proposition[9].

1. Le gouvernement a retiré sa proposition **dont** la conformité à la Constitution avait été remise en cause.
2. Ils ont enfin trouvé la maison **dont** ils rêvaient depuis longtemps.
3. À cette occasion, se sont réunis huit représentants **dont** notre Président.

[8] On entend ici par « proposition » et « syntagme », des unités purement de surface. Nous n'entrons pas dans la discussion sur la véritable nature de ces unités introduites par ces connecteurs, que certaines théories linguistiques traitent comme un phénomène d'ellipse.

[9] Nous trouvons tout de même, dans Grevisse (1993), deux types d'exemples – bien que qualifiés de rares – de « que » relatif introduisant une structure non phrastique : suivi d'un gérondif « ce QUE *voyant* (= en voyant cela) » d'une part, et dans le style juridique « TOUT CE QUE *dessus sera fait de suite* (Code civil, art. 976) » de l'autre.

Dans le cas de la définition adoptée par notre *tagger*, le même problème se pose entre les conjonctions de subordination « comme » et « que ». Par définition, « *Comme* est CS [= conjonction de subordination] dans les interrogatives indirectes, les subordonnées causales et les comparatives non réduites » (Abeillé & Clement, 2003) (souligné par nous-mêmes). En revanche, « QUE est conjonction de subordination, après un verbe (ou un nom ou un adjectif) à complétive, après une Prép, dans les comparatives ou les corrélatives (mêmes réduites) et dans les impératives » (souligné par nous-mêmes). Pour définir une grammaire conforme à cette définition, il serait déjà impossible de conserver ces étiquettes qui regroupent des éléments ayant des comportements syntaxiques différents.

La détermination de la pertinence de la distinction entre ces différentes étiquettes pour une tâche donnée est d'autant plus utile et même indispensable que cette opération est loin d'être aisée.

4.5.2 La typologie proposée par Le Goffic

Les études sur les subordonnées de Le Goffic se situent dans le cadre de ses travaux plus larges sur les termes en « qu- ». Il s'agit d'« une vieille famille indo-européenne en **kw-*, remarquablement conservée » de termes qui sont « fondamentalement des indéfinis, c'est-à-dire des marqueurs du parcours de toute la classe : classe des animés (*qui*), des lieux (*où*), des moments (*quand*), etc. » (Le Goffic, 1992). Sur la base de la thèse selon laquelle les termes en « qu- » sont des marqueurs désignant une variable, il essaie de « parvenir à une présentation unifiée et globale de l'ensemble des emplois des termes en *qu-*. » (Le Goffic, 2002).

Dans ces études, les connecteurs de subordination sont divisés en quatre types, percontatif, intégratif, relatif, complétif, qui correspondent respectivement à quatre types de propositions subordonnées différentes. Chaque connecteur constitue des subordonnées différentes :

1. percontative (interrogative indirecte)[10] :
   - je sais **qui a gagné la course** (où il est allé, quelle mouche l'a piqué).
   - Paul cherche **comment il pourrait faire**.
   - Paul se demande **s'il va réussir**.
2. intégrative :
   a) pronominale (relative sans antécédent)
      - **Qui dort** dîne.
      - Embrassez **qui vous voulez**.

   b) adverbiale (circonstancielle en « qu- » ou « si »)
      - **Quand on veut**, on peut.
      - **Si vous avez fini**, vous pouvez sortir.

[10] Les termes entre parenthèses sont des dénominations usuellement utilisées que l'auteur présente comme la correspondance de ses classes.

  - Marie est aussi jolie **qu'elle est gentille**. (corrélatif)[11].

3. relative (relative avec antécédent)
   - Le médecin **qui est venu**.
   - La maison **où je suis né**.
4. complétive (complétive)
   - Je crois **qu'il va pleuvoir**.
   - La peur **que le ciel leur tombe sur la tête**.
   - **Qu'elle fût bien ou mal coiffée**, je l'admirais.[12]

Examen critique du classement de Le Goffic

Avant d'aborder l'analyse de son classement des subordonnées, nous passons en revue sa définition des connecteurs, différente de la définition usuelle présentée dans la section précédente.

Définition des connecteurs dans les travaux de Le Goffic

Le Goffic (1993a,b) considère que les termes en « qu- » sont, avec « si », les seuls connecteurs du français et qu'ils appartiennent tous à une des trois catégories : pronoms, adjectif et adverbes.

- **pronoms** : qui, que, quoi, lequel ;
- **adjectif** : quel ;
- **adverbes** : où, quand, comme, comment, combien, que (homonyme du pronom), dont, pourquoi.

La principale particularité de cette définition réside dans l'absence de catégorie de conjonction. Le Goffic renonce également à la notion de locution conjonctive, en dénonçant le « caractère peu satisfaisant » de leur liste traditionnelle et l'absence de véritable analyse des propositions considérées comme introduites par ces locutions conjonctives.

[11] Les corrélatives sont considérées ici comme des circonstants, « au rebours de la tendance actuelle ». Aujourd'hui, beaucoup de linguistes excluent les corrélatives des circonstancielles et en font des constituants secondaires à l'instar des relatives « en considérant que *plus aimable que ne l'était sa sœur* forme un GAdj (= *aimable* + quantification par un GAdv discontinu *plus ... que* P) et que la corrélative n'a aucune autonomie de placement dans la phrase (ni même par rapport à son antécédent). » Malgré ce courant, Le Goffic défend sa position par le fait que « les corrélatives (toujours facultatives) sont en fait souvent séparées de leur antécédent, d'une façon incompatible avec la structure d'un groupe. » Toutefois, il signale également des cas d'exception : les corrélatives sont d'autant moins autonomes qu'elles sont elliptiques ; elles sont considérées comme des constituants secondaires dans le cas des locutions conjonctives du type « si bien que » où elles sont inséparables de leur antécédent.

[12] Les propositions introduites par le « que » complétif peuvent également avoir le statut de subordonnée paratactique ou de terme nominal proleptique. L'élément en prolepse (ou disloqué à gauche) est défini selon lui comme un élément détaché en début de phrase, repris par un pronom anaphorique qui en précise la fonction. L'élément détaché en fin de phrase est dit en reprise (ou disloqué à droite).

Dans ses travaux, les unités introduites par une locution conjonctive sont analysées comme des groupes adverbiaux ou des groupes prépositionnels comprenant une subordonnée introduite par un véritable connecteur en « qu- ». Par exemple, « pour que P » est analysé non pas comme une subordonnée, mais comme un groupe prépositionnel constitué de la préposition « pour » suivi d'une complétive introduite par « que » ; « du moment où P » est analysé comme un groupe prépositionnel contenant une relative introduite par « où » ; « aussitôt que P » est analysé comme un groupe adverbial constitué de l'adverbe « aussitôt » suivi d'une intégrative corrélative.

Cette analyse permet un traitement unifié et homogène des unités « propositions ». Mais son plus grand atout pour notre finalité est d'annuler, par exclusion de la catégorie de conjonction, le caractère polycatégoriel de la plupart des connecteurs, facilitant ainsi considérablement l'étiquetage automatique.

Inconvénients de la classification des subordonnées de Le Goffic

Malgré tout l'intérêt théorique qu'elle présente, la typologie de Le Goffic ne permet pas pour autant la conception d'un système simple de détection automatique des propositions. En effet, le problème est que, dans cette théorie, les connecteurs possèdent différents emplois dans lesquels chaque connecteur introduit différents types de subordonnées. Le tableau 4.4, reproduit de Le Goffic (2002), est une vue d'ensemble de leurs emplois.

	interrogatifs	**indéfinis**	**intégratifs emphatiques**	**intégratifs**	**relatifs**
+ h	*qui*	*qui…qui*	*qui*	*qui*	- Prép + *qui*
- h	*quoi* / *que*	-	*quoi*	-	- Prép + *quoi*
entité N (±h)	*quel* *lequel*	*quelque*	*quel* / *quelque*	-	*qui* / *que* / *lequel* Prép + *lequel* *dont*
Lieu	*où*	-	*où*	*où*	*où*
Temps	*quand*	-	-	*quand*	
Manière	*comment* *comme* excl.	-	-	*comme*	-
Quantité (Degré)	*combien* *que* (adv.) excl.	-	*quelque* (adv.)	*que* (adv.)	-

TAB. 4.4 – Emploi des marqueurs *qu-* du français

Autrement dit, tout en facilitant l'opération d'étiquetage, sa catégorisation des connecteurs ne permet pas directement de repérer chaque type de proposition qu'il définit et l'identification des subordonnées nécessite une étape supplémentaire dédiée à l'analyse de l'emploi exact du connecteur dans le contexte où il est

utilisé. Ce qui représente, finalement, une tâche aussi délicate que l'étiquetage avec la catégorisation classique des connecteurs.

4.5.3 Typologies selon la catégorie du mot simple équivalent

Jusqu'à la 11ème édition de *Le bon usage*, les subordonnées sont divisées en trois classes selon la nature du mot auquel elles sont assimilables et la fonction qu'elles remplissent dans la phrase.

1. **substantives** : assimilables à des noms et correspondant aux compléments d'objet ou aux compléments de l'adjectif ou de l'adverbe. Elles peuvent aussi être sujets, attributs ou termes complétifs d'un nom ou d'un pronom ;
2. **adjectives ou relatives** : assimilables à des adjectifs ou à des participes-adjectifs et correspondant aux compléments du nom ou du pronom ;
3. **adverbiales ou circonstancielles** : assimilables à des adverbes et correspondant aux compléments circonstanciels.

Biskri & Desclés (2005) proposent une typologie similaire basée sur la Grammaire Catégorielle Combinatoire Applicative qui, d'après les auteurs, favorise le traitement automatique des langues. Selon cette catégorisation, les relatives, les complétives et les interrogatives (les circonstancielles ne sont pas traitées dans leurs travaux) sont distinguées en deux grands types selon les opérateurs servant à construire les propositions. En effet, les connecteurs sont considérés ici comme des opérateurs qui rattachent la subordonnée à la principale et ils sont divisés en deux classes : les constructeurs de noms et les constructeurs de modifieurs. Les propositions construites avec ces opérateurs agissent de la même façon que des substantifs ou des adjectifs.

Dans cette perspective, les propositions relatives sans antécédent s'apparentent plus aux interrogatives qu'aux relatives avec antécédent. Pour justifier ce classement, les auteurs s'appuient d'abord sur les arguments de Le Goffic :

– la proposition interrogative emploie presque toujours la troisième personne du singulier alors que la proposition relative avec antécédent s'accorde en genre et en nombre avec l'antécédent ;
– les propositions relatives sans antécédent et les propositions interrogatives indirectes peuvent avoir la fonction d'objet direct.

Ils défendent ensuite la capacité de construction de noms des relatifs sans antécédent, interrogatifs et complétifs par le fait que « ces opérateurs permettent la construction de syntagmes référençant une partie de la réalité. »

Examen critique des classements selon la catégorie équivalente

Ces typologies qui ne se fondent pas sur les types de connecteurs qui introduisent les subordonnées semblent mieux adaptées à la définition d'une grammaire pour la détection des propositions. Cependant, il existe d'autres problèmes.

Comme le signalent Riegel et al. (1994), le parallélisme des catégories n'est que partiel : les relatives ne peuvent, par exemple, pas assurer la fonction d'attribut en dépit de leur apparente équivalence à l'adjectif.

Par ailleurs, sur le plan pratique dans le cadre de nos travaux, cette typologie qui classe les subordonnées dans seulement trois catégories, risque de multiplier le nombre d'analyses possibles d'une phrase.

Par exemple, une subordonnée en « que » peut être substantive, adjective et adverbiale, et avec ces trois possibilités, le nombre d'analyses possibles d'une phrase qui en contient une risque d'être très important, surtout avec la quantité restreinte d'information dont nous disposons pour l'analyse. La discrimination d'un type par rapport aux autres selon la fréquence est impossible car il n'existe pas d'homogénéité même à l'intérieur d'un type : une subordonnée substantive en « que », par exemple, est extrêmement fréquente à la fonction de complément mais elle l'est moins à la fonction de sujet. Ainsi, une fois tous les candidats calculés, une étape supplémentaire serait nécessaire pour choisir la réponse la plus probable.

Néanmoins, il nous semble possible, en définissant une typologie tenant compte d'autres critères, de contrôler plus efficacement le nombre d'analyses possibles et d'obtenir la réponse la plus probable sans ajout d'étape supplémentaire.

4.5.4 Typologies selon la fonction dans la racine

On peut trouver dans Grevisse (1969, 1990) un classement des propositions selon leur fonction. L'auteur y énumère douze fonctions que peut jouer une proposition.

1. **sujet** : il faut que l'on patiente.
2. **attribut** : le remède serait que tu vives dans la solitude.
3. **apposition** : ne renversons pas le principe que le droit prime la force.
4. **objet direct** : j'attends qu'il revienne.
5. **objet indirect** : je consens qu'il parte.
6. **compl. circonst.** : opposez-vous au mal avant qu'il s'enracine.
7. **compl. d'agent** : cet homme est aimé de quiconque le connaît.
8. **compl. détermin.** : la modestie qui procède de l'orgueil est détestable.
9. **compl. explicatif** : la modestie, qui relève si bien le mérite, sied aux savants.
10. **compl. d'adjectif** : certain qu'il vaincra, le lièvre se repose.
11. **compl. du comparatif** : Pierre est plus savant qu'on ne pense.
12. **compl. du présentatif** : voici que la nuit vient.

Chevalier et al. (1964) proposent également un classement des propositions subordonnées selon la fonction. Enfin, Delaveau (2001) propose une typologie un peu particulière réalisée non pas selon la fonction dans la racine, mais selon l'élément de la racine qui domine la proposition. Elle définit quatre classes : dominée par GN, dominée par GV, dominée par GA, dominée par GP.

Examen critique des classements selon la fonction

Ces typologies qui ne présupposent pas d'analyse correcte des types de connecteurs sont également favorables à notre tâche. Toutefois, elles possèdent aussi des inconvénients.

Ces typologies divisent les subordonnées en un nombre plus ou moins important de classes, alors qu'une économie des règles de grammaire serait sans doute envisageable par une restriction des types non pertinents pour notre opération de détection des propositions, réduisant ainsi les calculs nécessaires.

Mais le plus grand défaut de ces typologies, est que les structures de subordination n'y sont décrites que partiellement avec seulement des exemples triviaux, ne nous fournissant pas suffisamment d'informations. En effet, nous ne pouvons pas savoir exactement quelles sont les subordonnées de fonction sujet, complément, etc. Une description précise permettrait sans doute de mieux rendre compte des points communs et des divergences entre les types et, avec cette étude, de réorganiser la typologie afin d'en obtenir une plus économique et suffisamment efficace.

4.5.5 Éléments de solution

Nous devons considérer deux points pour le choix de la typologie des subordonnées. Premièrement, comme pour tous les travaux de traitement automatique, la description doit être systématique et précise. Deuxièmement, du fait de la difficulté d'étiquetage que nous avons abordée, il est préférable que la typologie ne présuppose pas une analyse correcte des connecteurs selon la catégorisation classique.

Tenant compte de ces deux prérequis spécifiques, nous avons défini une typologie des subordonnées selon les critères combinés de catégorie/position adaptée à notre opération de détection des propositions.

Nous avons d'abord distingué trois types de subordonnées : substantives, adjectives et adverbiales.

Ensuite, en étudiant pour chaque position dans la phrase les types de subordonnées susceptibles d'apparaître, nous avons réalisé une typologie finale selon la position. Nous distinguons cinq types de subordonnées selon leur position dans la phrase :

- position post-verbale ;
- autres positions pouvant être occupées par un SN ;
- positions initiale et finale ;
- position post-nominale ;
- positions post-adverbiale et post-adjective.

Enfin, chaque classe a été décrite de manière systématique et précise, à l'aide de la description détaillée des subordonnées de Le Goffic.

Notre typologie présente tout d'abord comme avantage l'indépendance vis-à-vis de la qualité d'analyse des connecteurs dans les catégories classiques ou de

leur emploi exact. De plus, le critère de position a permis d'obtenir une division plus optimisée pour définir une grammaire, qu'avec les deux critères traditionnellement utilisés, la catégorie et la fonction.

4.6 Notre typologie des subordonnées selon la position

Nous allons aborder dans cette section la définition globale et les caractéristiques de la typologie ainsi définie pour passer ensuite à la description détaillée de chaque classe.

Mais, avant d'entrer dans la présentation, faisons un point sur ce que nous appelons connecteur de subordination.

Définition préliminaire des connecteurs de subordination

Nous adoptons, dans un premier temps, la catégorisation des connecteurs sans classe de conjonction définie par Le Goffic (une autre plus adaptée à nos travaux sera présentée dans § 4.7). Toutefois, pour des raisons pratiques – notamment par souci d'utilité pour l'alignement –, nous conservons le statut de connecteur locutionnel que nous attribuons aux locutions conjonctives, regroupées et étiquetées comme CS (conjonction de subordination) par notre *tagger*.

4.6.1 Premier classement selon la catégorie

Selon notre typologie, les subordonnées sont d'abord divisées en trois types correspondant à trois catégories de mot simple :

1. **substantives** : auxquelles appartiennent les
   a) intégratives pronominales :
      Qui dort dîne
      Embrassez qui vous voulez
   b) complétives :
      Que vous ayez menti me déçoit
      Je pense qu'il viendra
   c) percontatives :
      Comment il a commis ce crime n'a jamais été établi
      Je me demande qui a fait cette bêtise
      Je me demande où il est parti
      Je me demande s'il est parti
      Il ne m'a pas dit quand il rentrerait
      Je ne vois pas à quoi tu fais allusion
      Je ne sais pas lequel de ces romans paraîtra le premier en livre de poche

Je ne comprends pas quels sont ses intérêts

Il ne m'a pas dit pourquoi il n'était pas venu

Je me moque de comment il a réussi

Je me demande combien de promesses il n'a pas tenues

Je ne comprends pas comment tu oses dire ces choses

Voyez comme c'est facile

2. **adjectives** : auxquelles appartiennent les
   a) relatives[13] :

      La peinture qui m'a fascinée

      La peinture dans laquelle notre maison était reproduite

      Ce à quoi je m'attendais

   b) complétives :

      L'idée que tout est fini

3. **adverbiales** : auxquelles appartiennent les
   - intégratives adverbiales :

      Quand je suis arrivé, il était déjà rentré

      Si tu ne manges pas, tu ne guériras pas

      Comme elle est écrite en chinois, il n'a pas pu lire cette lettre.

      Où il y a de la gène, il n'y a pas de plaisir. (repris de Le Goffic (1993a))

      Il était déjà rentré quand je suis arrivé.

      Tu ne guériras pas si tu ne manges pas.

      Il n'a pas pu lire cette lettre comme sa mère l'avait deviné.

      Tu peux poser ton manteau où tu veux.

      La maison est restée aussi conviviale qu'elle l'était avant.

      La nouvelle l'a tellement surprise qu'elle s'est mise à pleurer.

[13] Le Goffic (1993b) distingue les relatives pronominales des adverbiales introduites par « où ». Il reconnaît une équivalence fonctionnelle de ces premières avec le groupe nominal et celle de ces dernières, avec le groupe adverbial. Nous suivons, sur ce sujet, plutôt le modèle classique selon lequel les relatives sont considérées comme équivalentes à l'adjectif et nous ne faisons pas non plus de distinction particulière entre ces deux types.

Notre choix lié à l'équivalence fonctionnelle des relatives se base essentiellement, comme beaucoup d'autres travaux, sur leur statut de complément adjectif permettant de les coordonner à un adjectif.

Quant aux relatives introduites par « où », nous avons en outre une raison liée à la présence des intégratives en « où », équivalentes au groupe adverbial. Afin de distinguer les relatives des intégratives en « où » jouant le rôle de complément secondaire du nom – qui nous semble théoriquement possible encore que nous ne connaissions pas d'exemple –, nous voulons d'autant plus éviter de leur attribuer le statut adverbial.

4.6.2 Second classement selon la position : description de chaque classe

Tenant compte de ce premier classement, nous distinguons cinq types de subordonnées selon leur position dans la phrase :

1. **position post-verbale** : subordonnée complément en Qu- (subQ)
   assure une fonction de complément et concerne les propositions **substantives** ;
2. **autres positions pouvant être occupées par un SN** : subordonnée SN (subSN)
   assure une fonction de sujet ou autre et concerne les propositions **substantives** ;
3. **positions initiale et finale** : subordonnée circonstancielle ou périphérique (subP)
   assure une fonction accessoire et concerne les propositions **adverbiales** ;
4. **position post-nominale** : subordonnée déterminante ou relative (subR)
   asssure une fonction secondaire et concerne les propositions **adjectives** et **adverbiales** ;
5. **positions post-adverbiale et post-adjective** : subordonnée en « que », complétive, corrélative ou relative, généralement analysée comme une proposition introduite par une locution conjonctive.

Chaque type de subordonnée à une position donnée est caractérisé par sa fréquence, afin de pouvoir favoriser l'interprétation comme subordonnée courante par rapport aux subordonnées rares. Faute de données permettant d'obtenir des statistiques représentatives, la définition de ces fréquences est réalisée de manière empirique. La justesse de ces hypothèses est examinée dans l'évaluation (§ 9.3).

Afin d'élaborer la description de chaque type de manière à obtenir un caractère suffisamment complet pour fournir une base pour la définition d'une grammaire globale et formelle, nous nous sommes appuyés sur les travaux de Le Goffic.

Nous utilisons donc la typologie de Le Goffic dans la description de nos classes de subordonnées : cette terminologie servira également de passerelle entre notre classement et les théories traditionnelles. Nous allons maintenant présenter chaque classe de subordonnées.

4.6.3 Position post-verbale : subordonnée complément en Qu- (subQ)

À cette position, apparaissent les propositions substantives : complétives, intégratives pronominales, percontatives.

- **substantives**
  a) complétives

     Je pense qu'il viendra

  b) intégratives

Embrassez qui vous voulez

c) percontatives

Je me demande s'il est parti

Il ne m'a pas dit quand il rentrerait

Voyez comme c'est facile

4.6.4 Autres positions SN : subordonnée SN (subSN)

Les propositions substantives apparaissent également, bien qu'assez rarement, à d'autres positions où un syntagme nominal peut apparaître : position sujet, après une préposition, position initiale (termes en prolepse).

1. position sujet : **substantives** (rare)
   a) Intégrative

      Qui dort dîne

   b) Complétive

      Que vous ayez menti me déçoit

   c) Percontative

      Qui a commis ce crime n'a jamais été établi

      Comment il a commis ce crime n'a jamais été établi

      Pourquoi il a commis ce crime n'a jamais été établi
2. après une préposition : **substantives**
   a) Intégrative : (rare)[14]

      Je voterai pour qui me promettra moins d'impôts. (tiré de Hatier (1990))

      Le pouvoir est seulement entre les mains de qui détient des armes à feu, de qui possède les richesses.

      Pour qui appartient aux classes moyennes, le fait de partir de chez soi chaque matin est un combat.

   b) Percontative : (rare)[15]

      Dominique de Villepin n'a d'ailleurs guère laissé planer de doute sur qui prendrait la décision finale.

      Il faudra se poser la question de pourquoi nous avons été choisis.

      Plus récemment se pose la question de comment l'Etat doit considérer les groupes et minorités défavorisés, s'il souscrit à l'idéal de traiter tous les citoyens et citoyennes comme égaux, indépendamment de leur appartenance sexuelle, religieuse ou ethnique.

[14] Les deux derniers exemples d'intégrative sont tirés de « Le Monde Diplomatique ».

[15] Le premier exemple est emprunté d'un article publié sur « Yahoo ! France ». Les autres exemples sont des résultats de requêtes dans « Google ».

Ce n'était plus une question de "si" mais bien une question de "quand" une telle échéance allait se produire.

La Cour n'a pas jugé nécessaire de trancher la question de si les Québécois formaient ou non "un peuple".

Selon lui, il y a quelque chose dans le statut de l'objet de la science qui reste comme non élucidé dès sa naissance, et la question de si la psychanalyse est scientifique ou non, ne pourrait donc pas se résoudre jusqu'à ce qu'on arrive à modifier le statut de la science comme tel.

c) Complétive : (fréquente dans les locutions)

après que, avant que, depuis que, dès que, malgré que, pendant que, pour que, sans que, sauf que, selon que, ... etc.

3. position initiale en prolepse : **substantives**

a) Intégrative pronominale

Qui ferait cela, il agirait sagement (obsolète). (repris de Le Goffic (1993a))

b) Percontative

Comment il a fait, je vous le demande ! (repris de Le Goffic (1993a))

c) Complétive

Qu'il y eût en tout être, et en lui d'abord, un paranoïaque, il en était assuré depuis longtemps. (repris de Chevalier et al. (1964))

Propositions substantives après une préposition

Les intégratives et les percontatives sont rares à cette position, mais les complétives y sont utilisées très fréquemment et constituent les locutions dites conjonctives. Dans nos travaux, ces locutions étant regroupées et étiquetées par le *tagger* comme conjonctions de subordination, nous ne devrions pas rencontrer de complétives seules apparaissant à cette position. Or, la liste sur laquelle se base l'étiqueteur peut être incomplète. Nous gardons donc la possibilité d'avoir une complétive (non constituant d'une locution) après une préposition – avec, comme indication de fréquence, rare – afin de pouvoir détecter la proposition introduite par la locution conjonctive que l'étiqueteur n'a pas réussi à regrouper.

Propositions substantives en prolepse en position initiale

Un élément en prolepse est « jeté en avant, posé pour lui-même, hors fonction et hors structure, comme si l'énonciateur commençait par indiquer le ou les objet(s) de son discours, avant même d'avoir arrêté un projet de phrase syntaxique » (Le Goffic, 1993a). Leur extériorité est si forte que la percontative en prolepse, en particulier, « peut aussi être interprétée comme une interrogation indépendante » (Le Goffic, 1993a). Même remarque dans Chevalier et al. (1964, p. 120) :

> « elle [= la proposition interrogative] prend parfois tant d'indépendance qu'elle peut retrouver les tours de l'interrogation directe :
>
> *Ses projets commerciaux se mêlaient-ils à ses repentirs de bedeau, je n'en sais rien* (Jacob). »

Ces termes en prolepse sont généralement repris et intégrés syntaxiquement par une anaphore. Comme le signale Le Goffic, « le français a perdu depuis l'époque classique l'usage des intégratives pronominales en prolepse. »

4.6.5 Positions initiale et finale : subordonnée circonstancielle ou périphérique (subP)

Ces positions concernent l'ensemble des locutions conjonctives de subordination et les intégratives. Nous étudions seulement ces dernières. Par ailleurs, apparaissent également les propositions en « que » analysées souvent comme subordonnées paratactiques.

Les subordonnées apparaissant à cette position peuvent être caractérisées – à l'exception des subordonnées en « que » intégratif (corrélatif compris) qui n'apparaissent qu'en position finale – par leur liberté liée à la position de leur occurrence : elles peuvent apparaître non seulement aux positions initiale et finale, mais elles peuvent aussi être insérées entre différents éléments de la phrase, sous forme détachée par deux séparateurs de même type tout comme les propositions détachées-insérées.

1. **adverbiales** : intégratives adverbiales
   - Position initiale

     Quand je suis arrivé, il était déjà rentré.

     Si tu ne manges pas, tu ne guériras pas.

     Comme elle est écrite en chinois, il n'a pas pu lire cette lettre.

     Où il y a de la gène, il n'y a pas de plaisir. (repris de Le Goffic (1993a), rare)
   - Position finale

     Il était déjà rentré quand je suis arrivé.

     Tu ne guériras pas si tu ne manges pas.

     Il n'a pas pu lire cette lettre comme sa mère l'avait deviné.

     Tu peux poser ton manteau où tu veux.

     Il était à peine arrivé qu'il était déjà assailli. (repris de Le Goffic (1993a))

     Il aurait bu que je n'en serais pas surpris. (repris de Le Goffic (*Ibid.*))

     Viens ici, que je t'embrasse. (repris de Le Goffic (*Ibid.*))

     Le crocodile n'eut pas le temps de se demander ce que lui voulait ce lourdaud, que Gropopotin s'était déjà assis sur son dos.[16]

[16] Repris de *Gropopotin l'hippopotame*, « Wakou », numéro 206, mai 2006.

Mehdi a tout juste le temps de sauter sur son lit que déjà voilà Maman qui ouvre la porte.[17]

La maison est restée aussi conviviale qu'elle l'était avant.

La nouvelle l'a tellement surprise qu'elle s'est mise à pleurer.

2. **adverbiales ou substantives** : intégratives ou complétives (?[18])

Que le gouvernement propose une nouvelle loi, l'opposition crie au scandale.

Je pars, que cela vous plaise ou non.

Propositions en intégrative

Les propositions en intégrative adverbiale sont celles que Le Goffic considère comme les seules subordonnées méritant le nom de « circonstancielles » (Le Goffic, 1993b).

Comme le remarque Le Goffic, l'intégratif en « où » en position initiale est rare. Bien que « comme » apparaisse aussi bien en position initiale que finale, son interprétation diffère dans les deux cas. Celui apparaissant en fin de phrase est un adverbe de prédicat exprimant la manière, alors que celui en position initiale est un adverbe de phrase à valeur temporelle ou causale ou un adverbe d'énonciation[19].

Propositions paratactiques en « que »

C'est une proposition très délicate à analyser. Le Goffic (1993a) analyse ce « que » comme complétif. D'après sa théorie, les complétives sont équivalentes au groupe nominal. Or, ces propositions paratactiques se déplaçant librement nous donnent une impression plus proche de celle des adverbes circonstanciels. D'ailleurs, dans le cadre de la constitution d'un corpus de circonstants au sein du laboratoire ELSAP, elles sont considérées comme circonstancielles (Guimier, 1993, p. 30) :

> « Ont également été incluses dans cette catégorie certaines propositions circonstancielles en *que* : [...]

[17] Repris de *Le tapis magique*, « Histoires pour les petits », numéro 42, mai 2006.

[18] L'interprétation de ces propositions étant très délicate, nous laissons en suspens leur analyse exacte (cf. *infra.* « Propositions paratactiques en que »).

[19] Notons tout de même l'existence de quelques rares exceptions. Dans la phrase : « comme on fait son lit, on se couche », la subordonnée introduite par « comme » en position initiale est un adverbe exprimant la manière, que l'on trouve généralement en fin de phrase. La phrase : « comme il sonna la charge, il sonne la victoire », exemple type dans les grammaires traditionnelles, est un rare cas où la subordonnée introduite par « comme » située en position initiale a une valeur comparative. On peut trouver un autre exemple de ce type dans Wagner & Pinchon (1991, p. 541) :

> « *Comme Mademoiselle Lambercier avait pour nous l'affection d'une mère, elle en avait aussi l'autorité.* (J.-J. ROUSSEAU)
>
> [= rapport de comparaison.] »

> *Qu'on parle de l'environnement ou de la drogue, de la condition féminine ou des immigrés, de la crise urbaine ou de la gestion hospitalière, de l'échec scolaire ou des accidents de la route*, l'appel à l'État est de moins en moins suffisant (...) (POL11) »

Afin d'éviter tout jugement prématuré et non suffisamment étudié, nous laissons en suspens l'analyse exacte de cette proposition et du connecteur « que » apparaissant ici. Nous signalons seulement que si nous voulons conserver l'analyse comme complétive, nous devrons réexaminer la possibilité d'attribuer à la complétive une catégorie équivalente autre que le groupe nominal, ce sans quoi nous serions obligés de remettre en cause l'analyse comme complétive et réétudier la possibilité d'interprétation comme intégrative, par exemple.

4.6.6 Position post-nominale : subordonnée déterminante ou relative (subR)

À cette position, apparaissent non seulement les propositions adjectives (relatives, complétives), mais aussi, quoique rarement, les propositions adverbiales (intégratives adverbiales), et percontatives en « si ».

1. **adjectives** :
   a) relatives

      La peinture qui m'a fascinée

      La peinture dans laquelle notre maison était reproduite

      Ce à quoi je m'attendais

   b) complétive

      L'idée que tout est fini
2. **adverbiales** : intégratives adverbiales (rare)

   La déception du père quand il a entendu cette nouvelle
3. percontatives en « si » (rare)

   Son incertitude s'il devait obéir (repris de Le Goffic (1993a))

4.6.7 Autres positions : post-adjective et post-adverbiale

Ces positions ne concernent que les propositions en « que », complétif, corrélatif ou relatif. Elles sont généralement analysées comme des propositions introduites par une locution conjonctive. Notre description suit l'analyse faite par Le Goffic (cf. § 4.5.2).

- Post-adjective :
  1. **intégrative** (corrélative) : de même que ;
- Post-adverbiale :

1. **complétive**[20] : à moins que, loin que, cependant que, bien que, déjà que, encore que[21], même que, non que, sinon que, surtout que ;

2. **relative** : alors que, aujourd'hui que, dès lors que, maintenant que ;

3. **intégrative** (corrélative) : ainsi que, aussi longtemps que, aussitôt que, d'aussi loin que, d'autant plus que (d'autant moins que), d'autant que, plutôt que (pas plutôt que), si bien que, sitôt que, tant que.

Comme nous l'avons déjà abordé dans la section 4.4.3, les phrases du type « Si malin qu'il soit, ... » sont analysées par la règle traitant le syntagme adjectif suivi d'une proposition corrélative en fonction secondaire décrite ici, avant qu'il ne soit séparé de la proposition racine par la règle traitant les éléments extra-prédicatifs.

Tout comme dans le cas de la proposition substantive suivant une préposition, ces subordonnées utilisées seules (c'est-à-dire sans être constituant d'une locution) apparaissent rarement à ces positions. Nous gardons donc, principalement pour les propositions introduites par une locution conjonctive que l'étiqueteur n'a pas réussi à regrouper et pour quelques autres cas tels que « Si malin qu'il soit, ... », la possibilité d'une subordonnée en « que » (non constituant d'une locution) après un adverbe ou un adjectif – avec, comme indication de fréquence, rare.

4.6.8 Récapitulatif

Récapitulons maintenant toutes les unités à détecter que nous avons définies.

[20] Dans l'analyse de Le Goffic, les trois premières – dans lesquelles la complétive se rattache effectivement à l'adverbe – se distinguent clairement du reste des locutions. Dans les autres locutions, « la complétive n'est pas malgré les apparences régime de l'adverbe » (Le Goffic, 1993b, p. 92). Ces constructions sont expliquées par un mécanisme du type parataxe. « Le GAdv comportant la subordonnée, qui ne peut être qu'une complétive, est posé paratactiquement, sous la lumière modale indiquée par l'adverbe (cf. *Toujours est-il que P*) et par le mode de son verbe » (Le Goffic, 1993a, p. 416).

[21] Il est à noter que Fuchs (1992) signale l'existence de « encore que » avec « que » non complétif mais corrélatif, comme dans la phrase suivante (exemple tiré de Fuchs (*Ibid.*)).

> [...] *Elle se tord peu à peu, vacille, essaie de se redresser, puis, d'une seule masse, s'effondre dans les bras du séminariste qui la reçoit respectueusement,* ***encore que*** *s'effeuillent à son intention toutes les pivoines de la terre.* (Bazin)

Il s'agit de « encore que » à valeur circonstancielle, ayant un fonctionnement intermédiaire entre celui de la construction « encore (adv.) ... + que (corrélatif) » à valeur circonstancielle et celui de « encore + que (complétif) » à valeur notionnelle (concessive ou adversative). Mais l'auteur signale également l'instabilité de cette valeur qui glisse « facilement vers une valeur concessive, dès lors qu'un rapport oppositif peut être reconstruit entre P et Q. »

- propositions :

1. **racine** ;
2. **coordonnée** ;
3. **détachée-insérée** ;
4. **subordonnée** : introduite par un connecteur de subordination à un endroit déterminé et caractérisée par sa fréquence (cf. tableau 4.5) :
   a) subordonnée complément en Qu- (subQ) apparaissant en position post-verbale ;
   b) subordonnée SN (subSN) apparaissant à une autre position pouvant être occupée par un SN ;
   c) subordonnée circonstancielle ou périphérique (subP) apparaissant non seulement en positions initiale et finale mais aussi insérée à différentes places sous forme détachée ;
   d) subordonnée déterminante ou relative (subR) apparaissant en position post-nominale ;
   e) subordonnée apparaissant en positions post-adverbiale et post-adjective.

- éléments extra-prédicatifs

√ = fréquent
△ = moins fréquent ou rare

		post-V	**post-N**	**Int/Fin**	**autres SN**			**autres**	
					sujet	int.	prép.	adj.	adv.
sub.	intég-pro.	△			△	△	△		
	perc. *si*	√	△		△	△	△		
	compl.	√	√	√	△	△	△		△
adj.	relatives *que, où*		√						△
adv.	intég-adv.		△	√					
	où			△					
	que			*1)				△	△

[1] les intégratives en « que » n'apparaissent qu'en position finale.

TAB. 4.5 – Caractérisation des subordonnées par catégorie, position et fréquence

4.7 Notre typologie des connecteurs

4.7.1 Étiquettes classiques et avantages de la redéfinition d'un nouvel ensemble

Nous avons déjà abordé dans la section 4.2 le problème lié à la difficulté d'attribution des étiquettes classiques et nous avons remis en question la nécessité d'une telle distinction, difficilement réalisable par l'analyse limitée des étiqueteurs.

Nous présentons maintenant notre typologie des connecteurs définie sur la base du classement des subordonnées décrit précédemment.

4.7.2 Typologie des connecteurs basée sur la position d'apparition de la subordonnée

Avant de décrire notre typologie des connecteurs, catégorisons-les d'abord selon les types de subordonnées – définis par Le Goffic – qu'ils peuvent introduire (cf. tableau 4.6).

	Percontative	Intégrative		Complétive	Relative
		pronominale	adverbiale		
qui	√	√			√
que (pro.)				√	√
(adv.)			√		
dont					√
où	√		√		√
quand	√		√		
comme	√		√		
si	√		√		
quoi*	√				√
lequel*	√				√
quel	√				
combien	√				
comment	√				
pourquoi	√				

* ainsi que toutes leurs formes fléchies.

TAB. 4.6 – Connecteurs du français

En nous basant sur l'étude des positions d'apparition des subordonnées décrite dans la section précédente, nous avons réalisé une classification des connecteurs, mots en « qu- ». Le tableau 4.7 page suivante présente la synthèse de cette étude.

À partir de ce constat, nous avons défini les quatre types de connecteurs suivants :

1. connecteurs isolés :
   qui, que, dont, où

✔ = fréquent
△ = rare

	Position	post-V				post-N				Int /Fin				pos. SN				Autres			
		I	C	P	R	I	C	P	R	I	C	P	R	I	C	P	R	I	C	P	R
C. isolés	qui	△		✔					✔					△		△					
	que		✔				✔		✔		✔				△			△	△		△
	dont								✔												
	où			✔		?			✔	△						△					
C.amb.	quand			✔		△				✔						△					
	comme			✔		△				✔						△					
	si			✔				△		✔						△					
C. rel.	quoi			✔					✔							△					
	lequel			✔					✔							△					
Indicat. prop.	quel			✔												△					
	combien			✔												△					
	comment			✔												△					
	pourquoi			✔												△					

I = Intégrative, C = Complétive, P = Percontative, R = Relative

TAB. 4.7 – Typologie des connecteurs

ayant un comportement particulier et dont les positions d'occurrence ne sont comparables avec aucun des autres connecteurs ;

2. connecteurs ambigus :
   quand, comme, si
   apparaissant fréquemment aux deux positions (post-V, Int/Fin) et rarement aux deux positions (post-N, SN) ;

3. connecteurs relatifs :
   quoi, lequel (et ses formes fléchies)
   apparaissant fréquemment aux deux positions (post-V, post-N) et rarement aux positions SN ;

4. indicateurs de propositions :
   quel (et ses formes fléchies), **combien, comment, pourquoi**
   apparaissant fréquemment seulement en position post-V et rarement aux positions SN.

4.7.3 Connecteurs composés

« Quel » (et ses formes fléchies), « combien (de) » et « lequel (de) » (et ses formes fléchies), qui constituent parfois un syntagme avec un nom, doivent être traités différemment des autres. Lorsque ces connecteurs fonctionnent comme compléments secondaires du substantif qui les suit, nous les considérons comme connecteurs déterminants et considérons qu'ils constituent avec le syntagme nominal qui les suit un connecteur composé.

Par exemple, dans la phrase « combien d'habitants compte Tokyo », « combien d' » est un connecteur déterminant qui constitue avec le syntagme nominal qui le suit, « habitants », un connecteur composé « combien d'habitants ». L'analyse est similaire pour « lequel de ces romans » ou « quel chemin ».

En revanche, lorsqu'ils fonctionnent tout seuls comme dans la phrase : « combien coûtait cette bêtise » ou « quel était son intérêt », nous considérons qu'ils constituent un connecteur tout seuls.

4.8 Problèmes généraux de la détection des propositions

Nous examinons maintenant les problèmes généraux liés à l'opération de détection automatique des propositions.

Nous allons tout d'abord traiter les problèmes liés aux symboles de ponctuation susceptibles de marquer la frontière de propositions mais aussi de provoquer une erreur d'interprétation. Nous aborderons ensuite l'ambiguïté de rattachement des syntagmes en fin de phrase, souvent ambigus. Nous examinerons également deux autres problèmes liés aux deux types de structures où apparaissent particulièrement des ambiguïtés de rattachement ou des problèmes liés à l'ellipse, difficiles – voire impossibles – à résoudre sans contexte : les structures à dépendance à distance et les structures de coordination.

Nous ne pouvons, bien entendu, proposer aucune solution à ces questions, mais présentons tout de même l'existence de ces cas d'un point de vue théorique. Il serait intéressant d'étudier, d'un point de vue pratique, la conséquence de ces problèmes dans de futurs travaux afin d'essayer de trouver une piste prometteuse.

4.8.1 Problèmes liés aux symboles de ponctuation

Comme on peut le constater dans la définition des éléments extra-prédicatifs ou des propositions détachées-insérées ou encore des propositions circonstancielles insérées, nous accordons de l'importance aux symboles de ponctuation, tels que les virgules, les parenthèses et les tirets, qui sont souvent complètement ignorés dans beaucoup de grammaires formelles.

Beaucoup de structures ne peuvent pas être interprétées correctement sans interprétation correcte des virgules utilisées. À tel point que l'absence d'une virgule peut entraîner des ambiguïtés comme le montre Fuchs (1996, p. 110) :

> « Au sein d'un texte, par ailleurs ponctué, il suffit parfois de l'absence d'une virgule pour que la segmentation de la phrase en propositions devienne problématique :
>
> > *Quant à la réforme fiscale, on se demande qui en veut vraiment : "Les élus en parlent tant qu'ils n'ont pas à la voter" a dit le ministre.*
> > "Les élus en parlent tant [= tellement], qu'ils n'ont pas à la

> voter" / "Les élus en parlent, tant qu' [= aussi longtemps que] ils n'ont pas à la voter" »

Il faut cependant noter leur fiabilité également restreinte, notamment celle de la virgule, comme indicateurs syntaxiques. Le Goffic définit, dans la section consacrée à la ponctuation de Le Goffic (1993a), la virgule comme séparateur faible, du fait de son caractère polysémique et la qualifie de « séparateur à tout faire ».

Mais s'ils sont polysémiques, c'est bien qu'ils ont un/des sens. Nous devons, lors de l'analyse, non pas les ignorer totalement, mais explorer les indices que laissent ces symboles autant que possible et ce le plus correctement possible. Nous les considérons donc comme des indicateurs secondaires importants et en profitons dans les cas où la frontière indiquée par la virgule est relativement fiable.

Nous examinerons l'influence de l'importance accordée à ces symboles – notamment les virgules – sur les résultats de l'analyse automatique lors de l'évaluation du système dans la section 9.3.

4.8.2 Ambiguïté du rattachement des éléments en fin de phrase

Dans certains cas – avec ou sans détachement –, la phrase étant ambiguë, les éléments en fin de phrase peuvent être interprétés aussi bien comme des constituants de la subordonnée que de la racine.

On peut trouver des exemples de ce type dans l'ouvrage de Fuchs (1996) consacré aux problèmes des ambiguïtés :

1. *Il a dit qu'il donnerait son avis par fax.*
   - Il a dit [qu'il donnerait son avis par fax]
   - Il a dit [qu'il donnerait son avis] par fax
2. *Au zoo, on peut voir un lion qui terrifie les badauds et de pauvres petites antilopes.*
   - ... voir un lion [qui terrifie les badauds et de pauvres petites antilopes]
   - ... voir un lion [qui terrifie les badauds] et de pauvres petites antilopes

Ce type d'ambiguïté est malheureusement impossible à traiter de façon automatique. Il nous faudra donc décider de la position à adopter face à de telles ambiguïtés.

Dans un autre cas, le rattachement des éléments en fin de phrase peut être correctement analysé par l'introduction d'informations supplémentaires plus ou moins complexes, telles que celles sur la structure argumentale du verbe. Toutefois, sur le plan pratique, l'ajout de ce type d'information risque de démultiplier les calculs.

Dans le cadre de la présente thèse, nous ne considérons les propositions comme des subordonnées détachées-insérées que lorsqu'elles sont bien entourées et détachées par deux symboles de même type. Ainsi, dans le cas des exemples ambigus cités précédemment, les éléments en fin de phrase seraient analysés, non pas comme des compléments discontinus du verbe principal, mais comme des éléments de la subordonnée.

4.8.3 Structures à dépendance lointaine

Considérons d'abord une phrase ayant une relative dite « longue » ou « imbriquée ».

Ce philosophe **qu'**il faut **que** vous lisiez est très connu.

Delaveau (2001, p. 111) explique cette construction ainsi :

> « Comme dans les relatives avec *dont* et pronom, il y a une complétive dominée par la relative, [...], dans les cas de relatives longues, il y a un vide dans la complétive, lequel a la fonction que requiert le pronom relatif en tête. »

Cette dépendance à distance apparaît non seulement dans les relatives, mais aussi dans les percontatives, ainsi que dans les structures clivées :

- Dites-moi avec qui vous croyez que vit Marie.
- C'est à Marie que je veux que tu parles.

Il est à noter que les structures déclaratives posent un problème d'appartenance du circonstant. La phrase suivante reprise de Fuchs (1996) :

Lundi prochain, fais-moi penser qu'il faudra relire le manuscrit.

peut être interprétée comme

- Lundi prochain, fais-moi penser [A] + qu'il faudra relire le manuscrit (= A)
- fais-moi penser [A] + Lundi prochain, ... qu'il faudra relire le manuscrit (= A)

Dans une structure dans laquelle l'antécédent (ou l'élément extrait dans la structure clivée) a fonction de circonstant, cet antécédent peut être raccroché soit au prédicat de la relative ou de la percontative, soit au prédicat de la complétive enchâssée. Ainsi, dans la phrase (Fuchs, 1996) :

> Montre-moi l'endroit où tu as dit qu'il fallait chercher. (relative)
> Montre-moi à quel endroit tu as dit qu'il fallait chercher. (percontative)

les deux interprétations sont possibles :

- ... l'endroit de ta déclaration selon laquelle il fallait chercher.
- ... l'endroit de la recherche nécessaire, selon tes dires.

4.8.4 Structures de coordination

Outre la difficulté de représentation hiérarchique que nous avons abordée dans la section 4.4.2, la structure de coordination est intimement liée, comme le disent Fuchs & Victorri (1993b), au phénomène très délicat et difficile à traiter qu'est l'ellipse.

L'ellipse est « définie comme "l'effacement" de constituants *a priori* obligatoires » (Fuchs et Victorri, *Ibid.*). Par exemple, souvent, les propositions subordonnées coordonnées partagent des éléments constituants, comme le COD « leurs pains » dans la phrase suivante :

Les boulangers **qui** préparent et **qui** vendent leurs pains.

Dans cette phrase, la première proposition subordonnée est-elle terminée seulement par « préparent » ou partage-t-elle « leurs pains » avec la seconde ? Il est plus naturel de considérer que le syntagme « leurs pains » joue à la fois la fonction de COD pour la première et la seconde proposition sans pour autant être marqué à chaque fois.

L'ellipse a une lourde conséquence pour la détection des propositions. Étudions chaque cas d'ellipse en fonction du type d'élément non exprimé.

Omission d'un complément

C'est le cas de l'exemple précédent :

Les boulangers **qui** préparent et **qui** vendent leurs pains.

Pour nos travaux de détection des propositions, l'idéal serait sans doute d'indiquer la présence d'un complément non seulement dans la proposition précédant directement le complément, mais également dans la proposition éloignée.

Toutefois, la détermination du complément commun ou non est également délicate. Dans certains cas, la détermination est réalisable par consultation de l'identité lexicale des verbes concernés. Mais dans d'autres cas, elle est impossible comme dans les exemples présentés dans Fuchs (1996) et repris ci-dessous :

Il regarde et il admire Marie.
- Il regarde + et il admire Marie
- Il regarde (Marie) + et il admire Marie

Il est venu et il est reparti avec tristesse.
- Il est venu + et il est reparti avec tristesse
- Il est venu (avec tristesse) + et il est reparti avec tristesse.

Nous abandonnons, du moins dans le cadre de la présente thèse, la détermination des éléments communs et tenons compte uniquement des éléments explicites. La reconnaissance est donc réalisée en extrayant simplement la première subordonnée délimitée par un connecteur quelconque de la seconde subordonnée, sans nous préoccuper des problèmes de l'appartenance des compléments suivant la deuxième subordonnée. Cette représentation est peu cohérente dans le cas où le complément appartient effectivement à la première subordonnée aussi. Nous évaluerons les conséquences de ce choix dans les résultats d'alignement.

Omission du sujet

Nous constatons des problèmes similaires dans les structures de coordination de verbes comme :

Il **achète et revend** des vieux meubles.

Dans ce cas, la question concerne non seulement le complément mais aussi le sujet.

On peut donc considérer que dans l'exemple du paragraphe précédent « Les boulangers qui préparent et qui vendent leurs pains », le complément – commun

aux deux propositions – est omis dans la première proposition et que pour le dernier exemple, le complément « des vieux meubles » est omis dans la première proposition et le sujet « il » dans la seconde proposition.

La détection de la proposition subordonnée dans laquelle le sujet est omis est réalisable en posant comme condition minimale la présence d'un verbe fini et éventuellement d'un connecteur. Même l'indication du sujet implicite semble envisageable. Mais dans un premier temps nous appliquons au sujet implicite la même règle qu'avec le COD implicite, à savoir la représentation uniquement avec des éléments explicites.

Omission du verbe

Le problème de l'ellipse peut se rapporter non seulement aux compléments, mais aussi aux verbes.

Mon père est français et **ma mère japonaise**.

Les structures des propositions participiales entraînent également l'ellipse de « étant », produisant ainsi des sous-phrases nominales.

Exemples (tirés de Le Goffic (1993a)) :

– Cette affaire (étant) terminée, nous pouvons penser à la suite.
– Nous réglerons cette question le moment venu.
– Il est tombé la tête la première.

La détection des propositions dans le cas où le verbe est omis est plus délicate que dans les deux cas précédents, car nous ne pouvons plus utiliser comme repère la présence d'un verbe fini. La reconnaissance de ces propositions coordonnées sans verbe nécessite beaucoup de calculs comme pour la détection des constructions détachées à prédication seconde, que nous avons abordée dans la section 4.4.3. Tout en ayant conscience de leur importance, nous laissons de côté, dans le cadre de la présente thèse, leur détection qui demanderait une analyse fine de la structure de l'ensemble de la phrase.

4.9 Grammaire pour la détection des propositions

Les études linguistiques présentées jusqu'ici nous ont permis de définir une grammaire pour la détection des propositions de type CFG (*Context-Free Grammar*, cf. § 9.1) qui permet non seulement de reconnaître les frontières des propositions, mais aussi d'analyser leurs relations syntaxiques.

Notre grammaire ainsi conçue pour la détection des propositions est présentée dans l'annexe § B.

Cette grammaire – qui se base sur notre typologie des propositions, définie de sorte que leur identification ne nécessite pas la détermination correcte de la nature des éléments introducteurs – présente comme avantage la non-dépendance à l'analyse correcte des connecteurs, tâche très difficile à réaliser.

Elle est également caractérisée par l'importance accordée aux symboles de ponctuation, tels que les virgules, les parenthèses et les tirets, qui sont souvent complètement ignorés en dépit de la présence non négligeable d'indices sur la structure de la phrase, que nous laissent ces symboles.

CHAPITRE 5

NOTIONS PRÉLIMINAIRES DE LINGUISTIQUE JAPONAISE

Nous présentons dans ce chapitre les quelques notions de base qui nous semblent indispensables pour une discussion sur tout sujet traitant de linguistique japonaise dans une optique de traitement automatique, en particulier d'analyse syntaxique. Nous allons tout d'abord présenter brièvement les principaux travaux sur lesquels nous nous basons (§ 5.1) avant d'aborder les unités linguistiques de l'écrit (§ 5.2), les catégories des mots (§ 5.3) et les variations de forme des mots variables (§ 5.4). Nous examinerons ensuite les éléments constituant la phrase japonaise (§ 5.5) avant d'étudier l'ordre des mots (§ 5.6) et les moyens d'indication de la fonction syntaxique (§ 5.7). Enfin, la dernière partie du chapitre (§ 5.8) sera consacrée à l'exposé de la structure de la subordination déterminante dans la phrase japonaise.

5.1 Fondement des études

L'exposé est basé sur une version largement retravaillée du chapitre « Notions de linguistique japonaise » de nos travaux antérieurs (Nakamura-Delloye, 2003a).

Pour introduire l'ensemble des notions de base et pour repérer différentes problématiques existantes, nous allons utiliser la grammaire dite scolaire (学校文法, *gakkô-bunpô*) – grammaire basée sur la théorie de Hashimoto (1934), que les Japonais apprennent aujourd'hui à l'école – qui servira de point de départ à toutes nos discussions.

Nous allons également nous référer à des théories reconnues « classiques » constituant la base des travaux linguistiques contemporains proposés par de grands linguistes japonais, parmi lesquels ÔTSUKI Fumihiko, YAMADA Yoshio,

MATSUSHITA Daizaburô, HASHIMOTO Shinkichi, TOKIEDA Motoki, SAKUMA Kanae, MIO Isago ou MIKAMI Akira. Nos travaux s'appuient essentiellement sur ceux de ce dernier, Mikami, notamment pour les problèmes liés à la définition des éléments constituants de la phrase japonaise.

Nous nous appuyons bien entendu sur les travaux récents des linguistes contemporains tels que Teramura, Minami, et en particulier sur la grammaire publiée par Masuoka & Takubo (1992), grammaire très utilisée aujourd'hui aussi bien dans le domaine de la linguistique que dans celui du TAL au Japon.

5.2 Unités linguistiques de l'écrit

Nous présentons d'abord les unités élémentaires (§ 5.2.1). Nous aborderons ensuite d'une manière un peu plus détaillée les deux unités, mot (§ 5.2.2) et syntagme minimal (§ 5.2.3).

5.2.1 Unités élémentaires

Dans les travaux de linguistique japonaise, sont utilisées des unités pour lesquelles on trouve facilement un équivalent dans une grammaire classique du français : 文章 (*bunshô*, **texte**), 段落 (*danraku*, **paragraphe**), 文 (*bun*, **phrase**), 節 (*setsu*, **proposition**) et 単語 ou 語 (*tango ; go*, **mot**).

La définition de ces unités en japonais – notamment la phrase, la proposition et le mot – a fait couler beaucoup d'encre tout comme dans la linguistique française. Nous consacrerons un peu plus tard quelques pages à la définition de ces deux premières unités, la phrase et la proposition, qui concernent le plus nos présents travaux. Nous nous contentons dans cette section de présenter en quelques mots la définition usuelle de l'unité mot et les problèmes de segmentation en mots de la phrase japonaise.

5.2.2 Problèmes liés à la définition du mot

Difficultés de la définition

La grammaire usuelle définit le mot comme la plus petite unité constituant la phrase[1] et il est souvent comparé à une pièce détachée de l'ensemble qu'est la

[1] Cette définition du mot japonais peut évoquer celle du morphème. En effet, certaines unités japonaises équivalentes aux morphèmes non autonomes dans d'autres langues, telles que les terminaisons ou les particules de cas servant de relateur casuel, font partie des mots, 単語 (*tango*). Comme nous allons le voir, seules les unités qui interviennent dans la dérivation des mots sont considérées comme des unités n'appartenant pas aux mots. La divergence entre les mots japonais et les mots dans les langues telles que l'anglais ou le français est remarquée par les linguistes japonais et certains tels que Sakakura (1979) considèrent plutôt les unités *bunsetsu* (cf. 5.2.3) comme équivalentes aux mots dans ces langues européennes. Néanmoins, le parallélisme des mots anglais ou français avec les *bunsetsu* japonais est également trop simpliste, dans la mesure où certains *bunsetsu* correspondent non pas à un mot, mais à un syntagme prépositionnel en français.

phrase. Par ailleurs, il est également considéré comme l'unité de mémoire dans le cerveau (Hayashi et al., 1988). Cependant, tous ces propos ne donnent en fait aucun critère concret sur l'étendue d'un mot, et la frontière entre les morphèmes et les mots varient souvent selon les théories – en particulier pour certains types comme les auxiliaires dits *jodôshi* et les particules.

Ces deux types de mots sont des éléments non autonomes qui, suivant toujours un mot autonome, marquent sa fonction syntaxique, ou ajoutent une modalité ou une valeur énonciative. Certains considèrent les auxiliaires et même, bien que plus rarement, les particules comme des unités n'appartenant pas aux mots. Dans la grammaire scolaire, ces deux catégories sont incluses dans les mots, et sont distinguées des suffixes et des préfixes. Seules les unités qui interviennent dans la dérivation des mots sont considérées comme des *setsuji* (接辞, affixes), unités n'appartenant pas aux mots.

Problèmes de la segmentation en mots de la phrase japonaise

Contrairement au français pour lequel les problèmes de segmentation en mots de la phrase se résument notamment à la reconnaissance des mots discontinus séparés par un/des séparateur(s) graphique(s) mais qui constituent une même unité, dans le cas du japonais où il n'existe presque aucun séparateur, la question se pose d'abord dans le sens inverse : où doit-on tracer la frontière des mots dans une séquence entièrement continue ?

C'est seulement après cette première segmentation que nous sommes confrontés à des problèmes semblables à ceux du français, à savoir la reconstitution des mots composés de plusieurs unités susceptibles d'être chacune considérée comme unité indépendante, mais qui constituent dans un contexte spécifique une seule unité.

5.2.3 Unité *bunsetsu*

Outre les unités élémentaires présentées précédemment, les Japonais utilisent souvent une unité appelée 文節 (*bunsetsu*).

La notion de 文節 (*bunsetsu*) provient de la théorie de Hashimoto (1934). Il définit cette unité comme la première unité que l'on obtient en segmentant une phrase et qui peut être un constituant de phrase. Il dit également que c'est le plus petit élément obtenu en segmentant au maximum une phrase, tout en conservant le statut de langue de cet élément. Les *bunsetsu* sont caractérisés, sur le plan formel, par la présence de coupures de syllabes immédiatement avant et après eux.

桜　の｜花　は｜もう｜散った。
(*sakura* - *no* | *hana* - *wa* | *mô* | *chitta*)
(cerisier - de | fleur - [thème] | déjà | tomber [passé])
« Les cerisiers ont déjà perdu leurs fleurs. »

Par ailleurs, Garnier (1982) utilise le terme « segment minimal » qui désigne « le plus petit ensemble pouvant remplir une fonction syntaxique dans l'énoncé ».

Bien que le segment minimal de Garnier ne corresponde pas exactement à cette unité de Hashimoto, nous traduisons *bunsetsu* par **segment minimal** ou **syntagme minimal**.

5.3 Catégorisation des mots japonais

La catégorisation des mots de la grammaire scolaire, comme d'autres catégorisations proposées ailleurs, est largement critiquée. Cependant, nous ne pouvons proposer pour le moment aucune autre catégorisation semblant plus adéquate pour nos travaux. Nous adoptons donc cette catégorisation classique mais la réétudions chaque fois que nous en sentons le besoin comme nous allons le présenter dans cette section.

On classe dans la grammaire scolaire les mots en dix catégories grammaticales (cf. figure 5.1, reprise de Hayashi et al. (1988) et traduite en français). Les mots sont d'abord divisés en deux grandes classes : 自立語 (*jiritsugo*, mot autonome) et 付属語 (*fuzokugo*, mot annexe). Les *jiritsugo* sont définis comme des mots pouvant

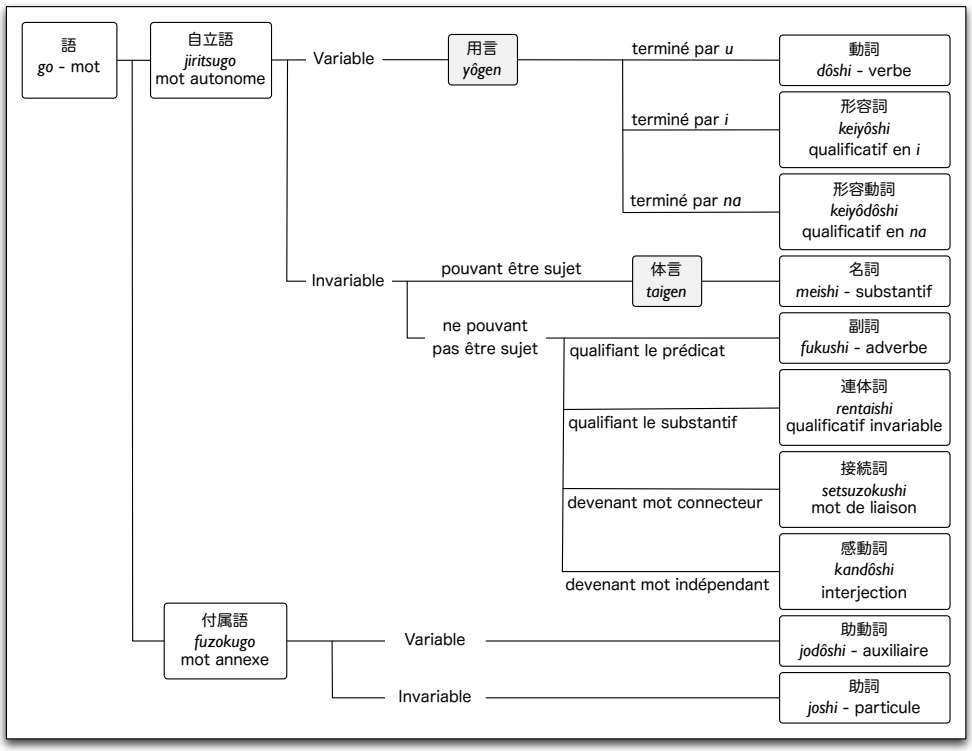


FIG. 5.1 – Catégorisation des mots dans la grammaire scolaire

constituer à eux seuls un syntagme minimal *bunsetsu*. Les *fuzokugo* sont définis selon Hashimoto comme des mots qui ne sont pas autonomes et qui sont toujours utilisés avec des mots qui doivent être autonomes.

Nous allons d'abord étudier les sous-catégories de *jiritsugo* (§ 5.3.1), puis celles de *fuzokugo* (§ 5.3.2).

5.3.1 Sous-catégories de *jiritsugo*

On distingue d'abord deux types de *jiritsugo*, variables et invariables.

Les mots autonomes variables sont traditionnellement appelés 用言 (*yôgen*). Les *yôgen* sont caractérisés par le fait qu'ils sont susceptibles d'être prédicat. Cette notion de *yôgen* s'oppose à celle de 体言 (*taigen*), qui désigne les unités susceptibles d'être sujet et qui correspond à la catégorie substantif.
On distingue dans la grammaire scolaire trois catégories de *yôgen* :

- unité exprimant l'action, l'effet ou l'existence :
  1. unité dont la forme de base se termine par *-u* : 動詞 (*dôshi*, verbe) ;
- unité exprimant la nature ou l'état :
  2. unité dont la forme de base se termine par *-i* : 形容詞 (*keiyôshi*, qualificatif en *i*) ;
  3. unité dont la forme de base se termine par *-da* ou *-desu* : 形容動詞 (*keiyôdôshi*, qualificatif en *na*).

Les mots autonomes invariables sont classés dans la grammaire scolaire en cinq catégories :

- unité pouvant être sujet, 体言 (*taigen*) :
  1. unité désignant un objet ou un évènement : 名詞 (*meishi*, substantif) ;
- unité pouvant qualifier une autre unité :
  2. unité qui qualifie les *yôgen* : 副詞 (*fukushi*, adverbe) ;
  3. unité qui qualifie les *taigen* : 連体詞 (*rentaishi*, qualificatif invariable) ;
- unité pouvant être mot de liaison :
  4. 接続詞 (*setsuzokushi*, mot de liaison) ;
- unité pouvant être énoncé indépendant :
  5. 感動詞 (*kandôshi*, interjection) ;

Nous renonçons au terme *yôgen* (qui désigne les mots autonomes variables). En effet, bien qu'il soit caractérisé par la possibilité d'être prédicat, n'en fait pas partie le substantif, capable également d'assurer le rôle de prédicat à l'aide de la copule – voire même parfois sans elle. Si bien que nous utilisons le terme **mot prédicatif** pour désigner non seulement les mots autonomes variables mais aussi le substantif lorsqu'il assure la fonction de prédicat dans la phrase.

5.3.2 Catégories de *fuzokugo*

Deux types de *fuzokugo* : particule et auxiliaire

Les *fuzokugo* sont divisés en deux catégories : ceux qui sont variables et ceux qui sont invariables. Les premiers désignent l'ensemble des 助動詞 (*jodôshi*), auxiliaires, et les seconds sont des particules, dits 助詞 (*joshi*).

Mais cette définition n'est pas cohérente avec la réalité : il existe des auxiliaires invariables que Kindaichi (1953) appelle 不変化助動詞, (*fuhenka-jodôshi*, auxiliaires invariables).

Avant l'examen de leur véritable différence, nous étudions les sous-catégories de particules dont le regroupement lui-même est parfois remis en question.

Sous-catégories de particules

Les particules regroupent différents types d'éléments. Si bien que certains tels que Okutsu et al. (1986) proposent même d'abandonner la catégorie « particule » en définissant à la place différentes classes plus précises. Sans aller jusqu'au renoncement total, les grammaires définissent généralement des sous-classes, mais la catégorisation varie d'une grammaire à l'autre.

Hashimoto (1969) définit 9 types :

1. 格助詞 (*kaku joshi*, particule de cas) ;
2. 間投助詞 (*kantô joshi*, particule interjective) ;
3. 終助詞 (*shû joshi*, particule finale) ;
4. 連体助詞 (*rentai joshi*, particule précédant le nom ou particule déterminante) ;
5. 係助詞 (*kakari joshi* ou *kei joshi*, particule relationnelle) ;
6. 副助詞 (*fuku joshi*, particule adverbiale) ;
7. 接続助詞 (*setsuzoku joshi*, particule conjonctive) ;
8. 準用助詞 (*jun'yô joshi*, particule assimilée à un *yôgen*) ;
9. 並立助詞 (*heiritsu joshi*, particule de coordination).

Les classes *kaku joshi* (particule de cas), *kantô joshi* (particule interjective) et *shû joshi* (particule finale), assez faciles à distinguer des autres, constituent les catégories les plus stables définies dans toutes les grammaires (même si sous des noms différents) .

Les *kaku joshi* (particule de cas) introduisent des syntagmes nominaux dans la phrase et servent de relateur casuel[2]. La description détaillée sera présentée plus loin (§ 5.7), et nous ne citons ici que les deux principaux : la particule *ga* usuellement considérée comme indicateur du nominatif (ou du sujet), et *wo* celui de l'accusatif (ou du complément d'objet direct).

[2]Certains comme Tamba & Terada (1991) soulignent la divergence entre les relations casuelles et celles que ces particules japonaises établissent entre le prédicat et le syntagme nominal qu'elles introduisent. Nous n'entrons pas dans les détails de ce problème dans le cadre de la présente thèse.

Malgré la stabilité de son statut, pour les membres de la catégorie des particules de cas, la définition varie parfois selon les linguistes. La particule de citation *to* (引用助詞, *in'yô joshi*), classée souvent dans la catégorie des particules de cas, constitue à elle-seule une classe dans la grammaire de Masuoka & Takubo (1992).

Il arrive également que la particule *no*, constituant la catégorie de particule déterminante, soit considérée comme une particule de cas comme l'a fait Mikami dans certains de ses travaux (voir note 12 page 209). Le traitement de cette particule est en effet assez varié. Alors que Teramura (1982b) distingue, comme Hashimoto, cette particule reliant deux mots des autres particules conjonctives qui, elles, servent à connecter deux propositions, Masuoka et Takubo la classent dans la catégorie des particules conjonctives.

Quant au reste des catégories, la définition est plus complexe, souvent floue, et le statut même de particule est souvent remis en cause pour certaines d'entre elles.

Comme le remarque Okutsu dans l'introduction de Okutsu et al. (1986), *kakari joshi* (particule relationnelle) et *fuku joshi* (particule adverbiale), termes largement utilisés depuis les travaux de Yamada et de Hashimoto, n'ont toutefois aucune définition précise et satisfaisante. D'ailleurs, les grammaires utilisées avant le niveau collège ne font pas la distinction entre ces deux catégories et les regroupent sous le nom de *fuku joshi* (particule adverbiale).

Dans les travaux contemporains, la catégorie 提題助詞 (*teidai joshi*, particule de thématisation) est de plus en plus reconnue. La particule de thématisation, *teidai joshi* est une notion apparue depuis les travaux de Matsushita et de Sakuma. Elle sert, comme son nom l'indique, à la thématisation d'un élément. La particule la plus représentative est la particule *wa*, qui, dans la grammaire de Hashimoto, appartient aux particules *kakari joshi*.

La catégorie de 取り立て助詞 (*toritate joshi*, particule de mise en relief) est également une catégorie de plus en plus adoptée par les grammairiens contemporains[3]. Elle comprend les particules classées dans les catégories *fuku joshi*, particule adverbiale, et/ou *kakari joshi*.

Nous suivons *grosso modo* la catégorisation de Hashimoto, excepté pour deux points : nous adoptons la catégorie de particule de thématisation que nous aborderons plus en détail (cf. § 6.4.5) ; sans faire de distinction entre les deux catégories de particules non bien définies, *kakari joshi* (particule relationnelle) et *fuku joshi* (particule adverbiale), nous les regroupons dans la classe *fuku joshi* (particule adverbiale) – les questions liées à ces catégories posant des problèmes directs dans nos travaux seront abordées de manière plus détaillée (cf. § 7.7).

Deux autres unités non autonomes

Il existe deux autres unités non autonomes qui ne font pas partie des mots : 接辞 (*setsuji*), affixes, intervenant dans la dérivation, et 活用語尾 (*katsuyô gobi*),

[3]Pour un état de l'art, voir Numata (1986).

terminaisons, qui constituent avec le radical les différentes formes des mots variables. La détermination de la frontière entre les mots non autonomes et ces unités inférieures est parfois très délicate. Étant donné leur caractère non autonome, certains comme Yoshikawa (1989)[4] considèrent aussi les *jodôshi*, auxiliaires, comme des unités n'appartenant pas aux mots.

Mais, les linguistes définissent généralement encore aujourd'hui les particules et les auxiliaires comme des mots. La grammaire de Masuoka-Takubo distingue les *fuzokugo*, les suffixes et les terminaisons, en définissant :

– comme des auxiliaires ou des particules les unités suivant le mot prédicatif à la forme autonome (qui peut constituer à lui-seul le prédicat de la phrase, voir § 5.4), tels que (mot souligné) :

  書く　だろう
  (*kaku* - *darô*)
  (écrire [forme autonome] - [conjecture ou invitation])

– comme des suffixes les unités constituant avec le mot prédicatif un autre mot, qui subit sa propre variation de forme :

  書き　たい
  (*kaki* - *tai*)
  (écrire [forme neutre] - [souhait])

– comme des terminaisons les éléments suivant le radical d'un mot prédicatif, qui n'appartiennent à aucune de ces deux catégories et qui constituent avec le radical une forme, tels que :

  書い　た
  (*kai ta*)
  (écrire [passé]).

Frontière floue

Nous avons vu, jusqu'ici, deux définitions valables (que nous adoptons) liées aux deux types de *fuzokugo*, particules et auxiliaires :

– les mots non autonomes invariables sont des particules (mais tous les auxiliaires ne sont pas variables) ;
– les auxiliaires sont ceux qui suivent directement une forme autonome des mots variables (et les éléments suivant une forme non autonome sont des suffixes).

Avec ces deux définitions, les deux extrémités sont bien définies mais il reste au milieu une zone floue (cf. tableau 5.2 page suivante).

On dit également que les auxiliaires suivent les mots variables, et que les particules suivent non seulement les mots variables mais aussi les substantifs. Mais les auxiliaires apparaissent en réalité également après les substantifs, et ce de manière non rare comme le dit Mikami (1955). L'examen des cas où les mots non

[4] Voir aussi `http://homepage3.nifty.com/taketoki/`.

Qui suit une forme autonome			Qui suit une forme non autonome	
variable	invariable			variable
	non dérivation		dérivation	
auxiliaire	***particule auxiliaire substantif formel***	***particule***	***suffixe***	

frontière floue

TAB. 5.2 – Frontière floue entre les particules et les auxiliaires

autonomes suivent un substantif ne nous fournirait que le même type de zone confuse.

Les mots appartenant à cette zone mal organisée concernent cependant étroitement la définition des propositions. Nous retravaillerons de manière plus poussée les mots suivant une forme autonome du mot variable – contexte plus lié à nos travaux sur la proposition – dans la section 7.6 afin de réaliser une catégorisation permettant une meilleure définition de la proposition.

Nouvelle catégorie : copule

Nous introduisons en outre une nouvelle classe **copule** (コピュラ, *kopyura*) que reconnaissent la plupart des travaux contemporains. Dans la phrase japonaise, elle constitue le prédicat en suivant un substantif ou un qualificatif en *na*. On l'appelle également 判定詞 (*hanteishi*, mot de jugement). Dans la grammaire scolaire, elle appartient à une sous-catégorie de *jodôshi*, dite 断定の助動詞 (*dantei no jodôshi*), *jodôshi* d'affirmation.

5.4 Variation de forme des mots variables

Les mots variables japonais changent de forme selon leur fonction syntaxique et selon la modalité et le temps. La décomposition et l'analyse des différentes formes des mots variables changent souvent selon les grammairiens : notre définition est basée sur les travaux de Teramura[5] et sur la grammaire de Masuoka & Takubo (1992).

Suivant Masuoka et Takubo, nous considérons comme des particules ou des auxiliaires les éléments suivant les mots variables à certaines formes, capables de constituer tout seuls le prédicat principal d'une phrase. Nous définissons comme des affixes les éléments formant avec leurs radicaux une autre unité autonome,

[5]La définition de Teramura est basée sur celles de Sakuma (1940a), Bloch (1946) et Mikami (1970).

verbe ou qualificatif. Les autres éléments sont considérés comme des terminaisons.

Nous étudions d'abord la variation des verbes (§ 5.4.1) et des qualificatifs (§ 5.4.2) avant de parler de celle des autres catégories (§ 5.4.3).

5.4.1 Verbes

Le tableau 5.3 présente les systèmes de variation des verbes.

Système	Radical	Neutre	Autonome	Condition	Volitive	Impérative
de base	ik	行き *iki*	行く *iku*	行けば *ikeba*	行こう *ikô*	行け *ike*
en *ta*	it	行って／行ったり *itte/ittari*	行った *itta*	行ったら *ittara*	行ったろう *ittarô*	– –

TAB. 5.3 – Verbe *iku* (aller)

Deux systèmes : de base et en *ta*

Tous les verbes possèdent deux types de système de variation de forme : système de base et système en *ta*. Certains verbes – dits verbes *go-dan* ou vocaliques – ont deux radicaux différents pour ces deux systèmes.

Chaque système a cinq formes – plus ou moins utilisées – représentant une modalité ou une fonction différente. Ces deux systèmes s'opposent, sauf pour la forme neutre, par l'aspect ou le temps qu'ils représentent : le système en *ta* est généralement employé pour représenter un temps passé ou un aspect accompli.

Forme neutre

La forme neutre[6] – dite *ren'yô* (連用, forme précédant le mot variable ou forme adverbiale) ou suspensive – est utilisée dans la position de complément adverbial. Elle n'a ni modalité ni temps en soi : ils sont déterminés par ceux du prédicat principal. Le système en *ta* a deux forme neutres : *itte* et *ittari*. Nous les appelons respectivement forme en *te* et forme en *tari*.

Par ailleurs, les formes neutres de base et en *te* peuvent constituer avec un autre verbe des verbes dits composés (複合動詞, *fukugô dôshi*) (Masuoka & Takubo, 1992). Certains verbes composés possèdent le sens obtenu par la conjonction du sens lexical des deux verbes tels que :

殴り　倒す
(*naguri - taosu*)
(donner des coups de poing [forme neutre de base] - faire tomber)
« faire tomber en donnant des coups de poing »

[6]Le terme neutre est la traduction de 中立形 (*chûritsu kei*) repris des travaux de Mikami.

Dans d'autres, le verbe post-posé perd plus ou moins sa fonction et son sens d'origine et ajoute seulement une valeur liée à l'aspect ou à la direction de l'action :

読んで　いる
(*yonde - iru*)
(lire [forme neutre en *te*] - [progressif])
« être en train de lire »

（宿題　を）手伝って　もらう
(*(shukudai - wo) - tetsudatte - morau*)
((devoir - [accusatif]) - aider [forme neutre en *te*] - [direction d'action : vers le locuteur])
« (m')aider à faire (mes devoirs) »

Dans le premier exemple, le verbe post-posé *iru* (sens d'origine « se trouver ») ajoute seulement au sens du verbe anté-posé, « lire », une valeur de progressivité. Dans le second, le verbe post-posé *morau* (sens d'origine « recevoir ») modifie le sens du verbe anté-posé, « aider » en précisant que la direction de l'action est dans le sens vers le locuteur.

Forme autonome

La forme autonome[7] (ou basique) est employée dans deux grandes fonctions différentes : déterminant des substantifs et position finale (prédicat principal). Elle est traditionnellement distinguée à chaque emploi par sa fonction et est appelée forme conclusive ou forme déterminante selon la fonction qu'elle assume dans l'occurrence effective. Nous appellerons également la forme du système en *ta* (*itta* dans le tableau) forme en *ta*.

Formes de condition, volitive et impérative

La forme de condition est employée dans les expressions de condition. Nous appellerons la forme du système en *ta* (*ittara* dans le tableau) forme en *tara*.

La forme volitive est utilisée pour exprimer une volonté et une conjecture. La forme du système en *ta* (*ittarô* dans le tableau) est aujourd'hui peu utilisée.

La forme impérative est utilisée pour exprimer un ordre et il n'existe pas de forme équivalente dans le système en *ta*.

Formes conclusives et connectives

Nous appelons également les trois formes (autonome, impérative et volitive) susceptibles d'indiquer la fin de phrase, formes conclusives, et les autres (neutre et de condition), formes connectives.

[7] Le terme autonome est la traduction de 自立形 (*jiritsu kei*) repris des travaux de Mikami.

5.4.2 Qualificatifs et copule

Qualificatifs en *i*

Les qualificatifs en *i* ont également deux systèmes basique et en *ta* qui ont chacun les formes neutres, autonome et de condition.

Qualificatifs en *na* et la copule

Les qualificatifs en *na* et la copule ont tout d'abord trois paradigmes selon le style d'énoncé : paradigmes en *da*, en *dearu* et en *desu*. Chaque paradigme a ensuite deux systèmes basique et en *ta* qui ont chacun les formes neutres, autonome et de condition.

5.4.3 Auxiliaires et suffixes variables

Auxiliaires variables

Les auxiliaires variables peuvent être distingués en trois types : ceux qui changent de forme selon le modèle de copule, selon le modèle de qualificatif en *na* et selon le modèle de qualificatif en *i*.

Suffixes variables

Les suffixes variables peuvent être distingués en deux types : ceux qui changent de forme selon les modèles de qualificatif et selon les modèles de verbe.

Outre ces deux grands types, Masuoka et Takubo définissent une classe particulière, celle de *nai* qui exprime la négation en suivant tous les types de mots prédicatifs. Ce suffixe de négation subit lui-même un changement de forme et a deux systèmes, basique et en *ta*, qui ont chacun les formes neutre, autonome et de condition.

5.4.4 Récapitulation

Forme		Verbe	Qualif. *-i*	Qualif. *-na* et copule
Conclusive	Autonome	行く ***ik** u*	寒い ***samu** i*	だ ***d** a*
		行った ***it** ta*	寒かった ***samu** katta*	だった ***d** atta*
	Volitive	行こう ***ik** ô*	寒かろう ***samu** karô*	だろう ***d** arô*
		行ったろう ***it** tarô*	寒かったろう ***samu** kattarô*	だったろう ***d** attarô*
	Impérative	行け ***ik** e*	---	---
Connective	Neutre	行き ***ik** i*	寒く ***samu** ku*	に‡ ***ni***
		行って ***it** te*	寒くて ***samu** kute*	で ***d** e*
		行ったり ***it** tari*	寒かったり ***samu** kattari*	だったり ***d** attari*
	Condition	行けば ***ik** keba*	寒ければ ***samu** kereba*	なら† ***nara***
		行ったら ***it** tara*	寒かったら ***samu** kattara*	だったなら ***d** attanara*
	Déterminante	---	---	な／の‡ ***na/no***

† absent de la définition de Masuoka-Takubo

‡ absent de la définition de Teramura

5.5 Éléments constituant la phrase japonaise

Soutenant la théorie de Mikami qui avançait tout au long de ses recherches linguistiques l'importance de l'introduction de ce concept, nous considérons que la structure fondamentale de la phrase japonaise est celle basée sur l'opposition thème-rhème.

Une fois que la phrase considérée est segmentée en thème et rhème, la partie rhème est analysée selon l'aspect fonctionnel. On constate alors dans cette partie la deuxième opposition : prédicat-compléments.

Avant d'entrer dans la discussion principale, nous introduisons tout d'abord un autre type d'opposition, plus conceptuelle, résidant dans la construction de la phrase : celle du dictum et du modus (§ 5.5.1). Nous abordons ensuite l'opposition thème-rhème (§ 5.5.2) puis celle prédicat-compléments (§ 5.5.3). Enfin, nous examinons également les éléments de phrases qui n'entrent pas dans ces oppositions (§ 5.5.4).

5.5.1 Opposition dictum-modus

Teramura (1982b) divise la phrase en deux parties : partie de la phrase qui décrit objectivement un fait ou une idée d'un côté, et partie indiquant la position du sujet parlant, qui prend la partie décrivant le fait ou l'idée comme matière, de l'autre.

Teramura considère la première comme correspondant aux *jojutsu naiyô* de Watanabe (cf. § 6.1.2), « dictum » de Bally et « proposition » de Fillmore et la seconde, comme correspondant aux termes, *chinjutsu*, « modus » et « *modality* », proposés par ces trois derniers.

Sur le modèle de Mikami, Teramura appelle la première *koto* (コト) et la second, *mûdo* (ムード). Aujourd'hui, dans la linguistique japonaise, la première est également appelée 言表事態 (*genpyô jitai*) ou encore 命題 (*meidai*) et la seconde, モダリティー (*modaritî*).

Nous adoptons également cette interprétation bien que peut-être trop grossière, et traduisons ces deux termes respectivement par « **dictum** » et « **modus** ».

Ces notions, en particulier celle du modus ou de la modalité, ne sont pas suffisamment étudiées et nous ne connaissons pas encore leur nature exacte. La définition de ces notions varie donc fortement d'un linguiste à l'autre.

Par exemple, Kudo (1989) définit la modalité comme l'expression grammaticale de la position du sujet parlant vis-à-vis du contenu descriptif de la phrase, de la réalité ou de l'interlocuteur et souligne que cette notion diffère fondamentalement de celle de « *modality* » de Fillmore ou de celle de *mûdo* (ムード) de Mikami et de Teramura, ces dernières incluant les éléments de temps, d'aspect et de voix.

Nous ne participons pas à ce débat sur les problèmes assez délicats liés à ces notions. Mais nous employons simplement le terme « modus » dans le sens relativement large de la définition de Teramura, en le distinguant du terme « **modalité** » que nous utilisons dans un sens plus restreint, celui de la définition de Kudo selon laquelle les éléments de temps, d'aspect et de voix n'appartiennent pas aux éléments de la modalité.

5.5.2 Structure fondamentale : opposition thème-rhème

Le Goffic considère que la distinction thème-rhème dans la phrase française est d'ordre psychologique, car elle repose « *essentiellement sur l'ordre des mots et la prosodie, qui n'offrent pas toujours d'indice formel d'interprétation sûre* ».

En revanche, le japonais dispose d'un mot grammatical indiquant le thème (ce dont on parle) – la particule *wa* (は) –, en plus de celui dédié à marquer la fonction dite « sujet » – la particule *ga* (が).

Du fait de cette particularité de la phrase japonaise, Mikami souligne l'importance d'établir une grammaire japonaise basée non pas sur l'opposition sujet-prédicat – un concept, d'après lui, particulier propre aux langues telles que l'anglais ou le français et qui ne convient pas au japonais –, mais sur la notion de thème. Il n'est cependant pas le premier linguiste à s'être rendu compte du statut tout à fait différent des particules *ga* et *wa*. On trouve déjà dans l'ouvrage de Matsushita (1928), une remarque sur cette différence.

Cette caractéristique de double structure fut également remarquée par des linguistes occidentaux, tels que Li & Thompson (1976). Dans cet article, ils définissent quatre types de langues selon la stratégie de construction des phrases, qui accorde de l'importance à la notion de thème ou de sujet. Le japonais est classé avec cette typologie dans la catégorie des langues ayant aussi bien le caractère de prédominance du sujet que celui de prédominance thématique. Les constructions japonaises caractéristiques liées à cette double structure, telles que celle appelée « double sujet », sont également étudiées par les chercheurs occidentaux comme Culioli (1999), qui a défini, notamment avec Desclés, une représentation formelle d'une des notions centrales de ces structures : la thématisation (Culioli & Desclés, 1982a,b).

Étant donné l'existence d'un élément syntaxique spécifique, l'introduction de l'opposition thème-rhème dès le niveau syntaxique semble indispensable, même flagrante, pour le japonais.

Définitions préliminaires du thème syntaxique et de la proposition syntaxique

Avant de poursuivre, définissons, ne serait-ce que brièvement, le thème, notion fondamentale pour les études sur la phrase japonaise, en attendant de traiter dans la section 6.4 le syntagme thématisé et la particule *wa* de manière plus approfondie.

Le thème est défini généralement, dans la linguistique japonaise, comme « ce à propos de quoi on parle ». Mais, comme nous le verrons plus loin, le thème peut être implicite ou réalisé sous une autre forme que le syntagme en *wa*. Dans la présente étude, nous appelons **thème** ou **thème syntaxique**, le thème – ce à propos de quoi on parle – explicite réalisé sous forme d'un syntagme en *wa*.

De même, la partie de phrase constituée autour d'un prédicat, susceptible de s'opposer au thème syntaxique, ne correspond pas forcément au rhème, propos sur le thème. Il est, au contraire, plus rare qu'elle soit entièrement rhématique. C'est pourquoi nous préférons désigner par **proposition** ou **proposition syntaxique**, le noyau structural constitué d'un prédicat et de ses compléments, constituant une phrase japonaise rentrant éventuellement en relation avec un thème syntaxique. Nous retravaillons également la notion de proposition dans le chapitre 7 consacré à la phrase complexe.

5.5.3 Constituants de la proposition : prédicat et compléments

Dans la proposition, s'organisent différents éléments autour du mot prédicatif.

On trouve un stemma de Mikami – reproduit figure 5.4 – dans Mikami (1953) représentant une phrase sans thème :

甲 が 乙 に 丙 を 紹介した
(*kô - ga - otsu - ni - hei - wo - shôkai shita*)
(X - [nominatif] - Y - [datif] - Z - [accusatif] - présenter [passé])
« X a présenté Z à Y »

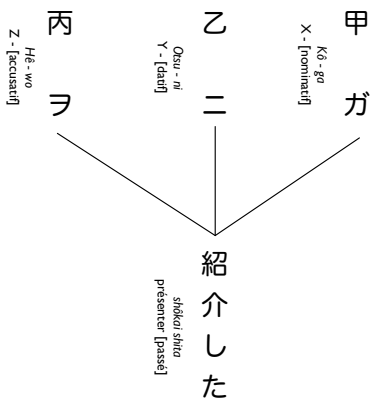


FIG. 5.4 – Stemma de Mikami

Trois syntagmes se terminant chacun par une particule marquant leur fonction s'accrochent au prédicat présent en bas du schéma.

Prédicat

Dans le cas du japonais, la fonction centrale du prédicat assurée par le verbe dans la phrase française est pourvue non pas systématiquement par les verbes mais par les mots prédicatifs, ensemble regroupant plusieurs catégories dont celle de verbe.

Concrètement, il existe quatre types de prédicat en japonais (Teramura, 1982b) :

1. verbe ;
2. qualificatif en *i* ;
3. qualificatif en *na* + copule ;
4. substantif + copule.

Ces prédicats se terminent souvent par plusieurs autres éléments tels que des auxiliaires – marquant le temps, l'aspect, la voix ou encore la modalité –, des suffixes ou des particules finales, constituant ainsi le véritable noyau de la proposition. Ce noyau est généralement appelé *jutsugo*, 述語 ou *jutsubu*, 述部, que nous traduisons par prédicat.

Pour être plus précis, nous appelons **prédicat** la partie s'étendant du dernier mot prédicatif – situé le plus postérieurement dans la phrase – jusqu'à la fin de la phrase. Dans l'exemple suivant :

マリー　が　日本　へ　行き　た　がって　いる　らしい
(*Mari - ga - nihon - e - iki - ta - gatte - iru - rashii*)
(Marie - [nominatif] - Japon - [direction] - aller - [vouloir] - [indication de la personne sujet de sentiment] - [état] - il semble que)
« Il semble que Marie souhaite partir pour le Japon »

le premier verbe « aller » suivi de deux suffixes, *iki - ta - gatte*, constitue avec le second verbe *iru*, un verbe composé qui représente le mot prédicatif, et forme finalement avec l'auxiliaire *rashii* le prédicat de la phrase.

Ces éléments suivant le mot prédicatif sont dits éléments du modus. Le prédicat japonais comporte donc les éléments du dictum (radical du mot prédicatif) et les éléments du modus (terminaisons et mots suivant le mot prédicatif).

Compléments primaires et secondaires

Tout comme pour le français, les compléments se distinguent d'abord en deux types : ceux rentrant directement en relation syntaxique avec le mot prédicatif et ceux qui sont en relation avec un substantif. Sur le modèle de la terminologie adoptée pour le français, nous appelons le premier type complément primaire et le second type, dit *rentai-shûshoku-go* (連体修飾語, mot qualifiant précédant le substantif), complément secondaire.

Les compléments primaires se distinguent encore en deux types : compléments essentiels et compléments accessoires.

Compléments essentiels

Les ensembles se terminant par une particule de cas, reliés au même mot prédicatif, sont appelés 補語 (*hogo*, complément), et nous les appelons compléments essentiels lorsque la distinction avec les compléments accessoires est nécessaire.

Dans l'exemple, 甲　が (*kô - ga*, X - [nominatif]), 乙　に (*otsu - ni*, Y - [datif]) et 丙　を (*hei - wo*, Z - [accusatif]) sont compléments du mot prédicatif.

Abandon de la notion de sujet

Comme on peut le constater dans le schéma de Mikami présenté précédemment, nous plaçons l'élément introduit par la particule *ga* (indicateur de la fonction dite « sujet ») sur le même plan que les autres compléments.

En effet, l'élément introduit par *ga* n'a pas de statut particulier – du moins de manière absolue comme dans la phrase française – par rapport aux autres compléments[8]. Il peut tout à fait être omis, comme les autres compléments, si le contexte permet une interprétation correcte. De plus, aucun accord n'étant nécessaire – sauf dans quelques structures particulières telles que celle de la politesse –, le verbe en est totalement indépendant.

Ainsi, nous renonçons au terme de « sujet » évoquant une supériorité par rapport aux autres compléments, terme que Mikami a considéré comme l'élément le plus nuisible au progrès de la recherche sur la syntaxe japonaise. Nous discuterons de l'appellation de chaque complément lors de la présentation des particules de cas dans la section 5.7.1.

Le schéma de Mikami que nous avons présenté au début de la section était fort semblable aux stemmas de Tesnière qui n'accordait pas de statut particulier au sujet dans la phrase française. Deux stemmas de Tesnière, reproduits dans les figures 5.5(a) et 5.5(b), représentent respectivement la phrase « Alfred parle » et celle « Alfred frappe Bernard ».

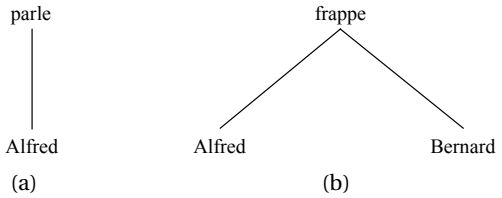


FIG. 5.5 – Stemmas de phrases françaises

La représentation en dépendance de ce type convient peut-être particulièrement à l'analyse du japonais pour lequel la prédominance syntaxique du nominatif par rapport aux autres compléments est peu marquée.

[8]Le statut du sujet dans la phrase japonaise est encore aujourd'hui un sujet d'actualité sur lequel les linguistes japonais n'arrivent pas à se mettre d'accord. La théorie de l'opposition – c'est-à-dire ceux qui veulent défendre l'utilité de la notion de sujet pour le japonais – la plus soutenue est sans doute celle avancée par Shibatani (1985). Il accorde de l'importance à la supériorité, ne serait-ce que relative, du sujet. Tout en reconnaissant certaines des particularités syntaxiques attribuées généralement au sujet, dans d'autres éléments que les syntagmes en *ga*, il propose une définition du sujet basée sur le concept de « prototype ». Se constituent alors d'une part la catégorie du sujet qui comporte quelques syntagmes non-*ga* et d'autre part un ensemble de syntagmes en *ga* qui n'appartiennent pas à la catégorie du sujet (Shibatani propose déjà ce principe dans ses travaux antérieurs (Shibatani, 1978)). Cette nouvelle définition tout à fait logique revient cependant à rajouter encore une nouvelle couche de catégories, puisqu'elle consiste en fait à réorganiser les classes de compléments catégorisés par la particule qui les introduit. Si nous avons adopté la théorie de Mikami plutôt que cette proposition, c'est que nous ne sentions pas, du moins dans le cadre de la présente thèse, l'utilité de l'introduction de cette nouvelle couche supplémentaire d'analyse.

Compléments accessoires

Il existe également un autre type de complément, dit 連用修飾語 (*renyô-shûshokugo*), que nous appelons désormais complément accessoire sur le modèle du terme adopté pour le français.

Ce sont notamment des adverbes ou des qualificatifs à une forme particulière marquant leur fonction de complément vis-à-vis du mot prédicatif.

Par exemple, dans la phrase :

部屋　を　手早く　片付ける
(*heya - wo - tebayaku - katazukeru*)
(chambre - [accusatif] - rapide [forme de complément] - ranger)
« (Je/On) range rapidement la chambre »

outre le complément essentiel *heya - wo* (chambre - [accusatif]), on trouve un autre élément, le qualificatif *tebayaku* (rapide) à la forme marquant sa fonction de complément, qui est un complément accessoire dépendant syntaxiquement, lui aussi, du mot prédicatif, le verbe *katazukeru* (ranger).

5.5.4 Éléments extérieurs à la structure thème-proposition

Il existe également des éléments qui n'appartiennent pas à cette opposition thème-proposition. Nous les appelons **éléments externes**.

Les éléments de phrase que les grammaires scolaires appellent 独立語 (*dokuritsu go*, mots indépendants) en sont un exemple type. Les éléments de liaison (接続語, *setsuzoku go*) qui établissent le lien avec les phrases précédentes sont également considérés comme extérieurs. Par ailleurs, beaucoup de linguistes reconnaissent aujourd'hui l'extériorité de certains adverbes, appelés adverbes de phrase, ainsi qu'une classe plus large d'éléments dits éléments d'évaluation, *hyôka-seibun*.

Éléments indépendants et éléments de liaison

Les 独立語 (*dokuritsugo*, **éléments indépendants**) définis dans la grammaire scolaire sont des interjections ou des mots tels que いいえ (*iie*, non), détachés en tête de phrase.

La grammaire scolaire en distingue quatre types :

1. Mots d'émotion :
   ああ、着いた。 (*aa - tsuita*, Ah - arriver [passé])
   « Ah, (je suis/nous sommes) arrivé(s) »
2. Adresse (vocatif, salutation) :
   さあさあ、急ぎなさい。 (*sâsâ - isogi nasai*, Allez - se dépêcher [ordre])
   « Allez, dépêchez-vous ! »
3. Réponse :
   はい、私です。 (*hai - watashi desu*, oui - moi [copule])
   « Oui, c'est moi »

4. Présentation d'une chose ou d'un fait :
   - 松、それ が 県 の 木 だ。
     (*matsu - sore - ga - ken - no - ki - da*)
     (pin - cela - [ga] - (notre) département - [no] - arbre - [copule])
     « Le pin, tel est l'arbre de notre département »
   - 卒業写真、それ は 私 の 好きな 曲 です。
     (*sotsugyô shashin - sore - wa - watashi - no - sukina - kyoku - desu*)
     (photo de fin de l'école - cela - [wa] - moi - [no] - favori - chanson - [copule])
     « "Sotsugyô shashin", telle est ma chanson préférée »
   - 九月九日、私は 一生 この日を 忘れない。
     ku gatsu kokonoka - watashi wa - isshô - kono hi wo - wasure nai)
     (le 9 septembre - moi [wa] - toute la vie - ce jour [wo] - oublier [négation])
     « Le 9 septembre, je n'oublierai jamais ce jour jusqu'à la fin de ma vie. »

Certains incluent les mots de liaison (接続詞, *setsuzoku shi*) dans cette catégorie, mais il est plus usuel de définir une autre classe distincte : les **éléments de liaison** (接続語, *setsuzoku go*). Bien que la catégorisation puisse susciter des débats, l'extériorité des éléments de liaison tout comme celle des éléments indépendants, semble être largement reconnue.

SN disloqué et SN en *wa* Autrefois, certains syntagmes en *wa* étaient également considérés comme faisant partie de la classe des éléments indépendants, mais aujourd'hui, on préfère les distinguer des syntagmes détachés sans *wa* – présentés dans l'exemple 4.

Contrairement au thème, qui est rarement repris dans la proposition par un pronom, les SN disloqués se caractérisent par le fait qu'ils sont toujours repris et implicitement insérés dans la structure syntaxique constituée autour du prédicat. Ce qui était justement la raison pour laquelle ils étaient considérés par Hashimoto (1938) comme extérieurs au reste de la phrase, la fonction syntaxique étant assurée par le pronom qui les reprenait. Fidèle à sa définition, Hashimoto considère également les syntagmes thématiques introduits par la particule *wa* comme des éléments indépendants lorsqu'ils sont repris par un moyen anaphorique.

Kitahara (1988) critique la définition de Hashimoto. Il dit que ces deux types de syntagmes doivent être distingués car les syntagmes en *wa* fonctionnent comme des thèmes dans la phrase alors que les SN disloqués – produisant un effet de « présentation » – ne sont que le fait (ou la chose) présenté, à propos duquel on parle dans le noyau de la phrase qui les suit. Cette explication ne clarifie cependant pas plus leur différence : elle ne dit rien sur la divergence entre le thème et la chose présentée à propos de laquelle on parle.

Néanmoins, nous avons l'intuition que ce sont bien deux éléments distincts. Selon la définition de Bonnot (1999) basée en premier lieu sur les travaux de Chafe (1976), le thème doit être non seulement « connu » mais aussi « donné » dans le contexte de l'énonciation. Les syntagmes thématiques en *wa* semblent effective-

ment être utilisés dans des contextes vérifiant ces conditions, tandis que les SN disloqués, connus ou non, sont donnés (ou mis en scène) pour la première fois par ce mécanisme de « présentation ».

Après cette mise en scène, la fonction de thème de la phrase proprement dite est assurée par le pronom (ou d'autres moyens anaphoriques) qui reprend ces SN disloqués. Ce mécanisme de présentation d'un SN sert spécifiquement et seulement à activer une notion dans l'esprit de l'interlocuteur.

Si bien que les SN disloqués sont indépendants à tel point qu'on peut les considérer comme des phrases indépendantes, contrairement au thème qui n'est pas aussi dissociable de son rhème. D'ailleurs, les SN disloqués, ou plus précisément les pronoms qui les reprennent, ne jouent pas forcément le rôle de thème dans la phrase comme le montre le dernier exemple.

Adverbes de phrase et éléments d'évaluation

On considère généralement que l'intériorité ou l'extériorité des adverbes par rapport au noyau structural thème-proposition est décidée selon leur nature lexicale. Si bien qu'il existe beaucoup d'études consacrées à leur catégorisation et à la caractérisation de chaque type. Bien que la nécessité ou la justesse de cette distinction soit approuvée par la plupart des linguistes, le détail du classement diverge.

Nous présentons la catégorisation de Yamada, représentant la base de toutes les études contemporaines sur les adverbes, la grammaire de Masuoka-Takubo et les travaux de Kudo, sur lesquels nous nous appuyons pleinement pour ce sujet[9].

Yamada (1936) a distingué les mots considérés aujourd'hui comme des adverbes en trois types, deux appartenant au dictum et un au modus :

– 属性副詞 (*zokusei fukushi*, adverbe attributif) : appartenant au dictum

  1. 情態副詞 (*jôtai fukushi*, adverbe de manière)
     ゆっくり (*yukkuri*, lentement), すぐ (*sugu*, tout de suite) ;
  2. 程度副詞 (*teido fukushi*, adverbe de degré)
     もっと (*motto*, encore ; plus), すごく (*sugoku*, extrêmement) ;

– 陳述副詞 (*chinjutsu fukushi*, adverbe du modus) : appartenant au modus
  おそらく (*osoraku*, probablement), もし (*moshi*, si [adverbe accompagnant l'expression de condition]) ;

La grammaire de Masuoka & Takubo (1992) distingue tout d'abord les adverbes en deux types : compléments du prédicat, correspondant *grosso modo* aux adverbes attributifs (*zokusei-fukushi*) de Yamada, et ceux entrant en relation avec l'ensemble de la phrase, adverbes du modus (*chinjutsu-fukushi*) de Yamada. Le premier type est considéré comme des adverbes au sens strict du terme et le second est appelé « adverbes de phrase ».

[9]Nous renvoyons pour un état de l'art sur l'étude des adverbes, aux ouvrages (Ichikawa, 1976 ; Kudo, 2000 ; Yazawa, 2000).

Les adverbes de phrase comprennent différents types, dont les deux principaux sont :

– 陳述副詞 (*chinjutsu fukushi*, adverbe du modus) : accompagnant les éléments du modus situés après le mot prédicatif

> どうも　知っている　ようだ
> (*dômo - shitteiru - yôda*)
> ([adverbe accompagnant "yôda"] - savoir [état] - il semble que)
> « (Il/Elle) semble être courant »

– 評価副詞 (*hyôka fukushi*, adverbe d'évaluation)

> **幸い**　無事だった
> (*saiwai - buji datta*)
> (par chance - être sain et sauf [passé])
> « Par chance, (je/il/elle/...) était sain et sauf »

Kudo (1997) travaille sur les adverbes d'évaluation dans un cadre plus large et tente d'élucider la classe syntaxico-sémantique d'éléments dits 評価成分 (*hyôka seibun*, éléments d'évaluation), concept provenant des travaux de Watanabe (1971) qui a introduit la notion de 誘導副詞 (*yûdô-fukushi*, adverbes de guide). Ces éléments d'évaluation sont définis comme des éléments situés en tête, qui sont indépendants du reste de phrase et qui expriment l'évaluation du locuteur pour le contenu de la phrase. Ces éléments, comptés à l'époque parmi les compléments du prédicat (*renyô-shûshokugo*), sont inclus dans les éléments indépendants par Suzuki (1972).

Les éléments d'évaluation peuvent être réalisés non seulement par des adverbes d'évaluation, mais aussi par des qualificatifs (en *i* et en *na*) à la forme adverbiale (連用形, *ren'yô-kei*) ou par des syntagmes constituant des locutions figées (Ichikawa, 1976) :

– adverbe d'évaluation (*hyôka fukushi*)

> **あいにく**　大粒の　雨が　降り出した
> (*ainiku - ôtsubu no - ame ga - furidashita*)
> (malheureusement - grosses gouttes - pluie [ga] - commencer à tomber [passé])
> « Malheureusement, il commença à pleuvoir à torrent »

– qualificatifs à la forme adverbiale

> **めずらしく**　東京に　大雪が　降った
> (*mezurashiku - tôkyô ni - ôyuki ga - futta*)
> (rare [à la forme adv] - Tokyo [locatif] - grosse neige [ga] - tomber [passé])
> « Fait rare, il a beaucoup neigé à Tokyo »

– syntagmes constituant des locutions

> **困ったことに**、さいふを　なくしてしまった
> (*komattakoto ni - saifu wo - nakushite shimatta*)

(embêtant - porte-feuille [wo] - perdre [passé])
« Quel ennui : j'ai perdu mon porte-feuille »

5.5.5 Récapitulatif

Récapitulons ce que nous venons d'étudier sur les éléments de la phrase japonaise.

1. Éléments externes : mots indépendants de la grammaire scolaire – tels que le mot いいえ (*iie*, non) –, adverbes d'évaluation, etc.
2. Thème : élément s'opposant à la proposition et qui se trouve sur un pied d'égalité avec la proposition.
3. Proposition
   a) éléments essentiels
      i. prédicat ;
   b) éléments complémentaires
      i. éléments complétant un mot prédicatif
         A. complément : substantif + particule de cas ;
         B. circonstanciel : adverbe ou qualificatif à la forme qualifiant le prédicat, etc. ;
      ii. éléments complétant un substantif : y compris les éléments coordonnés.

Nous présentons également deux figures comparatives, figures 5.6 et 5.7 (voir page suivante), représentant respectivement un schéma de la phrase française et un de la phrase japonaise.

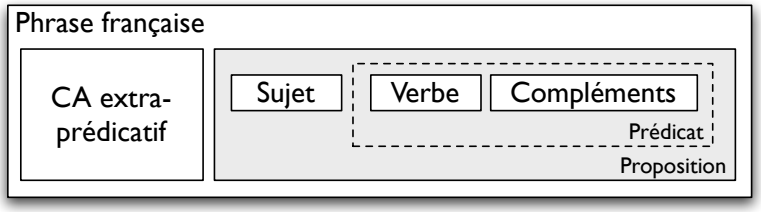


FIG. 5.6 – Structure de la phrase française

La définition du terme prédicat de la phrase japonaise – qui correspond plutôt à la notion de constante prédicative de la logique des prédicats – est différente de l'emploi de Le Goffic que nous utilisons pour l'analyse de la phrase française. Dans la mesure où nous conserverons ces deux définitions chacune exclusivement pour l'analyse de l'une ou l'autre langue – la première pour le japonais et la seconde pour le français – il n'y a probablement aucun risque de confusion.

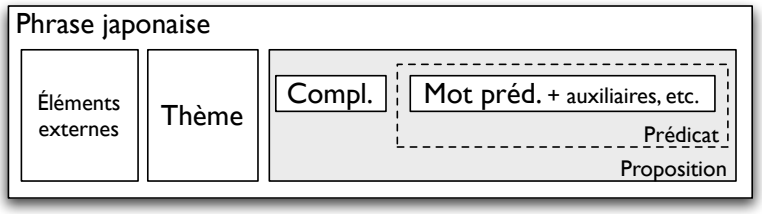


FIG. 5.7 – Structure de la phrase japonaise

5.6 Ordre des mots

Tesnière (1988) distingue les langues selon le sens du relevé linéaire des mots en connexion structurale, en deux classes : centrifuge et centripète. Lorsqu'on énonce d'abord le régissant et ensuite le subordonné, la langue est centrifuge, et dans le cas inverse centripète.

Dans la phrase en japonais – classé dans la catégorie des langues centripètes accusées –, l'élément subordonné est toujours antéposé à son régissant.

Ainsi, les compléments précèdent toujours le mot prédicatif comme dans la phrase :

日本語　を　流暢に　話す
(*nihongo - wo - ryûchôni - hanasu*)
(langue japonaise - [accusatif] - couramment - parler)
« parler couramment le japonais / (il) parle couramment le japonais »

où les compléments essentiel *nihongo - wo* (langue japonaise - [accusatif]) et accessoire *ryûchôni* (couramment) sont tous les deux placés avant le prédicat *hanasu* (parler).
Un qualificatif précède également le nom :

青い　空
(*aoi - sora*)
(bleu - ciel)
« ciel bleu »

Enfin, une subordonnée doit aussi être mise avant la principale :

雨　が　降ったら　出かけない
(*ame - ga - futtara - dekakenai*)
(pluie - [nominatif] - tomber [condition] - sortir [négation])
« S'il pleut, (je) ne sortirai pas »

5.6.1 Ordre absolu : régit - régissant

Mikami (1953) illustre l'ordre absolu « régit - régissant » du japonais par comparaison des structures complexes de l'expression hypothétique en anglais et en japonais :

> « Contrairement à l'anglais pour lequel l'ordre des propositions principale et subordonnée est assez libre, dans la phrase japonaise, l'ordre entre le qualifiant et le qualifié est fixé, à savoir le qualifiant d'abord et le qualifié après. Si bien que lorsqu'on compare les phrases suivantes :
>
> Will you go out, if it rains ?
> 出カケル　カイ、雨　ガ　降ッ　テモ？
> (*dekakeru - kai, - ame - ga - fut temo*)
> (sortir - [interrogation] - pluie - [nominatif] - tomber - même si)
>
> alors que la subordonnée anglaise semble pouvoir se placer librement à la guise du locuteur, dans la phrase japonaise, au vu de la nature de la forme des mots variables, il y a clairement eu inversion. »

Comme nous l'avons vu dans les études linguistiques sur le français, la subordonnée circonstantielle placée en tête, élément extérieur au prédicat, se distingue assez clairement de celle en fin de phrase, circonstant lié intra-prédicatif, même dans la phrase française.

Mais, dans le cas du japonais, l'ordre principale-subordonnée est théoriquement impossible et cette inversion forcée, sans doute coûteuse, est très peu utilisée en particulier à l'écrit[10].

5.6.2 Ordre libre entre les compléments

Néanmoins, les Japonais parlent souvent d'un ordre relativement libre des mots en japonais. En effet, comme Mikami (1953) l'explique, l'ordre entre les com-

[10] Kuno (1978) signale toutefois l'emploi fréquent, à l'oral, de la forme où un/des complément(s) et/ou le thème – éléments censés être mis en avant – sont mis derrière le prédicat telle que :

行ッテ　シマイ　マシタ　ヨ、　山田　ハ、　花子　ヲ　連レテ、東京　ニ。
(*itte - shimai - mashita - yo - yamada - wa - hanako - wo - tsurete - Tôkyô - ni*)
(aller - [accomplissement] - [politesse + passé] - [conviction] - Yamada - [thème] - [accusatif] - amener [forme neutre] - Tokyo - [locatif])
« (Il) est parti, Yamada, accompagné de Hanako, à Tokyo »

Il explique ce phénomène non pas par l'inversion, mais par l'omission de certaines informations dans la principale et par la reprise de ces éléments omis, disloqués après la principale, dans un but de rappel.

Par ailleurs, Kindaichi (1988) explique que cette inversion fréquente à l'oral a pour but – bien que les locuteurs n'en soient pas forcément conscients – de prononcer d'abord les informations importantes contenues dans les mots déterminés (plutôt que déterminants), et surtout dans les prédicats, qui se situent en fin de phrase avec l'ordre « normal » de l'énoncé japonais.

pléments du prédicat – y compris le nominatif – est beaucoup plus souple. En d'autres termes, la position des éléments dépendant du même élément dans le stemma de Tesnière est interchangeable dans l'ordre linéaire.

Par exemple, pour le contenu sémantique « ma petite sœur étudie le français », chacun des mots subordonnés « petite sœur » et « français » dépendant tous les deux du même régissant « étudier », peut être, dans la réalisation d'une phrase japonaise, relevé avant l'autre, seul le régissant « étudier » se plaçant obligatoirement derrière tous les éléments dépendant de lui. Ainsi, deux ordres linéaires sont possibles :

妹　が　フランス語　を　勉強する。
(*imôto - ga - furansu go - wo - benkyô suru*)
(ma petite sœur - [nominatif] - français - [accusatif] - étudier [non passé])

フランス語　を　妹　が　勉強する。
(*furansu go - wo - imôto - ga - benkyô suru*)
(français - [accusatif] - ma petite sœur - [nominatif] - étudier [non passé])

Toutefois, cette liberté de placement est beaucoup plus restreinte qu'on ne le croit intuitivement tout comme nous l'avons constaté pour les circonstants dans la phrase française. Selon notre étude antérieure (Nakamura-Delloye, 2003b)[11] , un ordre particulier des compléments produit une valeur énonciative précise et le choix est loin d'être aléatoire, fixé souvent selon le contexte.

5.7 Moyens d'indication de la fonction syntaxique dans la phrase japonaise

En japonais, la fonction syntaxique d'un syntagme est marquée essentiellement par deux moyens : à l'aide d'une particule de cas pour les syntagmes nominaux (§ 5.7.1) et par la variation de forme pour les syntagmes se terminant par un mot variable (§ 5.7.2).

5.7.1 Particules de cas et fonctions syntaxiques

La fonction syntaxique – « statut syntaxique » selon la terminologie de Garnier (1982) – de complément est marquée par certaines particules appartenant à une sous-catégorie de particules, appelée particules de cas (格助詞, *kaku joshi*). Mais la définition exacte des particules de cas varie selon les linguistes.

Définition des particules de cas

Mikami, après avoir défini les cas comme les catégories de fonctions qu'assure un substantif ou un groupe nominal, limite les fonctions attribuées par une parti-

[11] Kuno (1978) et Takami (1995) signalent également le changement de la nature informative des constituants selon leur position.

cule de cas à celles complétant les mots variables et celles qualifiant un substantif. Il définit dans (Mikami, 1963a) huit particules de cas qui peuvent constituer un complément du mot variable (*ga, wo, ni, de, to, e, kara, yori*), et une particule de cas (*no*) qui peut constituer un syntagme qualifiant un substantif[12], que nous allons étudier plus précisément dans la présente section.

En revanche, Masuoka & Takubo (1992) définissent dans les particules de cas comme des mots indiquant le rapport entre le prédicat et le complément qu'elles constituent avec le syntagme nominal précédent, excluant ainsi la particule indiquant la fonction qualificative. Ils reconnaissent neuf particules de cas (*ga, wo, ni, de, to, e, kara, made, yori*).

Catégorisations des cas : formelle ou sémantique

Les cas attribués par les particules de cas ont des appellations différentes selon que les critères sont formels ou sémantiques.

Avec les critères formels, la fonction attribuée par la particule *ga* est appelée ガ格 (*ga-kaku*, cas de *ga*), celle par la particule *wo* est appelée ヲ格 (*wo-kaku*, cas de *wo*), et ainsi de suite.

Avec les critères sémantiques, la fonction attribuée par la particule *ga* est appelé 主格 (*shu-kaku*, cas principal) si le référent de l'élément contenant cette particule est l'acteur, mais une fonction attribuée par la particule *no* peut également être *shu-kaku* dans un contexte enchâssé où la particule *no* remplace *ga*. Non seulement les types définis selon ces critères sémantiques varient énormément, mais en plus la catégorisation peut différer d'un linguiste à l'autre.

Mikami définit neuf cas (1 ~ 9) dans lesquels le substantif constitue avec la particule un complément du mot variable, et une particule de cas (10) où le substantif constitue avec elle un syntagme qualifiant un substantif, comme suit[13] :

1. 時の格 (*toki no kaku*), *case of time* : marqueur zéro ;
2. 主格 (*shu-kaku*), *nominative*[14] : ガ (*ga*) ;
3. 対格 (*tai-kaku*), *accusative* : ヲ (*wo*) ;
4. 位格 (*i-kaku*), *locative* : ニ (*ni*) ;

[12]Pour la particule de cas qualificatif, Mikami (1953) étudie différents sens de *no* (の, généralement traduit par la préposition « de » en français) et considère comme un élément autre que la particule de cas, le mot *no* du syntagme « X *no* Y » dans lequel X désigne la qualité de Y, tel que « *kenchiku-ka no ani* » (architecte - *no* - grand frère) que l'on peut paraphraser par « *kenchiku-ka dearu ani* » (architecte - [copule à la forme qualificative] - grand frère, « (mon) grand frère qui est architecte »).
Cependant, dans Mikami (1963a), il semble réaliser une catégorisation plus formelle et considère certains emplois de *no, ni, de, to* – jugés parfois comme des emplois d'un mot d'une autre catégorie – comme des emplois exceptionnels des particules de cas.

[13]Les traductions anglaises sont reproduites telles quelles du texte de Mikami (1955).

[14]Conscient de l'inexactitude de cette traduction, Mikami (1955) la conserve en renonçant à sa propre traduction « *subjective case* » qu'il avait utilisée dans Mikami (1959). Par ailleurs, dans Mikami (1960), il adopte le terme « *agentive* » pour le cas formé par *ga*.

5. 与格 (*yo-kaku*), *dative* : へ、ニ (*e, ni*) ;
6. 奪格 (*dak-kaku*), *ablative* : カラ、ニ (*kara, ni*) ;
7. 具格 (*gu-kaku*), *instrumental* : デ (*de*) ;
8. 共格 (*kyô-kaku*), *commitative* : ト (*to*) ;
9. 比較の格 (*hikaku no kaku*)[15] : ヨリ (*yori*) ;
10. 連体格 (*rentai-kaku*), *dominative* : ノ (*no*).

Mais, même parmi les ouvrages de Mikami, cette définition des cas varie.

Dans Mikami (1953), Mikami ne définit que sept cas de complément du prédicat parmi lesquels on ne trouve ni le cas de temps, ni le cas de comparaison. Dans Mikami (1955), il exclut toujours le cas de comparaison, définissant huit cas de complément. Puis dans Mikami (1963a), il inclut le cas de comparaison dans les cas de complément du prédicat, tout en jugeant certains emplois de la particule *yori* comme exceptionnels.

Sens des particules de cas

Masuoka & Takubo (1992) décrivent différents emplois des neuf particules de cas définies. Nous reprenons ici leurs travaux, auxquels nous avons apporté quelques modifications tenant compte des études de Mikami.

1. ガ (*ga*) indique : l'élément qui accomplit l'action ou dont on décrit l'état ; l'objet dont on décrit l'état ; la partie de l'élément dont on décrit l'état ;
2. ヲ (*wo*) indique : la destination de l'action ou du sentiment ; l'endroit de déplacement ; le point de départ du déplacement ;
3. ニ (*ni*) indique : l'endroit où existe l'être ou la chose ; le possesseur ; le point d'arrivée du déplacement ; le destinataire de l'action ; l'objet sur lequel on agit ; la partie de l'élément dont on décrit l'état ; la cause ; le but du déplacement ; le moment de l'évènement ;
4. カラ (*kara*) indique : le point de départ du déplacement ; le partenaire de l'action de réception ; l'élément qui accomplit l'action en tant que point de départ du déplacement ; le point de départ du moment ; la cause en tant que déclencheur de l'évènement ; le motif du jugement ; la matière première ;
5. ト (*to*) indique : le partenaire de l'action collective ; le partenaire dans une relation symétrique ;
6. デ (*de*) indique : l'endroit de l'action ou l'évènement ; le moyen ou l'outil ; les matériaux ; la cause ; l'étendue ; la limite ; le critère ; l'élément qui accomplit l'action ;
7. へ (*e*) indique : la destination ou la direction ;

[15] Nous n'avons pas trouvé la traduction de Mikami pour ce terme. Il s'agit d'un cas de comparaison.

8. マデ (*made*) indique : l'endroit où le déplacement se termine ; le moment où l'évènement se termine ;
9. ヨリ (*yori*) indique : le partenaire de la comparaison ; le point de départ dans le temps.

Notre définition

Nous définissons les neuf particules ci-dessus, ainsi que la particule *no*, comme particules de cas.

Étant donnée la nature de nos travaux, s'intéressant plus à la syntaxe, nous adoptons en principe les appellations basées sur les critères formels. Mais pour faciliter la compréhension des lecteurs non familiers avec le japonais, nous indiquerons si nécessaire leur sens plutôt que le cas, notamment pour les particules ayant des sens très divers.

5.7.2 Indication de la fonction syntaxique par les formes des mots variables

À la différence du français dans lequel les verbes et les adjectifs s'accordent en genre ainsi qu'en nombre, les variations morphologiques des mots variables japonais (cf. § 5.4) dépendent notamment de leur fonction syntaxique, donc de leur position dans la phrase.

Formes en fin de phrase

Le mot qui occupe la position de prédicat principal, régissant de tous les éléments, et qui se situe donc en fin de phrase – sauf dans le cas où il est suivi d'une particule finale – prend la forme dite « finale » ou « conclusive », marquant ainsi la fin de la phrase par sa forme elle-même. Les formes autonomes sont utilisées à cette fin.

本　を　買う。
(*hon* - *wo* - *kau*)
(livre - [accusatif] - acheter [forme autonome de base = non passé])
« (J')achète un/des livre(s) »

本　を　買った。
(*hon* - *wo* - *katta*)
(livre - [accusatif] - acheter [forme autonome en *ta* = passé])
« (J'ai) acheté un/des livre(s) »

Les formes impérative et volitive peuvent également marquer la fin de phrase, mais avec une valeur supplémentaire, à savoir celle de l'ordre et de la volonté ou de la supposition.

Formes précédant le prédicat

Un mot variable prend la forme dite « précédant le prédicat », s'il est prédicat de la subordonnée adverbiale. Les formes dédiées à cette fonction sont typiquement les formes neutres[16].

本　を　なくして、父　に　叱られた。
(*hon - wo - nakushite - chichi - ni - shikarareta*)
(livre - [accusatif] - perdre [forme neutre en *te*] - mon père - par - être grondé [passé])
« (J'ai) perdu le livre, (j'ai) été grondé par mon père »

Mais les formes de condition constituent également une subordonnée adverbiale à valeur de condition.

本　を　なくせば／なくしたら、父　に　叱られる。
(*hon - wo - nakuseba/nakushitara - chichi - ni - shikarareru*)
(livre - [accusatif] - perdre [forme de condition de base/en *ta*] - mon père - par - être grondé [non passé])
« Si (je) perds le livre, (je serai) grondé par mon père »

Formes précédant le substantif

Un mot variable situé avant un nom, c'est-à-dire à la fin d'une construction déterminant un nom où il y est prédicat, prend la forme « précédant – donc qualifiant – le substantif ». Les formes autonomes dédiées à la fonction conclusive sont également employées à cette fonction déterminante.

きのう　買った　本
(*kinô - katta - hon*)
(hier - acheter [forme autonome en *ta*] - livre)
« le livre que j'ai acheté hier »

Comme nous pouvons le constater dans l'exemple, il n'y a aucun connecteur du type pronom relatif en français qui marque l'introduction de la subordonnée déterminante dans la phrase. Nous allons étudier dans la section suivante de manière plus détaillée la structure de la subordination déterminante en japonais.

[16]Les subordonnées à la forme neutre, qui n'ajoutent aucune valeur spécifique, sont souvent considérées comme coordonnées plutôt que subordonnées. Nous les considérons comme subordonnées syntaxiquement. Nous discuterons ce sujet plus précisément dans les sections 7.9 et 7.10 consacrées à la typologie des subordonnées.

5.8 Structure de la subordination déterminante

Avant d'étudier la structure de la subordination déterminante en japonais, rappelons deux types de structures de subordination qui existent en français : structure à relatif et structure à intégratif.

5.8.1 Structure avec un pronom relatif

Dans la phrase française, la structure de subordination déterminante est principalement réalisée par l'utilisation d'un pronom relatif comme dans la phrase :

J'ai perdu le livre **que** mon père m'avait prêté.

La figure 5.8 représente le mécanisme de la subordination avec un pronom relatif.

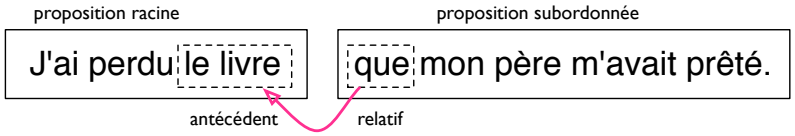


FIG. 5.8 – Structure avec un pronom relatif

La jonction des deux propositions (proposition racine et proposition subordonnée) dans une structure avec un relatif est assurée par un **pronom relatif**, qui, appartenant à la subordonnée, pointe un constituant de la racine, dit antécédent, en se plaçant derrière lui et marque sa fonction syntaxique dans la proposition subordonnée.

5.8.2 Structure avec un pronom intégratif

Les structures avec une subordonnée intégrative sont réalisées avec un autre mécanisme présenté dans la figure 5.9.

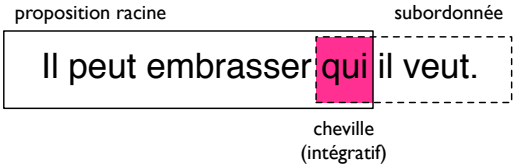


FIG. 5.9 – Structure avec une cheville

La jonction des deux propositions (proposition racine et proposition subordonnée) dans ce type de structure est assurée par un élément « **cheville**[17] » qui

[17] Selon Le Goffic (1993a), l'effet de « cheviller » est un terme utilisé dans Damourette & Pichon (1971).

est à la fois constituant des deux propositions. Dans l'exemple, le pronom « qui », appartenant à la fois aux deux propositions, fonctionne comme cheville et assure la jonction des deux propositions. En français, du moins moderne, seuls les pronoms intégratifs capables d'assurer le chevillage permettent la constitution d'une structure de subordination avec cheville.

5.8.3 Structure avec cheville en japonais

La structure de subordination déterminante en japonais est toujours réalisée selon ce mécanisme de chevillage[18]. L'élément assurant le chevillage ne se limite donc pas à quelques mots particuliers mais tous les substantifs peuvent jouer le rôle de cheville[19].

きのう　買った　本　を　なくした。
(*kinô - kat ta - hon - wo - nakushita*)
(hier - acheter [parfait] - livre - [accusatif] - perdre [passé])
« (J'ai) perdu le livre **que** j'avais acheté hier »

母さん　が　作った　カレー　は　日本一　だ。
(*kâsan - ga - tsukut ta - karê - wa - nihon ichi - da*)
(maman - [nominatif] - préparer [parfait] - curry - [thème] - le meilleur du Japon - [copule])
« Le curry **que** ma maman prépare est le meilleur du Japon »

Cette absence de structure à relatif ramène au fait qu'il n'existe pas dans la structure de subordination déterminante en japonais la notion d'« antécédent ». Nous avons seulement l'élément cheville, que nous appelons désormais **base**[20].

[18] Il est à noter que la fonction de l'élément cheville dans la proposition racine est marquée par la particule de cas qui le suit. En revanche, sa fonction dans la proposition subordonnée n'est aucunement indiquée comme dans la structure avec intégratif en français.

[19] Notre étude antérieure a montré, en revanche, que les mots japonais équivalents aux mots « qu- » français ne possédaient pas la fonction intégrative sauf quelques rares cas d'exception (Nakamura-Delloye, 2003c).

[20] Cette appellation est inspirée du terme « *base* » (en anglais) que Mikami (1959) a utilisé avec le mot japonais équivalent 底 (*soko*) pour désigner le substantif pour lequel on transforme une phrase en subordonnée déterminante de manière à le qualifier.

CHAPITRE 6

ÉTUDE DE LA PHRASE JAPONAISE

Nous abordons dans ce chapitre différents sujets clarifiant la notion de phrase japonaise. Comme dans beaucoup de langues, la définition de cette unité de base pour les études syntaxiques posait et pose encore aujourd'hui des problèmes aux chercheurs. Mais, sans pouvoir la définir exactement, les études sur la structure et les éléments constituants de cette unité ont tout de même connu une évolution considérable. Nous allons tout d'abord passer en revue l'évolution de la définition de la phrase japonaise (§ 6.1) et le développement des analyses de la structure de la phrase japonaise par une architecture multicouche (§ 6.2). Nous présenterons ensuite la typologie des phrases japonaises (§ 6.3) avant de passer aux études plus détaillées des deux éléments de la phrase : le syntagme thématisé et la particule wa *(§ 6.4) et les éléments préposés par rapport au thème (§ 6.5).*

6.1 État de l'art I : définitions

Nous présentons ici différentes définitions de la phrase proposées jusqu'aujourd'hui.

Nous présentons d'abord les définitions basées sur des critères formels (§ 6.1.1), puis les définitions plus conceptuelles (§ 6.1.2). Enfin, nous aborderons également le caractère incomplet de la phrase japonaise (§ 6.1.3).

6.1.1 Définitions basées sur des critères formels

Nous avons défini, dans le cadre de l'alignement des phrases, les phrases (graphiques) comme des unités entourées de deux séparateurs graphiques de phrase

préalablement définis (i.e. points final, interrogatif et exclamatif, deux points, point virgule et retour chariot).

On trouve dans le dictionnaire de linguistique japonaise (Hayashi et al., 1988) une définition de la phrase d'un point de vue purement formel identique à la nôtre :

> « Contrairement aux mots, la reconnaissance de chaque phrase n'est pas vraiment difficile pour le japonais non plus. C'est-à-dire, une phrase est un bloc segmenté par un point final : "。 ". »

La grammaire scolaire adopte une définition utilisant également des critères formels, non graphiques mais phonétiques/prosodiques, définition proposée par Hashimoto (1948) :

1. une phrase est une suite de sons ;
2. avant et après une phrase, il y a toujours une coupure de son ;
3. en fin de phrase, s'ajoute une mélodie particulière.

Cette définition est cependant largement critiquée. Kitahara fait remarquer dans Kitahara (1976) :

> « Ces caractéristiques sont des phénomènes que l'on peut reconnaître dans la conséquence de la constitution d'une phrase, mais elles sont loin de pouvoir constituer une définition de phrase. D'ailleurs, il faut dire que la constatation des caractères de ce type ne permet aucunement de cerner la véritable nature de la phrase. »

6.1.2 De la définition formelle à la définition conceptuelle

À la fin du XIX^{ème} siècle, sous l'influence des grammaires occidentales, beaucoup de grammairiens japonais définirent la phrase sur la base, d'une part de la présence d'un sujet et d'un prédicat, et d'autre part du caractère complet.

Mais, dès le début du XX^{ème} siècle, beaucoup de linguistes, à commencer par Yamada, affirmèrent l'inadaptation de cette définition à la langue japonaise, pour laquelle ces deux éléments jugés obligatoires dans les langues telles que le français ne sont pas toujours présents. En effet, le sujet est, bien au contraire, plus souvent absent dans la phrase japonaise.

Définition de Yamada

Les séquences constituées d'un seul mot telles que 「犬」(*inu*, chien), 「火事」 (*kaji*, incendie) étaient pour Yamada des phrases tout à fait complètes signifiant respectivement « (Il y a) un chien ! », « (Il y a) un incendie ! »[1].

Il analyse dans Yamada (1936) ces énoncés d'un seul mot comme suit :

[1] Aujourd'hui, nous distinguons les phrases contenant une structure prédicative (phrases développées, selon la terminologie de Masuoka & Takubo (1992)) des phrases sans noyau prédicatif (phrases non-développées) (cf. § 6.3.2). Les exemples de Yamada appartiennent aux phrases sans noyau prédicatif et ce type de phrase n'est pas particulier à la langue japonaise. On trouve

> « [...] 「犬」, 「火事」 sont des mots lorsqu'on les considère en tant que mot, mais ce sont également des séquences constituant une phrase quand on les analyse en tant que phrase. Si on peut les définir comme des phrases, c'est parce qu'on les utilise pour exprimer une pensée. Bien que leur apparence limitée à un seul mot soit simple, à l'intérieur de cette forme simple, ils possèdent l'activité complexe d'une pensée ; seulement, la représentation de cette pensée est réalisée avec un seul mot. »

Il considère ainsi que définir une phrase uniquement par des formes est impossible, proposant alors comme nouveau critère le « contenu », exprimé par une séquence qu'il appelle 思想 (*shisô*, pensée).

Selon Yamada, une pensée unifiée est composée de différentes idées assemblées par l'« effet d'unification » (統覚作用, *tôkaku sayô*). Cet effet provoqué par un mot variable situé à la position du prédicat dans une phrase est appelé *chinjutsu* (陳述, traduit par « capacité phrasogénératrice » par Garnier (1982)).

Yamada cherche à définir également la proposition par l'introduction de cette notion d'« effet d'unification ». Une proposition est la représentation linguistique d'une pensée unifiée par un seul déclenchement de cet effet d'unification. Ainsi, une phrase constituée d'une seule proposition est appelée phrase simple, et celle constituée de plus d'une proposition, phrase complexe. Cependant, cette définition de Yamada n'a pas permis de tracer la frontière exacte entre proposition et phrase.

Définition de Tokieda

Beaucoup de grammairiens japonais s'intéressèrent ensuite à définir la notion de phrase sur la base de cette théorie de Yamada, en particulier autour de la notion de *chinjutsu*, effet phrasogénérateur.

Tokieda (1950) définit comme caractères nécessaires à la reconnaissance d'une phrase les trois points suivants :

1. être l'expression d'une pensée concrète ;
2. posséder un caractère unifié ;
3. posséder un caractère terminé.

Une pensée concrète est constituée par la conjonction d'éléments objectifs et d'éléments subjectifs. Les éléments objectifs sont réunis par l'expression subjective, donnant ainsi un caractère unifié à la phrase. Enfin, le caractère terminé de la phrase est donné par le mot variable en fin de phrase dont la forme marque la terminaison.

par exemple dans Le Goffic (1993a) la description de ce type de phrase dans la langue française, qu'il appelle phrases sans verbe ou phrases nominales. En revanche, dans le cas du japonais, une construction sans mot variable – ou un substantif tout seul – peut, suivie de la copule, parfaitement constituer une phrase développée, ce qui est sans doute une caractéristique assez particulière du japonais.

L'introduction du caractère terminé par Tokieda permet de distinguer la phrase de la proposition. C'est une propriété dont n'est pas munie une proposition et propre à une phrase. Cependant, la différence entre les caractères unifié et terminé n'est pas suffisamment discutée dans sa théorie.

Cette nouvelle question est ensuite résolue par Watanabe qui distingua clairement la fonction produisant le caractère unifié de celle qui entraîne le caractère terminé dans une phrase.

Définition de Watanabe

Selon Watanabe, la constitution de la phrase est terminée lorsque la fonction d'effet phrasogénérateur, *chinjutsu*, opère sur la matière qu'est le contenu – 叙述内容 (*jojutsu naiyô*) –, qui représente une pensée ou un fait, unifié par des fonctions particulières.

Watanabe restreint ainsi la notion de *chinjutsu* qui était assez large et surtout floue. Selon lui (Watanabe, 1971), il s'agit :

> « [d']une fonction syntaxique qui détermine, en prenant le 叙述内容 (*jojutsu naiyô*) [le contenu unifié, d'une pensée ou d'un fait] comme matière, le rapport entre ce contenu et le sujet parlant. »

Le caractère terminé d'une phrase provient donc de cette réalisation de *chinjutsu*. En revanche, le caractère unifié est obtenu à l'intérieur de *jojutsu*, c'est-à-dire avant le déclenchement de *chinjutsu*.

L'élément *jojutsu* représentant une pensée ou un fait, est constitué d'éléments ayant une fonction de développement (展叙, *tenjo*) et d'autres ayant une fonction d'unification (統叙, *tôjo*).

Afin de mieux illustrer ces propos, considérons la phrase suivante.

桜　の　花　が　咲く。
(*sakura - no - hana - ga - saku*)
(cerisier - de - fleur - [*ga*] - s'épanouir [non passé])
(Les fleurs de cerisiers s'épanouissent.)

1. ***sakura* + *no*** : le deuxième élément, indicateur de relation *no*, s'attache au premier élément, représentant la matière « cerisier », constituant un nouvel élément doté d'une fonction de développement.
2. ***sakura - no*** **+ *hana*** : ce nouveau bloc constitue ensuite avec le troisième élément, représentant la matière « fleur », un nouvel élément qui représente lui-même la matière « fleur de cerisier ».
3. ***sakura - no - hana*** **+ *ga*** : le quatrième élément *ga* donne à son tour une fonction de développement à ce nouveau groupe.
4. ***sakura - no - hana - ga*** **+ *saku*** : ce bloc doté d'une fonction de développement s'associe avec le contenu du prédicat « s'épanouir ».

5. ***sakura - no - hana - ga - saku*** : la fonction d'unification que possède également le prédicat, regroupe l'ensemble de ces éléments, lui donnant ainsi un caractère unifié.

Ainsi, le caractère unifié d'une phrase provient de l'exécution de cette fonction d'unification.

En plus du contenu « s'épanouir » et de la fonction d'unification, le prédicat *saku* possède une capacité phrasogénératrice. Lorsque cet effet phrasogénérateur opère sur l'ensemble unifié « *sakura - no - hana - ga - saku* », cet ensemble obtient un caractère terminé, devenant ainsi une phrase.

Teramura considère cette notion de *jojutsu naiyô* de Watanabe comme une notion comparable aux « dictum » de Bally et « proposition » de Fillmore (cf. § 5.5.1).

6.1.3 Caractère incomplet de la phrase japonaise

Pour définir la phrase japonaise, Mikami, qui accorde toujours de l'importance à la forme, a remis en question la validité même du caractère complet de la phrase japonaise (Mikami, 1963a).

Après avoir cité des passages de Meillet (1903) et de Bloomfield (1970) parlant du caractère autonome et complet de la phrase, Mikami affirme la nécessité de modifier cette partie de la définition de la phrase pour le japonais. En effet, la portée de *wa* peut franchir facilement la limite de phrase (le point final à l'écrit). Mikami appelle souvent cette particularité du thème, capacité à « dépasser le point final ».

De plus, les éléments susceptibles d'être omis ne se limitent pas au thème. Le japonais est en fait une langue dont la dépendance au contexte et à la situation est extrêmement élevée, et grossièrement, tous les éléments jugés connus par l'interlocuteur peuvent ne pas être exprimés explicitement[2].

Cette dépendance, une des caractéristiques principales du japonais[3], pose des problèmes cruciaux lors de l'analyse automatique, surtout lorsqu'il s'agit d'une application multilingue avec des langues dans lesquelles tous les éléments obligatoires syntaxiquement sont marqués explicitement, telles que le français[4].

[2] On trouve dans Mikami (1970) des études plus approfondies très intéressantes sur les règles d'omission des éléments de la phrase.

[3] La présence d'éléments ayant une portée plus large qu'une phrase n'est pas un caractère spécifique de la phrase japonaise. Les introducteurs du cadre, par exemple, qui « sont à même de fixer des cadres regroupant une ou plusieurs propositions » (Charolles, 1997), peuvent avoir une portée beaucoup plus large qu'une phrase. En revanche, le fait que les éléments dépassant les frontières de phrases ne se limitent pas à des constituants particuliers tels que les introducteurs du cadre ou les thèmes, constitue une particularité de la langue japonaise.

[4] Il est toutefois à noter que certaines études, en ne citant qu'une des plus récentes (Akihiro, 2004), montrent que dans des contextes particuliers, très limités et probablement bien définis, le complément peut ne pas être réalisé dans la phrase française. Cependant, il est important de souligner que ces omissions ne sont pas du même ordre en français et en japonais. Différente en japonais dans laquelle elle est très fréquente voire systématique, cette omission est très restreinte en français,

6.2 État de l'art II : structure multicouche de la phrase japonaise

La distinction entre le dictum et le modus a amené les linguistes à formuler l'hypothèse que dans la phrase japonaise, le dictum constitue le centre, le modus l'enrobant en se mettant à sa périphérie. Alors, se sont développées des études sur la structure multicouche de la phrase japonaise[5].

Ces travaux ont d'abord donné comme résultats quelques distinctions élémentaires des éléments appartenant au modus avant de proposer quatre couches concrètes qui constituent la phrase japonaise.

Nous allons maintenant passer en revue les travaux pionniers avant de nous intéresser aux études définissant ces quatre couches.

6.2.1 Les premiers travaux

Haga propose d'abord de distinguer les éléments du modus en deux types.

Par ailleurs, dans les travaux de Kitahara et Sakakura, on trouve des schémas présentant clairement la structure multicouche de la phrase japonaise. Cependant, dans ces travaux, ne figurait pas encore de définition des niveaux de ces couches.

Haga

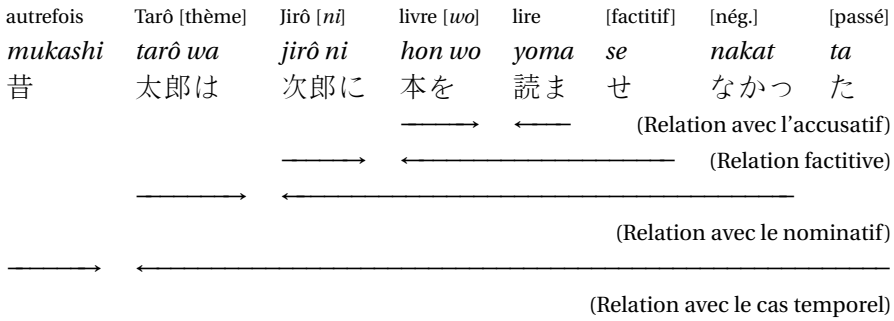
En retravaillant sur la notion de modus de Watanabe (cf. § 6.1.2), Haga (1954) a distingué les éléments du modus en deux classes : éléments représentant les attitudes du locuteur sur le contenu (conjecture, etc.) ; éléments visant à produire directement un effet sur l'interlocuteur (ordre, invitation, etc.). Cette distinction est reprise par beaucoup de linguistes. Nous appelons par commodité le premier type « modus orienté vers le contenu » et le second « modus orienté vers l'interlocuteur ».

Kitahara

Sur la base de la théorie de Watanabe, Kitahara (1976) analyse la phrase « Autrefois Tarô ne laissait pas Jirô lire de livres » comme une structure à quatre couches :

ne pouvant apparaître que sous certaines conditions. Le choix sur le complément à supprimer est particulièrement contraignant à tel point que l'on peut parler de l'incorporation du complément dans le verbe, voire de l'intransitivation des verbes.

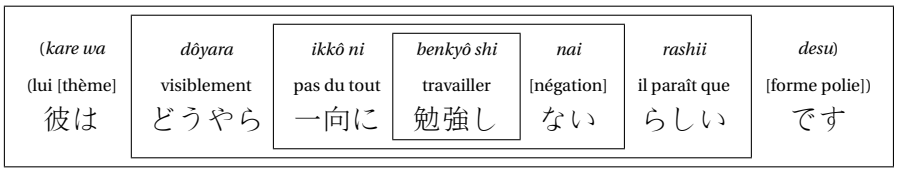
[5]Nous trouvons une description détaillée et comparative de l'ensemble de ces travaux dans les ouvrages de Minami (1974, 1993).



La racine du mot prédicatif « *yoma* (lire) » s'associe avec le complément accusatif « *hon wo* (livre [wo]) » pour constituer la toute première couche. Ce premier noyau transformé en factitif par l'ajout du suffixe « *se* » entre en relation avec le complément en *ni* « *jirô ni* » indiquant le destinataire et constitue ainsi la deuxième couche. À cette dernière, s'ajoute l'élément de négation « *nakatt* » et l'ensemble ainsi constitué prend le nominatif « *tarô wa* » pour composer la troisième couche. Enfin, elle est complétée par l'élément de temps « *ta* » marquant le caractère temporel passé (ou accompli) et entre en relation avec le complément de temps « *mukashi* (autrefois) ».

Sakakura

Sakakura (1979) analyse également les phrases à partir des couches qui les constituent. Mais, il s'intéresse également aux adverbes qui s'accompagnent d'éléments modaux. La phrase « Il paraît qu'il ne travaille pas du tout » est analysée comme suit :



La racine du mot prédicatif « *benkyôshi* (étudier) » est tout d'abord enrobée par l'élément de négation « *nai* » et l'adverbe accompagnant la négation « *ikkôni* (pas du tout) ». Cet ensemble est enchâssé à son tour par l'élément de modalité « *rashii* (il paraît que) » et l'adverbe accompagnant cet élément « *dôyara* (visiblement) ». Enfin, cette construction est recouverte par la dernière couche y ajoutant le thème « *kare wa* (lui [wa]) » et l'élément marquant la politesse « *desu* ».

6.2.2 Définition des quatre niveaux constituant la phrase japonaise

L'analyse par la structure multicouche de la phrase japonaise a conduit à la définition de quatre niveaux d'éléments de la phrase japonaise. Ces niveaux comprennent cependant des éléments plus ou moins différents selon les linguistes.

Hayashi

Hayashi (1960) scinde la structure du prédicat japonais en quatre couches et définit les contenus représentés dans chaque couche comme suit :

1. niveau de description ;
2. niveau de jugement : assertion/négation, possibilité/impossibilité, temps, conjecture, interrogation, etc. ;
3. niveau d'expression : admiration, espoir, souhait, inquiétude, volonté, décision, etc. ;
4. niveau de communication : transmission simple, ordre, exigence, demande, question, etc.

Mikami

Pour comprendre la syntaxe de la phrase japonaise, Mikami a pensé qu'il était indispensable de comprendre les différentes relations qu'entretiennent les syntagmes, et notamment les relations entre les syntagmes ayant comme tête un mot variable (syntagme à mot variable ci-après).

Il définit cinq types de connexions :

1. style simple : connexion se limitant à l'intérieur d'un syntagme à mot variable ;
2. style complexe : connexion entre un syntagme à mot variable et les éléments extérieurs à ce syntagme ;
   a) style souple : connexion pouvant glisser facilement vers le style simple ;
   b) style dur : connexion entre deux éléments ne réalisant jamais de connexion de style simple ;
3. style flottant : connexion entre un syntagme à mot variable et un élément n'entretenant aucune relation syntaxique avec lui ;
4. style fermé : connexion du type déterminant et connexion du discours rapporté.

Les particules de cas appartiennent selon l'auteur au style simple, alors que les particules de mise en relief appartiennent au style complexe. Les interjections et les mots de liaison, qui constituent les éléments indépendants selon la grammaire scolaire, sont considérés comme des éléments du style flottant.

Mikami met l'accent en particulier sur la détermination de la nature de la connexion que produit chaque forme de mot variable (cf. § 7.3.1). Dans ce but, Mikami a proposé trois critères concrets :

– la forme de mot variable concernée empêche-t-elle ses compléments de franchir la limite de la structure dans laquelle elle est prédicat ?
– la structure qu'elle forme en tant que prédicat avec ses compléments peut-elle appartenir à une subordonnée déterminante ?

– se transforme-t-elle en forme polie lorsque le prédicat principal est transformé en forme polie ?

Ces travaux de Mikami ne concrétisent pas encore la structure en couches de la phrase japonaise, mais ses résultats et ses réflexions ont inspiré Minami qui a, par la suite, construit la base de cette théorie soutenue par la plupart des linguistes japonais d'aujourd'hui.

La plus grande contribution de Mikami sur ce sujet a sans doute été le fait que ses travaux montraient la méthodologie correcte, permettant ainsi de rendre concrète et descriptible cette idée de différentes couches, idée qui aurait pu retomber autrement dans une notion conceptuelle.

Minami

Minami (1974, 1993) définit lui aussi quatre couches constituant l'ensemble de la structure phrastique : A (niveau de description), B (niveau de jugement), C (niveau d'émission) et D (niveau de manifestation). Il travaille, dans ce but, sur la possibilité d'apparition ou non de certains éléments dans différentes structures subordonnées. Il analyse la structure de chaque type de phrase de manière très détaillée. L'analyse générale des différents éléments de phrase est présentée figure 6.1.

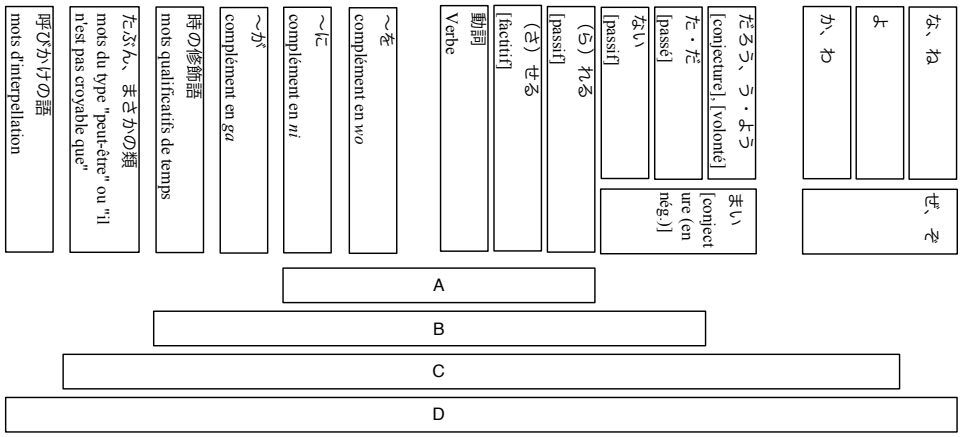


FIG. 6.1 – Analyse de la structure de phrase par Minami

Teramura

Teramura distingue d'abord la phrase en deux parties, le dictum et le modus. Comme nous pouvons le constater figure 6.2 (voir page suivante), il exclut le thème et le temps du dictum. Le niveau auquel appartiennent le thème et le temps

est celui du modus de déclaration où le locuteur exprime un jugement concret sur un fait, sortant du cadre du dictum, abstrait et conceptuel. Ce niveau semble correspondre au niveau du jugement de Minami, mais la différence est que ce dernier inclut le thème non pas dans le niveau du jugement mais dans celui de l'émission, un niveau supérieur à celui du jugement.

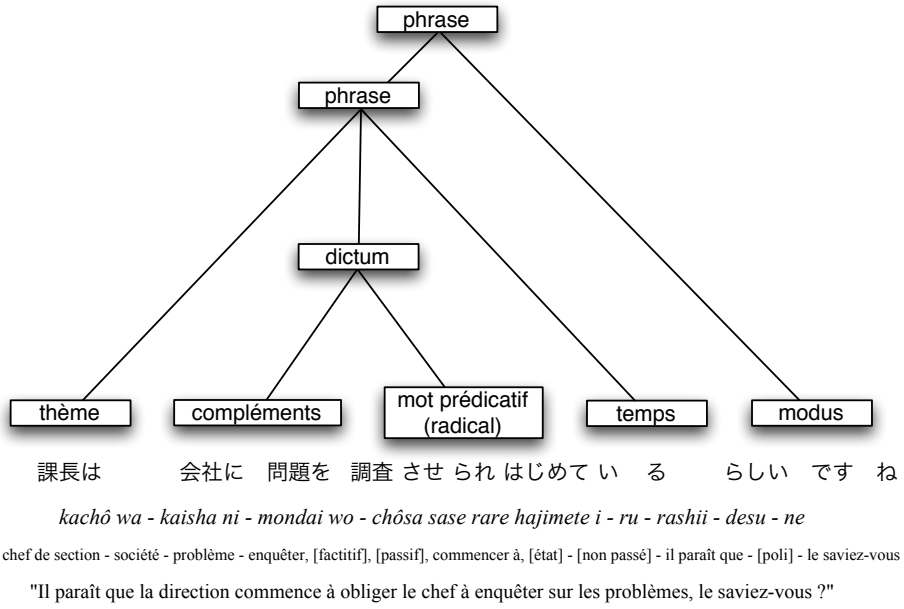


FIG. 6.2 – Analyse de la structure de phrase par Teramura

Dans la théorie de Teramura (1982b, 1984), le modus est catégorisé en trois types : modus primaire, modus secondaire et modus tertiaire.

Le modus de déclaration tel que le temps est appelé modus primaire. C'est le seul type de modus obligatoire au niveau syntaxique et il est représenté par différentes formes de mots variables. Le modus secondaire est celui orienté vers le contenu et le modus tertiaire, celui orienté vers l'interlocuteur.

Grammaire fonctionnelle de Dik

Masuoka (1997) introduit la notion des quatre niveaux conceptuels de Minami dans ses travaux sur la phrase complexe du japonais. Tout en affirmant que ce concept de structure en couches de la phrase développée est original et particulier à la linguistique japonaise, il ne le considère pas comme une particularité de la langue japonaise. Ainsi, il insiste sur le fait qu'on trouve des travaux non japonais présentant une notion assez proche, en particulier les travaux de la grammaire fonctionnelle de Dik (1989). Dik définit quatre niveaux dans la phrase (cf. tableau

6.3), dont le deuxième, le troisième et le quatrième contiennent respectivement des éléments marquant le temps, des éléments exprimant le jugement du locuteur sur le contenu propositionnel et des éléments exprimant la force illocutoire (*illocutionary force*). Ce qui correspond quasiment aux quatre couches définies par Minami.

	Layer	Entity type
1	predicate	property/relation
	term	entity
2	predication	state of affairs
3	proposition	possible fact
4	clause	speech act

TAB. 6.3 – Niveaux et types d'entités dans la grammaire fonctionnelle

6.3 Typologie des phrases japonaises

6.3.1 Opposition des phrases « avec-thème » et « sans-thème »

Beaucoup de linguistes soulignent l'importance de la distinction entre les phrases avec thème et sans thème pour le japonais. En effet, bien qu'il soit possible de considérer l'opposition thème-rhème comme la base de la structure de la phrase japonaise, il en existe certaines démunies de thème. Ce sont des phrases appelées généralement « de description »[6].

Toutefois, les phrases avec thème ne contiennent pas forcément un syntagme en *wa*. Mikami définit alors quatre types de phrases – trois types de phrases avec thème et un sans thème – comme suit (repris de Mikami (1970)) :

1. Phrase avec thème :
    a) phrase avec thème (explicite) ;
    b) phrase à thème caché (implicite) ;
    c) phrase à thème omis (elliptique) ;
2. Phrase sans thème (entièrement rhématique).

[6]Mikami fait un rapprochement entre les phrases sans thème de la linguistique japonaise et les phrases constituées entièrement d'information nouvelle de Chafe dans Mikami (1970) (avec comme référence Chafe (1970)) :

> W. Chafe de l'Université de Californie a distingué les éléments de phrase en ancien et nouveau, et il a affirmé que la phrase *John opened the door* peut être entièrement nouvelle (d'après une lettre personnelle de Monsieur Teramura). Les occidentaux semblent également avoir enfin découvert la notion de sans-thème.

Phrases à thème explicite, implicite et phrase sans thème

Mikami (1953) illustre leur différence (sauf la phrase elliptique) avec trois phrases ayant toutes le même contenu sémantique « Henri est arrivé » mais possédant des différences formelles – notamment la présence ou non de la particule *wa* – selon le contexte (traduction de Kawamoto (1958)) :

1. **phrase avec thème explicite**
   - 扁理は　どうした？
     (henri wa - dôshita)
     Qu'est-ce que devient Henri ?
   - **扁理は**　到着しました
     (henri wa - tôchaku shimasita)
     Henri est arrivé. (Henri, il est arrivé, ou Il est arrivé, Henri.)
2. **phrase à thème caché**
   - 誰が　到着した？
     (dare ga - tôchaku shita)
     Qui est arrivé ?
   - **扁理が**　到着したんです
     (henri ga - tôchaku shita n desu)
     Henri est arrivé. (C'est Henri qui est arrivé.)
3. **phrase sans thème**
   - 何か　ニュースは　ないか？
     (nanika - nyûsu wa - naika)
     Y a-t-il quelque chose de nouveau ?
   - **扁理が**　到着しました
     (henri ga - tôchaku shimashita)
     Henri est arrivé.

Dans la phrase à thème caché, le thème est implicite car elle peut être paraphrasée par une phrase à thème explicite. La phrase d'exemple **扁理が**到着したんです *(**henri ga** tôchaku shita n desu)* peut être paraphrasée par :

到着した　の　は　扁理です
(tôchaku shita - no - wa - henri desu)
(arriver [passé] - [nominalisateur] - wa - Henri [copule, politesse])
Celui qui est arrivé est Henri.

Si bien qu'on considère la phrase **扁理が**到着したんです *(**henri ga** tôchaku shita n desu)*, malgré l'absence de syntagme thématisé par *wa*, non pas comme sans thème mais comme à thème caché.

Dans ce type de phrase : « N *ga* V » interprétée comme « ce qui réalise l'acte V est N », la particule *ga* joue une fonction de désignation. Kuroda (1973) a appelé cet effet provoqué par *ga* de désignation « effet de focus » qu'il a emprunté à Chomsky. Mais cet effet de focus n'est pas la spécialité exclusive de la particule

ga. Mikami parle par exemple du cas où il se produit sur le syntagme en *ni* dans Mikami (1963b).

La désignation entraîne souvent une nuance d'exclusivité qui nie l'appartenance des autres candidats à la catégorie caractérisée par la réalisation de V. Dans le cas de l'exemple, cette fonction de *ga* peut, selon le contexte, sous-entendre le rejet d'autres candidats que « Henri ».

Phrase à thème omis

Ce sont des phrases dénuées de *wa* qui portent sur un thème déjà présenté dans le contexte, c'est-à-dire sur le thème de la phrase précédente ou même plus antérieure.

Mikami (1963a) explique la phrase à thème omis comme suit :

« a) 鯨　wa　けもの　だ。　魚　ではない。
(*kujira - wa - kemono - da. sakana - de wa nai*)
(baleine - [thème] - animal - [copule]. poisson - [copule, négative].)
« Les baleines sont des animaux. Ce ne sont pas des poissons. »

⋮

Dans l'exemple (a), on peut considérer que "鯨wa" (*kujira - wa*) fonctionne deux fois et forme le sens de "鯨　wa　魚　ではない。" (*kujira - wa - sakana - de wa nai*, "Les baleines ne sont pas des poissons."). Nous appelons les phrases telles que cette deuxième, des **phrases à thème omis**. »

Toutefois, il nous semble difficile de distinguer les phrases à thème omis de celles sans thème. En effet, l'exemple précédemment présenté :

– 何か　ニュースは　ないか？
(nanika - nyûsu wa - naika)
Y a-t-il quelque chose de nouveau ? (ou plus littéralement « Y a-t-il des nouvelles ? »)
– **扁理が**　到着しました
(henri ga - tôchaku shimashita)
Henri est arrivé.

est considéré comme un exemple de phrase sans thème, mais il est possible, semble-t-il, de considérer ニュース (nyûsu, nouvelles) comme le thème et dans ce cas ce serait plutôt un exemple de phrase à thème omis.

6.3.2 Typologie selon la catégorie du prédicat

La phrase japonaise se caractérise également selon la nature de ses éléments constituants. Plusieurs linguistes (Yamada, 1908 ; kokuritsu kokugo kenkyû jo, 1963 ; Mikami, 1963a ; Minami, 1993) la scindent d'abord en deux types : phrase de mot indépendant et phrase avec prédicat. Masuoka & Takubo (1992) appellent

les premières 未分化文 (*mibunkabun*, phrases non-développées) et les secondes 分化文 (*bunkabun*, phrases développées).

Le second type, phrase avec prédicat, se décompose lui-même en trois types selon le mot prédicatif qu'il contient.

1. phrase de mot indépendant ;
2. phrase avec prédicat :
   a) phrase verbale ;
   b) phrase substantive ;
   c) phrase adjective : ayant comme mot prédicatif un qualificatif en *i* ou un qualificatif en *na*.

Phrase de mot indépendant et phrase avec prédicat

Les phrases du premier type, celles démunies de mot variable, *dokuritsugo-bun* (独立語文, phrase de mot indépendant), expriment un certain sentiment exclamatif ou un appel[7] :

アッチッチ。
(*attitti*)
« Aïe ! C'est chaud »

オーイ、下田君。
(*ôi - shimoda kun*)
(coucou - Shimoda (nom de famille) [titre pour un camarade ou un collègue])
« Eh ! M. Shimoda ! »

ウウン。
(*uun*)
(non [familier])
« Non »

Les phrases du second type appelées *jutsugo-bun* (述語文, phrase avec prédicat), regroupent les phrases munies comme élément central d'un mot prédicatif telles que :

雨　が　降って　いる。
(*ame - ga - fut te - iru*)
(pluie - [nominatif] - tomber - [aspect progressif])
« Il pleut »

バルザック　は　フランス人　だ。
(*baruzakku - wa - furansu jin - da*)
(Balzac - [thème] - français - [copule])
« Balzac est français »

[7] Exemples tirés de Minami (1974).

パリ　の　街　は　とても　美しい。
(*pari - no - machi - wa - totemo - utsukushii*)
(Paris - de - ville - [thème] - très - beau)
« La ville de Paris est très belle »

Selon le mot prédicatif qu'elles contiennent (verbe, substantif ou qualificatif), la nature de ces phrases diffère.

Étudions maintenant chacun de ces trois types de phrases avec prédicat.

Phrases verbales

La particularité des phrases verbales est qu'elles peuvent être sans thème. Du fait de l'existence de ces phrases verbales sans thème, Mikami considère que seul le prédicat constitue le pilier essentiel de la phrase japonaise.

Il existe en effet beaucoup de phrases verbales véritablement démunies de thème. Une phrase décrivant une scène qui se déroule sous nos yeux ne contient pas la particule *wa* ; c'est une phrase sans thème. Par exemple, quand on voit un canard voler, on dit :

あっ、鴨　が　飛んで　いる。
(*at - kamo - ga - tonde - iru*)
(tiens ! - canard - [nominatif] - voler - [progressif])
« Tiens ! Il y a un canard qui vole »

De même, les articles d'une dépêche destinée à la transmission d'informations nouvelles sont des phrases dénuées de *wa* sans thème.

三宅島　で　小規模の　噴火　が　発生し　た。
(*miyake jima - de - shôkibo no - funka - ga - hassei shi - ta*)
(île de Miyake - [lieu] - de petite taille - éruption - [nominatif] - se produire - [passé])
« Il y a eu une petite éruption sur l'île de Miyake »

一日午後、近畿地方　で　地震　が　あり、
京都府　や　滋賀県　など　で　震度３　の　揺れ　を　記録し　た。
(*tsuitachi gogo - kinki chihô - de - jishin - ga - ari -*
kyôtofu - ya - shigaken - nado - de - shindo san - no - yure - wo - kiroku shi - ta)
(après-midi du premier (du mois) - région du Kinki - [lieu] - séisme - [nominatif] - avoir lieu - dép. Kyoto - et - dép. Shiga - etc. - [lieu] - niveau 3 - de - secousse - [accusatif] - enregistrer - [passé])
« Le premier dans l'après-midi, il y a eu un tremblement de terre dans la région du Kinki ;
des secousses de niveau 3 ont été enregistrées dans les départements de Kyoto et de Shiga »

Néanmoins, la détermination – surtout automatique – de la nature des phrases verbales sans syntagme en *wa* est très délicate, car elles peuvent également être des phrases à thème caché si elles ont un ou plusieurs compléments ou des phrases à thème omis si elles sont précédées d'autres phrases.

Phrases substantives

Contrairement aux phrases verbales, Mikami ne reconnaît pas de phrase substantive sans thème. Il considère qu'une phrase substantive est constituée sur la

base de l'opposition thème-rhème et qu'on constate nécessairement un effet particulier dans les phrases dénuées de *wa*.

Le modèle pour la phrase substantive japonaise est donc : « SN_1 *wa* SN_2 *da* ».

Phrases de base avec *wa* Elles se classent en deux types : phrases exprimant un jugement sur SN_1 et phrases exprimant une désignation concernant SN_1.

1. phrases de jugement : A *wa* B *da*.
   a) A = B
      東京　は　日本　の　首都　だ。
      (*tôkyô - wa - nihon - no - shuto - da*)
      (Tokyo - [thème] - Japon - de - capitale - [copule])
      « Tokyo est la capitale du Japon »
   b) A ∈ B
      クジラ　は　哺乳類　だ。
      (*kujira - wa - honyûrui - da*)
      (baleine - [thème] - mammifère - [copule])
      « Les baleines sont des mammifères »
2. phrases de désignation : B *wa* A *da*.
   a) A = B
      日本　の　首都　は　東京　だ。
      (*nihon - no - shuto - wa - tôkyô - da*)
      (Japon - de - capitale - [thème] - Tokyo - [copule])
      « La capitale du Japon, c'est Tokyo »
   b) A ∈ B
      ? 哺乳類　は　クジラ　だ。
      (*honyûrui - wa - kujira - da*)
      (mammifère - [thème] - baleine - [copule])
      « Les mammifères, ce sont les baleines » (en situation : par exemple, pour répondre à la question demandant lesquels des baleines et des requins sont des mammifères.)
      Lorsque A est un élément de l'ensemble B, la phrase de désignation « B *wa* A *da* » est impossible en phrase générique sans un contexte particulier. La détermination de B telle que 海に住む哺乳類 (*uminisumu honyûrui*, mammifères vivant dans la mer), qui rend l'ensemble B restreint voire réduit à un seul élément comme 日本の首都 (*nihon no shuto*, capitale du Japon), rend valide la phrase générique de désignation « B *wa* A *da* », mais la relation entre A et B n'est plus A ∈ B mais plutôt A = B.

Considérons maintenant les même exemples où la particule *wa* est remplacée par *ga* afin d'étudier les effets particuliers qui se produisent.

Phrases avec *ga* Considérons les phrases suivantes créées à partir des phrases 1a, 1b, 2a et 2b, dans lesquelles la particule *wa* est remplacée par *ga*.

1'. A *ga* B *da*. (A = B ou A ∈ B) ≃ phrase de désignation « B *wa* A *da* »

a) 東京　が　日本　の　首都　だ。
(*tôkyô - ga - nihon - no - shuto - da*)
(Tokyo - [ga] - Japon - de - capitale - [copule])
« C'est Tokyo qui est la capitale du Japon »
C'est par exemple une réponse à la question : quelle ville est la capitale du Japon ? On peut la paraphraser par la phrase 2a. Mikami considère ce type de phrase comme phrase à thème caché, comme pour les phrases verbales.

b) ? クジラ　が　哺乳類　だ。
(*kujira - ga - honyûrui - da*)
(baleine - [ga] - mammifère - [copule])
« Ce sont les baleines qui sont des mammifères »
Les contextes dans lesquels cette phrase peut être valide (où l'on désigne les baleines en cherchant les espèces de mammifères) sont limités tout comme la phrase de désignation équivalente 2b.

2'. B *ga* A *da*. (A = B ou A ∈ B) ? ≃ phrase de jugement « A *wa* B *da* »

a) 日本　の　首都　が　東京　だ。
(*nihon - no - shuto - ga - tôkyô - da*)
(Japon - de - capitale - [ga] - Tokyo - [copule])
« La capitale du Japon. C'est ça Tokyo »
Il semble apparaître un autre effet particulier dans cette structure, notamment lorsque B est un nom propre : on parle de B, de plusieurs de ses particularités et on présente finalement A comme le point le plus caractéristique de B. C'est une paraphrase de l'exemple 1a avec un sens supplémentaire du type « avant tout ».

b) ? 哺乳類　が　クジラ　だ。
(*honyûrui - ga - kujira - da*)
(mammifère - [ga] - baleine - [copule])
« C'est le mammifère, la baleine » (en situation : par exemple, on demande en montrant les images d'une baleine et d'un requin d'indiquer le poisson et le mammifère. Une des personnes interrogées affirme « *gyorui wa kujira de honyûrui wa same da* (le poisson, c'est la baleine et le mammifère, c'est le requin) ». Alors une autre récuse cette réponse : « *iya, gyorui ga same de honyûrui ga kujira da* (non, c'est le poisson le requin, et c'est le mammifère, la baleine) »)
L'interprétation de cette phrase comme dans l'exemple précédent du type « avant tout » est impossible. Le seul contexte possible est la situation décrite après la traduction de l'exemple. Mais, la détermination exacte du sens ou de l'effet énonciatif produit par cette forme nécessite plus d'études spécifiques, que nous ne réaliserons pas dans le cadre de la présente thèse.

Ainsi, lorsque *wa* disparaît dans une phrase substantive, et que la particule *ga* apparaît, on constate un effet particulier sur l'élément introduit par *ga*.

Il est à noter que Minami (1993) exclut de cette catégorie certaines phrases formellement substantives et les classe dans une autre catégorie qu'il appelle « **phrase pseudo-substantive** ». Cette classe contient deux types de phrases. Le premier type de phrase est celui ayant comme prédicat un type de substantif particulier que Matsushita a appelé verbe invariable. Ce sont généralement des mots provenant du chinois, exprimant une action tels que 出張 (*shutchô*, déplacement professionnel) :

父　は　明日　から　大阪　へ　出張　だ。
(*chichi - wa - asu - kara - ôsaka - e - shutchô - da*)
(mon père - [thème] - demain - [point de départ] - Osaka - [destination] - déplacement professionnel - [copule])
« Mon père sera en déplacement professionnel à Osaka à partir de demain »

Il en existe également ayant comme prédicat un mot emprunt d'une langue autre que le chinois tel que ストップ (*sutoppu*, stop ; arrêt) :

経済成長率　は　3.5　パーセント　で　ストップ　だ。
(*keizai seichô ritsu - wa - 3,5 - pâsento - de - sutoppu - da*)
(taux de croissance économique - [thème] - 3,5 - pour cent - [lieu] - stop - [copule])
« Le taux de croissance économique s'est arrêté à 3,5 pour cent »

Le deuxième type correspond à des phrases que Noda (1998) appelle phrase avec thème à cas cassé (破格主題文, *hakaku shudai bun*, voir dans la section 6.4.3 la page 238). Ces phrases sont caractérisées par le fait que la relation syntaxique entre le syntagme thématisé et les éléments du prédicat n'est pas évidente. La relation sémantique entre ces éléments n'est déductible que de connaissances extra-linguistiques, voire du contexte.

Minami caractérise ces phrases par le fait que la structure du rhème ressemble à celle de la phrase verbale.

Par ailleurs, Mikami fait remarquer que certaines phrases verbales, telles que les phrases négatives, se rapprochent de la phrase substantive.

Phrases adjectives

Mikami inclut les phrases adjectives, comme les phrases substantives, dans la classe des phrases nominatives. Il définit ainsi comme forme de base pour les phrases adjectives, le modèle « A *wa* B », tout comme pour les phrases substantives.

Mais, pour la description d'un ciel bleu, la phrase « *sora* (ciel) *ga aoi* (bleu) » semble plus adéquate que la phrase « *sora* (ciel) *wa aoi* (bleu) » : cette dernière est plutôt interprétable comme un jugement générique sur la couleur du ciel.

Nous ne nous prononcerons pas sur le rattachement ou non des phrases adjectives aux phrases nominatives, ce point n'ayant pas d'influence sur la suite de nos travaux.

6.4 Syntagme thématisé et particule *Wa*

Nous allons tout d'abord rappeler la différence entre les particules de cas et les particules adverbiales (§ 6.4.1). Nous présenterons ensuite notre position vis-à-vis de la question du mécanisme de génération du thème (§ 6.4.2) avant d'éclaircir la double fonction du syntagme thématisé (§ 6.4.3). Enfin, dans le dernier paragraphe, nous traiterons des syntagmes en *wa* non-thématiques (§ 6.4.4) et du thème réalisé par un autre moyen que la particule *wa* (§ 6.4.5).

6.4.1 Particules de cas et particules adverbiales

Comme nous l'avons déjà mentionné, le japonais dispose d'un mot grammatical indiquant le thème (ce dont on parle) – la particule *wa* (は) –, en plus de celui dédié à marquer la fonction dite « sujet » – la particule *ga* (が).

Beaucoup de linguistes travaillaient et travaillent encore sur la différence entre ces deux particules, *wa* et *ga*. Certains de ces travaux sont très connus même en dehors du Japon, comme par exemple l'article de Kuroda (1973) qui a essayé de montrer l'exactitude de la distinction du jugement thétique et du jugement catégorique faite par Franz Bretano et Anton Marty. Mais cette façon de traiter ce problème entraîne parfois une confusion : en effet, ces deux particules sont des unités de niveau fort différent.

La particule *ga* appartient à la sous-catégorie des particules appelée 格助詞 (*kaku-joshi*, particule de cas), qui regroupe les particules indiquant la fonction syntaxique du syntagme qui les précède, comme nous l'avons déjà présenté dans la section 5.7.1. En revanche, la particule *wa* appartient, selon la grammaire scolaire, à la sous-catégorie des particules appelée 副助詞 (*fuku-joshi*, particule adverbiale) qui regroupe les particules ayant comme fonction d'ajouter une valeur sémantique ou énonciative supplémentaire[8]. La particule *wa* a comme fonction de transformer le syntagme qu'elle précède en thème à propos duquel on parle[9].

Les particules adverbiales peuvent suivre différents syntagmes, y compris les syntagmes postpositionnels terminés par une particule de cas. Si bien que l'on peut tout à fait avoir une séquence avec une particule de cas suivie d'une ou même deux particules adverbiales. Cependant, les particules *ga* et *wo*, lorsqu'elles sont suivies de la particule *wa*, disparaissent, ce qui provoque souvent une confusion quant aux niveaux des particules de cas et des particules adverbiales, en fait très différents.

En tenant compte de cette différence entre particule de cas et particule adverbiale, nous devons, pour être corrects, parler non pas de la différence entre les

[8] La catégorisation des particules varie également selon les linguistes. Certains définissent, en plus des particules de cas, deux, voire trois types. Nous soulignons ici la distinction la plus importante entre les particules de cas et les autres et présentons seulement la définition de la grammaire scolaire qui regroupe toutes les particules autres que celles de cas sous le nom de particules adverbiales.

[9] On parle également de la fonction contrastive de la particule *wa*. Nous aborderons, comme nous l'avons annoncé, cette autre fonction de *wa* dans la section 6.4.4.

phrases avec le syntagme en *ga* et celles avec le syntagme en *wa*, mais de la différence entre les phrases avec et sans syntagme en *wa*.

6.4.2 Génération du thème

La génération du thème est également une question fondamentale pour une analyse des phrases à structure thème-rhème et pour laquelle les linguistes japonais ne partagent pas tous la même interprétation. D'après les études réalisées par Noda (1998), existent quatre théories pour la façon dont se génère le syntagme thématisé, comme suit[10] :

1. génération par déplacement (Kuroda, 1965) ;
2. génération par reproduction (Muraki, 1974) ;
3. génération à la base (Kuno, 1973) ;
4. méthode combinée des générations par déplacement et à la base (Shibatani, 1978).

Dans la génération par déplacement (cf. figure 6.4) ou par reproduction (cf. figure 6.5 page suivante), est d'abord produite une structure avec un mot prédicatif et ses compléments, puis la mise en thème d'un certain élément est réalisée par une procédure supplémentaire (respectivement déplacement et reproduction).

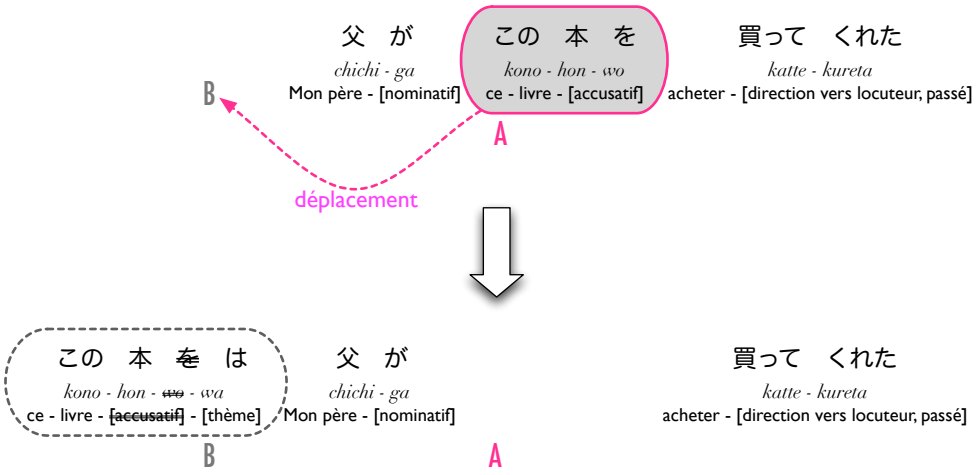


FIG. 6.4 – Génération du thème par déplacement

En revanche, selon la théorie de la génération à la base, le thème existe déjà au niveau profond (c.f. figure 6.6 page ci-contre).

Kuno (1973) défend cette position du fait, d'une part, de l'existence de phrases dans lesquelles le syntagme thématisé est repris par un pronom :

[10] Les références sont également reproduites de Noda (1998).

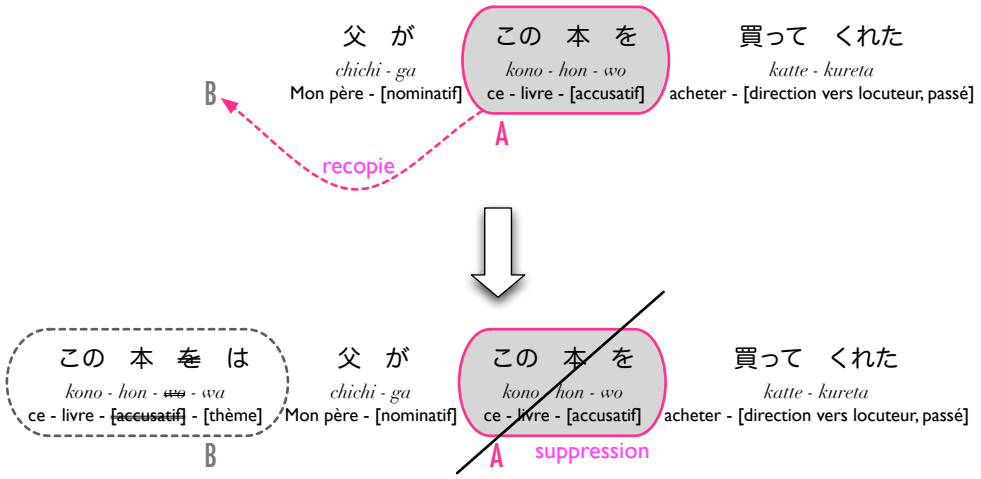


FIG. 6.5 – Génération du thème par reproduction

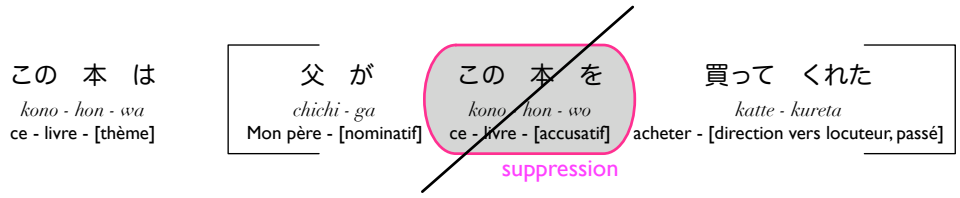


FIG. 6.6 – Génération du thème à la base

太郎は 彼が 書いた 本が ベストセラーになっている
(*tarô wa - kare ga - kaita - hon ga - besutoserâ ni natteiru*)
(Tarô [thème] - il [nominatif] - écrire [passé] - livre [nominatif] - être devenu un best-seller)
« Tarô, le livre qu'il a écrit est devenu un best-seller »

et, d'autre part, de la présence de phrases dans lesquelles la fonction syntaxique du syntagme thématisé vis-à-vis du mot prédicatif est difficile à déterminer comme dans la phrase :

魚は 鯛が いい
(*sakana wa - tai ga - ii*)
(poisson [thème] - dorade [nominatif] - bien)
« Les poissons, c'est la dorade qui est la meilleure ».

Noda (1998) qui soutient la thèse de la génération par reproduction, analyse ce type de phrase par une omission ou par une redondance et propose de les analyser comme des anomalies de surface.

Nous, qui considérons l'opposition thème-rhème comme la structure fondamentale de la phrase japonaise, soutenons la théorie de la génération à la base.

Mikami illustre à juste titre ce statut fondamental du thème avec la possibilité de construction d'une phrase interrogative uniquement par un syntagme thématisé.

ここ　に　あった　新聞　は？
(*koko - ni - at ta - shinbun - wa*)
(ici - [lieu] - se trouver [passé] - journal - [*wa*])
« (Où est / qu'est-ce que tu as fait de) le journal qui était là ? »

Dans cet exemple, la fonction syntaxique du syntagme thématisé vis-à-vis du mot prédicatif est en suspens, n'étant fixée que dans la réponse à la question. Si on répond « *katazuke mashita* » (ranger - [passé, poli]), le thème est au cas accusatif vis-à-vis du verbe « ranger », et si la réponse est « *tsukue no ue ni arimasu* » (bureau - de - dessus - [lieu] - se trouver [poli]) le thème est au cas nominatif vis-à-vis du verbe « se trouver ».

Aussi, Mikami considère-t-il la fonction syntaxique du syntagme thématique vis-à-vis du mot prédicatif comme secondaire, soulignant l'importance de sa première fonction : être le thème de la phrase.

Il dit dans Mikami (1963b) :

> « [...] le locuteur ou l'auteur, lorsqu'il prononce "*Hamamatsu wa*[11]", ne pense pas forcément à une telle distinction[12]. D'ailleurs, c'est cette capacité à être employé sans nécessiter ce genre de prévision contraignante qui est la particularité du syntagme "X *wa*". »

6.4.3 Double fonction du syntagme thématisé

Fonction principale et fonction cumulative

Comme nous l'avons vu dans la section précédente, le syntagme thématisé joue, en général, en plus de la fonction de thème, une fonction syntaxique vis-à-vis du mot prédicatif. Kitahara (1976) explique la fonction du syntagme thématisé comme suit :

> « Le thème "X *wa*" entraîne l'achèvement de la construction d'une phrase par interaction mutuelle avec la fin de phrase située dans le prédicat, tout en conservant (ou "représentant" selon Mikami) la fonction grammaticale "X *ga*", "X *wo*", "X *ni*" ou "X *de*" qu'il occuperait dans la phrase en *koto*[13]. Étant donné que le thème entraîne par interaction mutuelle avec le prédicat l'accomplissement de la

[11] Hamamatsu (nom de ville) - particule *wa*.

[12] La distinction entre les différents cas que le syntagme thématique va assurer vis-à-vis du prédicat.

[13] Afin de connaître la particule marquant la fonction du syntagme, omise du fait de l'utilisation de la particule *wa*, on transforme la phrase en une structure enchâssée, comme avec コト (*koto*), structure dans laquelle la particule *wa* ne peut pas apparaître. Mikami a introduit ce terme *koto* comme la traduction japonaise de « dictum » de Bally (1965), en même temps que le terme モード (*môdo*, modalité), traduction de « modus » (cf. § 5.5.1).

construction d'une phrase, il correspond au sujet dans les langues comme l'anglais.

Dans la phrase :

この本　は、父　が　買ってくれました。
(*kono hon - wa, - chichi - ga - katte kure mashi ta*)
(ce livre - [*wa*] - mon père - [*ga*] - m'acheter [passé])
« Ce livre, mon père me l'a acheté »

on introduit une information nouvelle "*chichi ga katte kure mashi ta* (mon père me l'a acheté)" à propos du thème (information connue) "*kono hon wa* (ce livre)". »

Telle est la double fonction du syntagme thématisé. Mikami appelle la fonction de thème **première fonction** ou **fonction principale**, et la fonction grammaticale qu'il occuperait vis-à-vis du mot prédicatif (ou plus précisément du radical du mot prédicatif), **fonction cumulative**.

Fonction cumulative : fonction de complément secondaire

Le syntagme thématisé assure aussi bien la fonction de cas nominatif que d'autres fonctions essentielles. Mais, il est également possible qu'il assume d'autres fonctions[14], en particulier la fonction secondaire d'un complément du mot prédicatif.

田中さん　は　家族　が　フランス　に　住んでいる。
(*tanaka san - wa - kazoku - ga - furansu - ni - sunde iru*)
(M. Tanaka - [*wa*] - famille - [nominatif] - France - [lieu] - habiter)
« M. Tanaka, sa famille habite en France »

Dans cet exemple, le syntagme thématisé assure le cas de *no* (qualification d'un substantif) vis-à-vis du mot « famille » et constitue avec lui le syntagme assurant le cas nominatif vis-à-vis du prédicat.

Chevauchement entre la fonction cumulative et la fonction de complément

On trouve également des phrases dans lesquelles la fonction du syntagme suivi de *wa* coïncide avec la fonction assurée par un autre syntagme.

新聞　は、　朝日　を　読んで　います。
(*shinbun - wa - asahi - wo - yonde - imasu*)
(journal - [*wa*] - Asahi - [accusatif] - lire - [habitude])
« Comme quotidien, je lis le Asahi »

[14]Pour des études détaillées sur les différentes fonctions cumulatives du thème, nous pouvons citer Mikami (1960), Noda (1998) et Kikuchi (1995).

Détermination difficile de la fonction cumulative

Cependant, il existe beaucoup de phrases dans lesquelles la fonction grammaticale du syntagme suivi de *wa* vis-à-vis du prédicat, est difficile à définir (exemples tirés de Mikami (1960)).

（犯行　の）場所　は、屋内　説　が　圧倒的だった。
(*(hankô - no) basho - wa - okunai - setsu - ga - attôtekidat ta*)
((crime - de) lieu - [*wa*] - intérieur du bâtiment - thèse - [nominatif] - être majoritaire [passé])
« Quant au lieu (du crime), la thèse d'un déroulement à l'intérieur était majoritaire »

新聞　を　読みたい　人　は、ここ　に　あります　よ。
(*shinbun - wo - yomitai - hito - wa - koko - ni - arimasu - yo*)
(journal - [accusatif] - avoir envie de lire - personne - [*wa*] - ici - [lieu] - se trouver - [appel[15]])
« Pour les gens qui veulent lire le journal, (ils) sont posés là ! »

Ces exemples sont des cas où la fonction grammaticale du syntagme suivi de *wa* vis-à-vis du prédicat, est très difficile à déterminer. La reconstitution de la forme avant la thématisation semble nécessiter beaucoup d'éléments implicites.

Ces phrases – que Noda (1998) appelle phrases avec thème à cas cassé (破格主題文, *hakaku shudai bun*) – sont caractérisées par le fait que la relation syntaxique entre le syntagme thématisé et les éléments du prédicat n'est pas évidente. Entre les deux éléments que le locuteur met en relation – l'un introduit par *wa* et l'autre exprimé par l'ensemble des éléments associés par le prédicat –, on ne peut constater, comme le dit Mikami, que « l'existence d'une relation quelconque »[16]. La relation sémantique entre ces éléments n'est déductible que de connaissances extra-linguistiques, voire du contexte.

Les phrases *unagi*, qui font couler beaucoup d'encre dans la linguistique japonaise depuis Okutsu (1978), sont un type particulier de ces phrases avec thème :

僕　は　ウナギ　だ。
(*boku - wa - unagi - da*)
(moi - [thème] - anguille - [copule])

Cette phrase ne signifie absolument pas que le locuteur est une anguille mais la particule *wa* introduit le thème comme « quant à N » ou « en ce qui concerne N »

[15] Il s'agit d'un autre type de particule appelé *shûjoshi*, particule finale. Ces particules figurent au rang des éléments de la modalité. Elles sont utilisées principalement pour rappeler l'attention de l'interlocuteur, ou renforcer le ton.

[16] On constate également en français oral des mises en relation thème-rhème de deux éléments assez vagues par juxtaposition (exemples tirés de Blanche-Benveniste (2000)) :
- Le lendemain, grande surprise
- Ce soir, pas moyen
- En centre ville, d'accord
- mais ailleurs, non

en français. La phrase peut donc être traduite par « Moi, de l'anguille. » et on la considère comme la forme réduite de la phrase : « *boku wa unagi wo taberu* » (moi - [thème] - anguille - [accusatif] - manger / Je prends de l'anguille) ou « *boku ga tabetai no wa unagi da* » (moi - [nominatif] - vouloir manger - [pronom neutre] - [thème] - anguille - [copule] / Ce que je veux prendre, c'est de l'anguille).

Noda (1998) essaye d'expliquer tous les cas particuliers par trois types d'anomalies de surface : la réduction, la répétition ou le thème vague.

Mikami (1960) parle par ailleurs des thématisations de situation et de résultat.

この　におい　は　ガス　が　もれている　にちがいない。
(*kono - nioi - wa - gasu - ga - moreteiru - ni chigainai*)
(ce - odeur - [*wa*] - gaz - [nominatif] - fuire - il est sûr que)
« Cette odeur (vu cette odeur), il doit y avoir une fuite de gaz »

Il ne parle que des thématisations de résultat et de situation, mais on pourrait y ajouter celle de cause, et la phrase suivante semble en être un exemple :

ビール　は　太る。
(*bîru - wa - futoru*)
(bière - [thème] - (on) grossit)

Il est difficile de dire quelle est la construction explicite de cette phrase[17]. On peut tout de même imaginer une phrase initiale exprimant une cause « *bîru wa futoru gen'in da* » (bière - [wa] - (on) grossit - cause - [copule] / La bière est une cause de prise de poids).

Ces phrases particulières montrent cependant clairement la particularité de la particule *wa*. Cette particule est parfois comparée avec la copule[18], mais les relations entre deux éléments reliés par cette particule sont beaucoup plus larges que celles exprimées par la copule du français ou de l'anglais : tous les éléments reliés par le verbe « être » peuvent être mis en rapport par *wa*, mais tous les éléments mis en relation par *wa* ne sont pas reliables par le verbe « être ».

Plusieurs fonctions cumulatives pour un syntagme thématisé

Dans certains cas, le syntagme en *wa* a même plus d'une fonction cumulative : une vis-à-vis du mot prédicatif principal et une pour le mot prédicatif de chaque

[17] On constate le même problème dans la construction des relatives. La détermination de la relation sémantique entre les deux éléments devient un problème crucial lors par exemple d'une traduction. Le syntagme nominal créé sous forme de relative : *kodomo ga futoru tabemono* (enfants - [nominatif] - grossit - aliment), doit être traduit en français par « l'aliment ☐ les enfants grossissent » avec un relatif dans le cadre. Trouver un relatif adéquat en déduisant la relation sémantique implicite à l'aide de connaissances extra-linguistiques puis transformer en phrase naturelle : « l'aliment qui fait grossir les enfants » n'est pas toujours une opération évidente même pour un être humain.

[18] Hayami (1923) dit qu'il arrive en japonais que la particule *wa* fonctionne comme la copule. Mikami (1963b) soutient cette théorie en expliquant que le mot *dearu* – considéré généralement comme copule – est un élément optionnel qui peut tout à fait être remplacé par une particule finale.

proposition subordonnée ou coordonnée. (Les prédicats – sauf celui de la subordonnée déterminante – sont soulignés pour faciliter la compréhension.)

構造計算　は、計算者　の　裁量　に　任される　部分　が　多く、
すべて　法律　で　定められている　わけではありません。
(*kozô keisan - wa - keisan sha - no - sairyô - ni - makasareru - bubun - ga - ôku*
- subete - hôritsu - de - sadame rarete iru - wake de wa arimasen)
(calcul structural - [*wa*] - personne qui calcule - de - jugement - [*ni*] - confier [passif] - partie - [*ga*] - nombreux
- tout - loi - [moyen] - décider [passif] [état] - ce n'est pas que)
« Une grande partie du calcul structural est laissée à l'appréciation de la personne qui calcule et tout n'est pas défini par la loi » (corpus Yomiuri)

Le syntagme thématisé de la phrase d'exemple assure à la fois le cas de *no* vis-à-vis du mot « *bubun* (partie) » qui est au cas nominatif du premier mot prédicatif « *ôku* (nombreux) » et le cas de *ga* (nominatif) vis-à-vis du second mot prédicatif « *sadame rarete iru* (être défini) ».

Cas ambigus

皇室　の　養蚕　は　日本書紀　に　登場する　が、
１８７１年　に　昭憲皇太后　が　復興し、
貞明、香淳、今　の　皇后さま　と　受け継がれてきた。
(*kôshitsu - no - yôsan - wa - nihonshoki - ni - tôjô suru - ga*
- 1871 nen - ni - shôken kôtaigô - ga - fukkôshi
- teimei - kôjun - ima - no - kôgô sama - to - uketsu garete kita)
(famille impériale - de - sériciculture - [*wa*] - Chronique du Japon - [*ni*] - apparaître - [connexion]
- année 1871 - [*ni*] - Impératrice Shôken - [*ga*] - rétablir
- (Impératrice) Teimei - (Impératrice) Kôjun - actuel - de - Impératrice - [*to*] - prendre la suite de)
« La Chronique du Japon [NdT. datée de 720] parle déjà de la sériciculture chez la famille impériale, mais en 1871 l'impératrice Shôken l'a rétablie et les impératrices Teimei, Kôjun, puis l'actuelle impératrice en ont pris la suite jusqu'aujourd'hui » (corpus Yomiuri)

La fonction cumulative du syntagme thématisé assume vis-à-vis du premier prédicat « apparaître » le cas *ga* (nominatif) et vis-à-vis du deuxième prédicat « rétablir » le cas *wo* (accusatif). Mais sa fonction vis-à-vis du troisième « prendre la suite de » est ambigu. En effet, la forme « *uketsu garete* » peut être interprétée aussi bien comme passive que comme honorifique et dans le premier cas le syntagme thématisé assume le cas *ga* tel que « la sériciculture est transmise d'impératrice en impératrice, depuis trois générations », et dans le second le cas *wo*, « les trois impératrices héritaient la sériciculture »[19].

[19]Cette ambiguïté est également due à la particule *to* qui introduit les noms des trois impératrices. Cette particule – que nous analysons plutôt comme *to* de citation (ou très proche) que *to* de coordination – sert vraisemblablement à énumérer un ensemble de paradigmes d'un complément ou à introduire des exemples. N'appartenant pas aux particules de cas, elle n'indique pas la fonction, mais aussi en permettant d'omettre les particules de cas, elle rend complètement implicite la fonction du syntagme qu'elle introduit. Si bien qu'il est possible de desambiguïser la phrase en utilisant une particule de cas à la place de *to* : le remplacement de *to* par *ga* oblige l'interprétation de la

6.4.4 *Wa* non-thème

Nous n'avons parlé jusqu'ici que du thème introduit par la particule *wa*. Mais, il existe également des syntagmes en *wa* qui ne sont pas le thème de la phrase, et la thématisation en japonais n'est pas toujours réalisée avec la particule *wa*.

En effet, dans la grammaire de Masuoka-Takubo par exemple, il existe une particule *wa* qui est classée dans la catégorie des particules de mise en relief (cf. § 5.3.2), et la catégorie des particules de thématisation comporte également non seulement la particule *wa* mais aussi quelques autres particules.

Nous examinons dans cette section l'autre fonction de la particule *wa* (ou la fonction d'une autre particule *wa*), et dans la section suivante, d'autres moyens de thématisation.

Particule de *toritate*, mise en relief

La grammaire de Masuoka & Takubo (1992) définit les particules de *toritate*, comme celles qui servent à mettre en avant un élément parmi son paradigme. Si bien que quand on parle de X spécialement introduit par une particule de mise en relief, l'existence d'autres éléments Y, Z, ..., etc. est présupposée.

La particule *wa* de mise en relief permet de produire un effet contrastif.

鮫は　魚類で、鯨は　ほ乳類だ
(*same wa - gyorui de - kujira wa - honyûrui da*)
(requin [wa] - poisson [copule à la forme neutre] - baleine [wa] - mammifère [copule])
« Les requins sont des poissons et les baleines des mammifères. »

ワインは　飲む　が、ビールは　飲まない
(*wain wa - nomu - ga - bîru - wa - nomanai*)
(vin [wa] - boire - [opposition] - bière [wa] - boire [négation])
« (La personne en question) boit du vin mais ne boit pas de bière. »

Dans le premier exemple, les requins et les baleines, tous les deux introduits par la particule *wa*, sont comparés et dans le second, le vin et la bière[20].

Enfin, il est à noter qu'il existe un autre type de *wa* très proche, voire un sous-type, de la particule *wa* de contraste, qui est celle s'accompagnant de la négation comme :

豚肉は　食べない

forme du prédicat comme honorifique, et celui par *ni* impose l'interprétation comme forme passive. Dans le cas de l'utilisation de la particule *ni*, il est possible de conserver également la particule *to* après cette particule de cas. Il est également possible que la particule *to* précède la particule de cas *ni*, et dans ce cas, la particule *to* doit être interprétée comme celle de coordination. Mais la particule *to* de coordination semble difficilement fonctionner toute seule sans particule de cas, raison pour laquelle nous avons, dans l'analyse de l'exemple, retenu l'interprétation de *to* comme particule de citation.

[20] Les éléments mis en relief par la particule de *toritate* ne se limitent pas aux syntagmes directement introduits par ces particules, qui mettent même parfois en relief l'ensemble du prédicat en introduisant seulement un des compléments. Pour les questions sur la portée des particules de mise en relief, voir par exemple Numata & Jo (1995) ; Numata (2000).

(*butaniku wa - tabenai*)
(porc [wa] - manger [négation])
« (La personne en question) ne mange pas de porc. »

Difficulté de distinction

Toutefois, la distinction entre la particule *wa* de thématisation et celle de mise en relief n'est pas toujours évidente.

En effet, dans le premier exemple, la particule *wa* produit non seulement l'effet de contraste mais aussi celui de thématisation. Si bien que certains travaux comme Masuoka (1991) ou Noda (1998) définissent entre les deux types de *wa*, ceux de thème et de contraste, un *wa* à bi-usage, en reconnaissant la continuité de ces deux emplois.

D'autres comme Teramura (1991) renoncent à la distinction de *wa*. Ce dernier considère que la fonction principale de cette particule est la production de l'effet de contraste et que dans certaines conditions, elle provoque la thématisation sans produire aucun effet de contraste.

Mais, il semble difficile d'éviter leur distinction, car il existe une contrainte syntaxique différente qui fait que seul *wa* de contraste peut apparaître dans une proposition subordonnée déterminante.

Mikami (1953) distingue également les deux emplois de la particule *wa* et il les appelle le thème et le thème secondaire. D'après lui, la portée du thème secondaire est plus restreinte que le thème qui peut avoir une influence sur plusieurs phrases. Mais, comme dit Mikami (1953), « on a pu confirmer que la distinction de ces deux types est nécessaire, mais on ne peut pas déterminer les conditions qui les différencient. » Il signale toutefois que les particules *wa* apparaissant après la première occurrence de *wa* dans la phrase servent souvent non pas à la thématisation, mais plutôt à l'effet de contraste[21]. Mais le thème pouvant être omis selon le contexte, le *wa* de contraste peut tout à fait apparaître à la première position, ce qui empêche la détermination de la nature selon la position[22].

Thème terminé par une particule de cas + *wa*

La particule *wa* peut suivre un syntagme terminé par une particule de cas (PC ci-dessous). Mikami signale dans ses ouvrages tels que Mikami (1960, 1963b) que les syntagmes « N + PC + *wa* » apparaissent en général à une position postérieure à la première occurrence de *wa* et que leur portée est donc également plus restreinte que celle du syntagme « N + *wa* »[23].

[21] La même remarque se trouve dans Kuno (1973).

[22] Mikami (1963b) dit également dans qu'à l'oral, le syntagme en *wa* accentué est contrastif et celui non accentué thématique.

[23] Il est à noter l'existence des travaux de Klingler (2003) sur les syntagmes « N + *de/ni* (PC locative) + *wa* » que l'auteur interprète comme des circonstants cadratifs (voir aussi § 6.5.9).

6.4.5 Thème non-*wa*

Autres particules de thématisation

La grammaire de Masuoka-Takubo énumère comme particules de thématisation, は (*wa*), なら (*nara*), って (*tte*), ったら (*ttara*).

Cependant, ces mots (excepté *wa*) ne sont pas toujours considérés comme des particules. En revanche, l'effet de mise en thème qu'ils produisent est généralement reconnu.

Mikami (1963b), tout en affirmant que la particule *wa* est la seule particule de thématisation, énumère comme constructions qui permettent également de réaliser la thématisation :

1. les constructions de l'expression hypothétique :
   なら (*nara*), etc. ;
2. les expressions contenant la construction de citation « *to iu* » :
   って (*tte*), ったら (*ttara*), ってば (*tteba*), と（いうの）は (*to (iuno) wa*) ;
3. certaines expressions contenant une construction équivalente aux particules de cas :
   のために (*no tame ni*), とともに (*to tomo ni*), をもって (*wo motte*), について (*ni tsuite*), にとって (*ni totte*), において (*ni oite*), に対して (*ni taishite*).

La liste étant incomplète, la construction d'un lexique exhaustif serait, sans aucun doute, indispensable pour l'analyse syntaxique du japonais.

Particule *mo*

La particule *mo* est également une particule de *toritate*, mais elle est parfois considérée comme une particule de thématisation.

À côté de la particule de *toritate wa* qui met en contraste deux éléments du même paradigme, elle introduit un autre élément du même paradigme, semblable au premier élément déjà présenté[24] comme dans l'exemple suivant :

父は　日本人で、母も　日本人です。
(*chichi - wa - nihonjin de - haha mo - nihonjin desu*)
(mon père [wa] - japonais [copule] - ma mère [mo] - japonais [copule, politesse])
« Mon père est japonais et ma mère aussi est japonaise ».

Comme l'explique Mikami (1963b), la fonction de *mo* est généralement équivalente au précédent élément appartenant au même paradigme que l'élément introduit par *mo* et avec lequel ce dernier est comparé. Dans l'exemple, les deux éléments comparables « mon père » et « ma mère » sont introduits l'un après l'autre. « mon père » est d'abord introduit par *wa*, puis « ma mère » par *mo*. « ma mère + *mo* » est donc le thème comme l'est « mon père + *wa* ».

[24] La particule *mo* a d'autres fonctions que l'introduction d'éléments semblables. Pour des études précises, voir Numata (1986).

Comme nous venons de le voir, par cette propriété, la particule *mo* peut également indiquer le thème, mais elle peut aussi introduire un élément non-thème du fait de cette même propriété.

私は　ワインが　好きです　が、ビールも　飲みます。
(*watashi wa - wain ga - sukidesu - ga - biiru mo - nomimasu*)
(moi [wa] - vin [ga] - apprécié [politesse] - [connexion] - bière [mo] - boire [politesse])
« J'aime le vin, mais je bois aussi de la bière »

mo introduit « la bière » en comparaison avec « le vin ». « le vin + *ga* » étant au cas *ga* non-thème, « la bière + *mo* » n'est pas non plus le thème.

Mikami estime que la particule *mo* jouant le rôle d'indicateur de thème se limite à moins d'un tiers de ses emplois. Mais, les deux éléments comparés peuvent apparaître dans deux phrases différentes et la possibilité d'omission des éléments complique également la détermination automatique de sa nature. Nous ne considérons donc pas, dans le cadre de la présente thèse, la particule *mo* comme une particule de thématisation. Nous examinerons la conséquence de ce choix lors de l'évaluation de nos analyses automatiques.

6.4.6 Notre position pour l'analyse syntaxique des syntagmes en *wa*

Étant donné la nature de nos travaux, il est difficile de réaliser une distinction quand il n'existe aucun repère formel entre les éléments à distinguer. La distinction entre les syntagmes en *wa* de thème et en *wa* de contraste est donc une tâche très délicate.

Nous définissons tout de même deux types de syntagmes en *wa*, thème fort et thème faible, afin de pouvoir distinguer ces deux constituants nous paraissant assez différents, à l'aide de tout indice formel exploitable.

Nous posons comme hypothèses les règles suivantes :

1. Le syntagme introduit par la particule *wa* utilisée seule diffère de celui introduit par la particule *wa* suivant une/des particule(s) de cas. Le premier est défini comme thème fort et le second comme thème faible.
2. Le thème fort entre en relation avec le dernier prédicat, c'est-à-dire le prédicat principal de la phrase, excepté le cas suivant.
3. Quand un autre thème fort apparaît dans la phrase et qu'il existe un/des prédicat(s) entre le premier et ce second thème fort, le premier thème fort entre en relation avec le(s) prédicat(s) situé(s) antérieurement au second thème fort et perd sa fonction de thème sans étendre sa portée jusqu'à la fin de phrase.
4. Si un ou plusieurs thèmes forts apparaissent après le premier thème fort de la phrase sans aucun prédicat entre eux, ces premiers doivent être considérés comme des thèmes faibles.

5. Les thèmes faibles n'entrent en relation qu'avec le premier prédicat rencontré.

Nous réaliserons des analyses de phrases japonaises avec les règles ci-dessus, et examinerons leurs avantages et inconvénients à partir des résultats.

6.5 Éléments préposés par rapport au thème

Comme nous l'avons déjà répété un certain nombre de fois, nos travaux se basent particulièrement sur la thèse que la phrase japonaise est de structure thème-rhème. Nous nous appuyons également sur une autre hypothèse qui est, semble-t-il, peu adoptée par les linguistes japonais : certaines places dans la phrase sont liées à des fonctions syntaxiques particulières.

Nous pensons, en particulier, que la partie initiale de la phrase – antérieure au thème – est une place réservée aux éléments qui ne participent pas à la constitution de la structure d'opposition thème-proposition. Afin de défendre notre hypothèse, nous examinons maintenant les éléments apparaissant à la place initiale de la phrase japonaise, afin de tenter d'élucider le rapport entre cette place et la fonction externe au noyau syntaxique.

Nous nous intéressons d'abord aux moyens d'indication de la fonction pour les éléments considérés externes par les grammaires usuelles, avant de présenter nos études de corpus sur les éléments apparaissant à la place initiale.

6.5.1 Moyens d'indication de la fonction externe

Comme déjà défini dans la section 5.5.4, nous appelons éléments externes les éléments qui n'appartiennent pas au noyau syntaxique de la phrase japonaise, composé du thème et de la proposition constituée autour du mot prédicatif. Les grammaires japonaises reconnaissent l'existence dans la phrase japonaise d'éléments extérieurs à la structure syntaxique centrale. Ce sont des éléments de phrase que les grammaires scolaires appellent 独立語 (*dokuritsu go*, mots indépendants) ou des éléments de liaison (接続語, *setsuzoku go*) qui établissent le lien avec les phrases précédentes (cf. § 5.5.4). Beaucoup de linguistes ont également reconnu l'extériorité de certains adverbes, appelés aujourd'hui adverbes de phrase, ainsi qu'une classe plus large d'éléments dits éléments d'évaluation, *hyôka-seibun* (cf. § 5.5.4).

Comme nous l'avons vu, l'extériorité des éléments extra-prédicatifs en français est marquée par détachement et/ou positionnement en tête de phrase. Quels sont alors les moyens d'indication de la fonction externe pour ces éléments externes de la phrase japonaise ?

Nous constatons trois types d'indication de la fonction externe : catégorie, moyen morpho-lexical et place.

Catégorie = Fonction

L'intériorité ou extériorité d'un élément semble être considérée comme fixe selon sa nature lexicale, c'est-à-dire sa catégorie. C'est le cas des mots d'émotion (感動詞, *kandô-shi*) qui constituent la plupart des éléments indépendants.

Le statut des adverbes de phrase est beaucoup moins stable, comme nous allons le voir, et la définition de la classe des adverbes selon ce critère semble être beaucoup plus difficile.

Moyen morpho-lexical

Il existe également un autre type de moyen d'indication, morpho-lexical. Kudo (1997) montre qu'un même adverbe peut être complément du prédicat ou complément de phrase et il analyse cette différence de fonction comme étant réalisée par l'ajout de la particule *mo* qui suit l'adverbe lorsque celui-ci joue la fonction du complément de phrase :

– 太郎は、地図を　書いて、道順を　親切に　教えてくれた。
(*tarô wa - chizu wo - kaite - michijun wo - shinsetsuni - oshiete kureta*)
(Tarô [wa] - plan [wo] - écrire - chemin [wo] - gentiment - m'apprendre)
« Tarô a dessiné un plan et m'a gentiment indiqué le chemin. »

– 太郎は、親切にも、地図を　書いて道順を　教えてくれた。
(*tarô wa - shinsetsuni mo - chizu wo - kaite - michijun wo - oshiete kureta*)
(Tarô [wa] - gentiment [mo] - plan [wo] - écrire - chemin [wo] - m'apprendre)
« Gentil, Tarô a dessiné un plan et m'a indiqué le chemin. »

Dans la première phrase, l'adverbe « gentiment » mis juste devant le prédicat est un complément du prédicat et ce qui est gentil est la manière dont Tarô a indiqué le chemin. En revanche, dans le second exemple, le même adverbe mis dans une position antérieure est un complément de phrase et sa portée est l'ensemble de la proposition, ce qui est qualifié de gentil étant l'ensemble des services qu'a rendu Tarô au locuteur, ou encore Tarô lui-même.

Kudo analyse cette différence de fonction comme étant réalisée par l'ajout de la particule *mo* qui suit l'adverbe lorsque celui-ci joue la fonction de complément de phrase. Il définit ainsi trois types d'adverbes (ou de qualificatifs à la forme adverbiale) d'évaluation :

1. mots exprimant toujours l'évaluation **sans *mo***;
2. mots toujours utilisés **avec *mo*** exprimant presque toujours l'évaluation ;
3. mots à double emploi, jouant le complément du prédicat **sans *mo*** et le complément de phrase **avec *mo***.

Place et/ou détachement en position initiale

Un autre moyen d'indication que nous avons pu observer est la place et le détachement. Les substantifs et les mots de liaison jouent une fonction externe, en position initiale.

Les mots indépendants et les mots de liaison sont généralement placés en tête de phrase, mais leur extériorité est déjà fortement marquée au niveau lexical de leurs constituants. La plupart des mots indépendants sont, comme nous l'avons déjà signalé, réalisés par des mots de la catégorie particulière 感動詞 (*kandôshi*, mots d'émotion), qui sont, quelle que soit leur place, extérieurs au noyau syntaxique. Toutefois, les mots indépendants appelés *teijigo* (mot présentant une chose ou un fait) étant réalisés par un substantif, leur fonction est marquée non pas par les constituants eux-mêmes – lexicalement –, mais par des moyens syntaxiques : position initiale et détachement :

卒業写真、それ　は　私　の　好きな　曲　です。
(*sotsugyô shashin - sore - wa - watashi - no - sukina - kyoku - desu*)
(photo de fin de l'école - cela - [wa] - moi - [no] - favori - chanson - [copule])
« "Sotsugyô shashin", telle est ma chanson préférée »

Les mots de liaison sont également réalisés par des mots de la catégorie particulière 接続詞 (*setsuzokushi*, mots de liaison). Mais, contrairement aux mots d'émotion, les mots de liaison peuvent assurer aussi les fonctions internes au noyau syntaxique :

文学、歴史　**そして**　教育　に　興味がある
(*bungaku - rekishi - soshite - kyôiku ni - kyômi ga aru*)
(littérature - histoire - et - éducation [ni] - avoir intérêt)
« (Je/Il/Elle... m'/s'/...) intéresse(ent) à la littérature, à l'histoire, ainsi qu'à l'éducation »

Leur fonction peut donc être marquée par des moyens syntaxiques : position initiale et/ou détachement :

そして、夏は　終わった。(*soshite - natsu wa - owatta*)
(puis - été [wa] - finir [passé])
« Puis/Ainsi, l'été est fini »

La place peut-elle marquer la fonction ?

Ces constats sur les moyens d'indication de la fonction externe nous ont ramené aux interrogations suivantes : est-il possible de généraliser le rapport entre la place initiale et la fonction externe ? Y a-t-il d'autres types d'éléments qui apparaissent en tête de phrase et qui jouent des fonctions externes ?

Dans l'article de Kudo cité précédemment, on peut trouver un exemple qui donne une réponse à notre première question.

– 親切に（も）　かれは　道を　ていねいに　教えてくれた。
(*shinsetsuni (mo) - kare wa - michi wo - teinei ni - oshiete kureta*)
(gentiment ([mo]) - lui [wa] - chemin [wo] - avec soin - m'apprendre)
« Gentil, il m'a indiqué le chemin avec soin. »

Dans cet exemple, l'adverbe « gentiment » joue la fonction de complément de phrase. L'adverbe est mis en tête et la particule *mo*, entre parenthèses, y est considérée comme facultative. L'extériorité de l'adverbe est donc ici marquée notamment par la mise en position initiale.

Dans l'article de Kudo, d'ailleurs, figure un passage laissant entendre l'existence d'une relation non négligeable entre la place et la portée, quoique l'auteur semble éviter soigneusement toute expression ferme :

> « On pourrait dire que plus [les mots d'évaluation] se situent à une place antérieure, c'est-à-dire plus ils sont loin du verbe, plus ils perdent leur caractère subordonné et leur valeur de restriction alors que leur caractère indépendant et leur valeur d'évaluation augmentent. L'ordre (place dans la phrase) pourrait en effet jouer un rôle dans l'attribution d'un caractère indépendant aux éléments de phrase, rôle que peut jouer la particule *mo*. »

Nous avons alors posé comme hypothèse que la partie initiale de la phrase – antérieure au thème syntaxique explicite – était une place réservée aux éléments qui ne participent pas à la constitution du noyau syntaxique composé du thème et de la proposition. Nous considérons que la partie initiale de la phrase représente, tout comme dans la phrase française, une zone de liberté pour l'énonciateur. Mais, comme la phrase japonaise ne dispose pas d'élément susceptible de marquer la frontière de la structure syntaxique centrale (cf. sujet ou terme en qu- en français), nous avons posé une autre hypothèse : le syntagme en *wa* trace une certaine limite dans la phrase. Nous avons ainsi décidé d'étudier les éléments préposés par rapport au syntagme thématisé par la particule *wa*.

Beaucoup de travaux qui parlent de l'ordre des mots dans la phrase japonaise existent dans la littérature. Cependant, les études se limitent souvent à celles sur la place la plus fréquente de chaque catégorie de mot, sans parler de la différence entre deux cas où le même mot apparaît à différentes positions dans la même phrase[25]. On trouve comme étude poussée sur l'ordre des mots celle de Saeki (1998) qui présente la tendance générale de l'ordre des mots comme :

> Interjection → mot de liaison → mot de situation, temps et lieu → thème → mot d'évaluation → compléments essentiels et accessoires → prédicat

L'auteur aborde également les conditions d'inversion de cet ordre « ordinaire » ou « canonique », mais il travaille surtout sur des phénomènes précis liés à certains couples de catégories de mots, et ne cherche pas, semble-t-il, à associer à une certaine place une fonction particulière.

Nous avons donc décidé d'observer les réelles occurrences des éléments apparaissant avant le thème afin de déterminer quels types d'éléments étaient placés

[25] Il est cependant à noter que des études intéressantes ont été réalisées dans le cadre de la grammaire fonctionnelle et elles proposent l'analyse du changement des natures informatives des constituants selon leur position (Kuno, 1978 ; Takami, 1995).

en position initiale pour jouer quelle fonction, et si cette place était vraiment réservée aux éléments extérieurs.

6.5.2 Études sur corpus : méthodologie et données

Nous avons collecté environ 150 exemples d'éléments (non phrastiques) précédant le thème syntaxique explicite. Nous n'avons extrait que des éléments non phrastiques – c'est-à-dire sans mot variable, sauf ceux avec un mot variable dont l'emploi est plus ou moins figé (cf. Xについて, *x ni tsuite*, « concernant X ») – afin d'éviter l'intervention de divers facteurs, notamment la longueur, considérée souvent comme à l'origine de certains changements d'ordre.

Les exemples sont extraits des corpus Yomiuri, Fujiwara, Murakami-kaze et Tsutsui[26].

6.5.3 Éléments pré-thèmes extraits du corpus

Différents éléments sont constatés en position pré-thème dans les exemples étudiés. Nous avons trouvé bien entendu des éléments qui sont considérés comme externes par les grammaires usuelles : éléments indépendants, éléments de liaison, adverbes de phrase et éléments d'évaluation.

En outre, nous avons également rencontré des compléments de temps qui ouvrent les cadres temporels, des compléments de lieu qui ouvrent les cadres spatiaux, des introducteurs d'autres cadres comme les cadres d'énonciation et le cadre thématique. Enfin, les éléments avec des particules de cas sont aussi présents à cette position.

Nous allons maintenant examiner de plus près quelques exemples de ces éléments observés.

6.5.4 Éléments indépendants

Nous avons recueilli un seul exemple contenant des éléments indépendants.

川とテニス･コート、ゴルフ・コース、ずらりと並んだ広い屋敷、
壁そして壁、幾つかの小綺麗なレストラン、ブティック、
古い図書館、月見草の繁った野原、猿の檻のある公園、
街は　いつも同じだった。 (MURAKAMI)
(*kawa to tenisukôto, gorufukôsu, zurarito naranda hiroi yashiki,*
kabe sosite kabe, ikutsuka no kogirei na resutoran, butikku,
hurui toshokan, tsukimisô no shigetta nohara, saru no ori no aru kôen
machi wa - itsumo onaji datta)
(rivière et court de tennis, terrain de golf, beaucoup de grandes résidences s'alignant
côte à côte, murs et encore murs, quelques restaurants assez chics, boutiques, vieille
bibliothèque, prairie remplie d'onagres, parc où il y a une cage de singes -
ville [wa] - être toujours le même [passé])

[26] Pour le contenu détaillé de chaque corpus, voir la liste des corpus utilisés (page 550).

> « Rivière et court de tennis, terrain de golf, beaucoup de grandes résidences s'alignant côte à côte, murs et encore murs, quelques restaurants assez chics, boutiques, vieille bibliothèque, prairie remplie d'onagres, parc avec une cage à singes. La ville restait toujours la même. »

Ce sont des substantifs référant aux choses qui sont la preuve que la ville n'a pas changé. Ils ne jouent aucune fonction syntaxique vis-à-vis du mot prédicatif.

6.5.5 Éléments de liaison

Beaucoup de phrases comportent des éléments de liaison. Ces éléments établissant le lien avec le contexte antérieur ne se limitent pas aux mots *setsuzokushi*. Différentes constructions semblent être employées pour établir une bonne articulation du texte.

Mots de liaison, *setsuzoku-shi*

Dans nos exemples, les mots de liaison apparaissant en tête de phrase servent tous, comme nous l'avons imaginé, à lier un ensemble de phrases aux phrases précédentes.

> あるいは、一方、さらに、しかし、そして、それでも、ただ、ところが、また
> *(aruiwa - ippô - sarani - shikashi - soshite - soredemo - tada - tokoroga - mata)*
> (ou - par ailleurs - de plus - cependant - puis - toutefois - seulement - mais - et)

Constructions avec *kyûchakugo*

Nous avons constaté des constructions avec les mots dont la catégorisation varie assez largement selon les linguistes et que Sakuma (1940b) regroupe sous le nom de 吸着語 (*kyûchaku go*, mots agglutinants).

> 結局のところ、このため、そんなわけで
> *(kekkyoku no tokoro - kono tqme - son'na wake de)*
> (en fin de compte - pour cela - du fait de cela)

Autres constructions

> それにもかかわらず、これに対し、これに伴い
> *(sorenimo kakawarazu - koreni taishi - koreni tomonai)*
> (malgré cela - par rapport à cela - parallèlement à cela)

La nature de ces constructions destinée à établir le lien avec la/les phrases précédentes peut être constatée, comme pour les constructions avec *kyûchakugo*, par la présence d'éléments anaphoriques.

6.5.6 Adverbes de phrase

Les adverbes apparaissant en tête de phrase sont ceux classés dans la catégorie des adverbes de phrase :

むろん、もちろん、やはり
(muron - mochiron - yahari)
(bien entendu - bien entendu - comme on peut s'en douter)

Nous avons également constaté des constructions adverbiales plus ou moins figées comme 予定通り (*yotei dôri*, comme prévu). Ces adverbes sont classés dans la catégorie des adverbes de phrase. Mais, comme nous l'avons déjà signalé, l'extériorité des adverbes de phrase est beaucoup moins stable qu'on l'estime généralement et la définition de la classe des adverbes selon ce critère semble être assez difficile. Néanmoins, afin de déterminer leur vraie nature, nos exemples sont encore insuffisants et surtout l'étude des travaux existants sur les adverbes japonais, sans doute assez nombreux, serait indispensable.

6.5.7 Éléments d'évaluation

Il y avait dans nos exemples un élément d'évaluation préposé par rapport au thème.

さいわいなことに　**野球部員たちは**、そこにあらわれたのが　たまたま　七瀬という、教務課職員とはいえ　私立手部高校随一の美人であったが為に、異変への関心をすぐ失った。
(TSUTSUI)
(*saiwaina koto ni - yakyû buin tachi wa - sokoniarawareta no ga - tamatama - nanase to iu - kyômuka shokuin towaie - shiritsu tebe kôkô zuiitsu ni bijin de atta ga tameni - ihen eno kanshin wo sugu ushinatta*)
(par chance - équipes du club de baseball [wa] - personne apparaissant là [ga] - par hasard - appelé Nanase - bien qu'une employée administrative - être la plus belle femme du lycée privé Tebe [cause] - perdre l'intérêt pour l'événement extraordinaire [passé])
« Par chance, du fait que la personne apparue était par hasard Nanase, la plus belle fille du lycée privé de Tebe, bien qu'employée administrative, les équipes du club de baseball perdirent tout de suite tout intérêt pour cet événement extraordinaire. »

Comme nous l'avons déjà vu, les éléments d'évaluation sont des éléments externes par définition. Mais si l'on déplace l'élément d'évaluation de la phrase d'exemple après le thème comme :

野球部員たちは、さいわいなことに　そこにあらわれたのが　たまたま　七瀬という、教務課職員とはいえ　私立手部高校随一の美人であったが為に、異変への関心をすぐ失った。
(TSUTSUI)
(*yakyû buin tachi wa - saiwaina koto ni - sokoniarawareta no ga - tamatama - nanase*

to iu - kyômuka shokuin towaie - shiritsu tebe kôkô zuiitsu ni bijin de atta ga tameni - ihen eno kanshin wo sugu ushinatta)
(équipes du club de baseball [wa] - par chance - personne apparaissant là [ga] - par hasard - appelé Nanase - bien qu'une employée administrative - être la plus belle femme du lycée privé Tebe [cause] - perdre l'intérêt pour l'événement extraordinaire [passé])
« Du fait que, par chance, la personne apparue était par hasard Nanase, la plus belle fille du lycée privé de Tebe, bien qu'employée administrative, les équipes du club de baseball perdirent tout de suite tout intérêt pour cet événement extraordinaire. »

la portée de cet élément se limite à la première proposition de cause et ce qui est chanceux, c'est le fait que « la personne étant apparue était Nanase ».

La mise en tête de l'élément favorise l'interprétation par la portée large, c'est-à-dire la portée sur l'ensemble de la phrase : ce qui est chanceux, c'est le fait que « les équipes ont perdu tout intérêt pour l'événement ». Cette interprétation semble plus adéquate d'après l'histoire où Nanase, dotée de pouvoirs psychiques, tente de cacher la nature extrasensorielle de l'événement.

6.5.8 Compléments temporels

En japonais, il existe deux types de compléments de temps : adverbe (ou substantif à emploi adverbial) sans particule et substantif introduit par la particule de cas *ni*. La différence entre ces deux types est étudiée dans de nombreux travaux. Saeki (1998) consacre un chapitre à ce sujet et notamment à leur ordre d'apparition.

Dans ses travaux, il constate que les compléments nus (i.e. sans particule) ont une portée plus large que ceux introduits par la particule de cas *ni*, et que les premiers précèdent les seconds. De plus, il signale que les compléments nus peuvent avoir une portée plus large qu'une phrase, contrairement aux compléments avec la particule de cas *ni* pour lesquels il n'a trouvé aucun exemple dans ses corpus.

Dans les présents travaux, nous nous intéressons non pas à la différence entre les constructions avec et sans moyen morpho-lexical – à savoir la particule de cas *ni* –, mais à l'influence de la place.

Comme d'autres éléments, l'abord sur la place des compléments de temps dans la littérature se limite souvent à celui de l'ordre canonique. Par exemple, Minami (1993) définit la place des compléments de temps et de lieu dans une phrase à prédicat verbal, entre le thème et les compléments essentiels du verbe précédant les compléments *ren'yô-shûshokugo*.

Mais nous défendons la relation étroite entre la place et la portée. Nous considérons notamment que tous les compléments de temps et de lieu apparaissant à une place antérieure au thème jouent le rôle d'introducteur du cadre de discours (Charolles, 1997).

Nous examinons les compléments de temps sans particule (SN nus ou constructions avec *kyûchakugo* sans particule) apparaissant en position initiale,

mais aussi à une position postérieure. Les exemples étudiés semblent valider cette hypothèse.

> 放課後、**七瀬は**　職員室へ行く用があり、運よく国語の教師の浜口が自分の席にいるのを見かけた。 (TSUTSUI)
> (*hôkago* - ***nanase wa*** *- shokuin-shitsu e iku yô ga ari - un'yoku kokugo no kyôshi no hamaguchi ga jibu no seki ni iruno wo mikaeta*)
> (après la classe - Nanase [wa] - avoir à faire dans la salle des professeurs - trouver par chance le professeur de japonais Hamaguchi assis à son bureau)
> « Après la classe, en passant dans la salle des professeurs pour une tâche, Nanase a trouvé par chance le professeur de japonais Hamaguchi assis à son bureau. »

Avant cette phrase, est décrite une scène dans un bureau administratif du lycée lors de la pause déjeuner où Nanase, jeune employée administrative très belle, est harcelée par des propos ironiques d'un professeur, jaloux depuis qu'il a vu cette dernière dans un café avec un étudiant. La scène change d'un coup par l'installation d'un nouveau cadre ouvert par l'introducteur du cadre temporel, groupe nominal détaché en tête « *hôkago* » de la phrase d'exemple. Après cette phrase, les descriptions de la nouvelle scène se poursuivent avec cet arrière-plan, après la classe.

Considérons un autre exemple.

> 小さい頃、**僕は**　ひどく無口な少年だった。
> (*chiisai koro* - ***boku wa*** *- hidoku mukuchina shônen datta*)
> (quand (j'étais) petit - moi [wa] - être garçon parlant extrêmement peu [passé])
> « Quand j'étais petit, j'étais un garçon qui parlait extrêmement peu. »

...

> 14歳になった春、信じられないことだが、まるで堰を切ったように　**僕は**　突然しゃべり始めた。 (MURAKAMI)
> (*jûyon sai ni natta haru - shinjirarenai koto daga - marude seki wo kitta yôni -* ***boku wa*** *- totsuzen shaberi hajimeta*)
> (le printemps où j'ai eu 14 ans - c'est incroyable, mais - comme si on ouvrait le barrage - moi [wa] - commencer à parler tout à coup [passé])
> « Le printemps où j'ai eu 14 ans, c'était incroyable, mais je commençai, comme si on ouvrait un barrage, à parler tout à coup. »

Après la première phrase, située au début d'un chapitre du roman, se poursuivent les descriptions de l'époque où les parents s'inquiétaient au point de l'envoyer chez un psychiatre une fois par semaine afin de suivre des séances. Après la description assez longue de ces séances, apparaît la seconde phrase qui change complètement l'arrière-plan par l'installation du nouveau cadre avec le SN détaché en tête, *jûyon sai ni natta haru*.

En revanche, lorsque des SN fonctionnant comme les compléments de temps apparaissent après le thème, l'ouverture du nouveau cadre peut se révéler beaucoup moins nette.

僕は　以前、人間の存在理由をテーマにした短かい小説を書こうとしたことがある。
(***boku wa*** *- izen - ningen no sonzai riyû wo têma ni shita mijikai shôsetsu wo kakô to shita koto ga aru*)
(moi [wa] - autrefois - tenter d'écrire un petit roman ayant comme thème la raison d'être des hommes [passé])
« J'ai tenté d'écrire autrefois un petit roman ayant comme thème la raison d'être des hommes. »
結局　小説は完成しなかったのだけれど、その間じゅう　僕は　人間のレーゾン・デートゥルについて考え続け、おかげで奇妙な性癖にとりつかれることになった。 (MURAKAMI)
(*kekkyoku - shôsetsu wa kansei shinakatta keredo - sono aida jû - bokuwa rezon-dêtoru ni tsuite kangae tsuzuke - okagede kimyôna - seiheki ni toritsukareru koto ni natta*)
(finalement - le roman n'a pas vu le jour, mais - pendant ce temps-là - moi [wa] - penser continuellement à la raison d'être des hommes - ainsi finir par développer une habitude bizarre [passé])
« Finalement, le roman n'a pas vu le jour, mais pendant cette période je pensais continuellement à la raison d'être des hommes et j'ai ainsi fini par développer une habitude bizarre. »

Le SN de temps de la première phrase, *izen* (autrefois), ne fonctionne, nous semble-t-il, pas comme l'arrière-plan pour la phrase qui la suit. La présence dans la seconde phrase d'un autre SN de temps, *sono aida jû* (pendant ce temps-là), semble d'ailleurs confirmer l'absence de cadre temporel.

En l'état actuel de nos travaux, il est encore difficile de trancher le statut du SN nu fonctionnant comme complément de temps apparaissant après le thème. En revanche, les SN nus apparaissant avant le thème peuvent être considérés comme des introducteurs du cadre temporel.

6.5.9 Compléments spatiaux

Il existe en japonais deux particules de cas introduisant des compléments de lieu : les particules *ni* et *de*. Contrairement à la différence des deux types de compléments de temps résidant plutôt sur un plan énonciatif, leur différence se trouve ici sur un plan syntactico-sémantique. Le choix entre ces deux particules de cas est donc presque fixe selon la nature du verbe et un mauvais emploi produit des phrases agrammaticales. De par cette nature pratique, les études sur leurs différences sont abondantes, en particulier dans le domaine de l'enseignement du japonais langue étrangère[27].

Toutefois, les études sur la relation entre la place et la portée sont, encore une fois, quasi absentes – sinon nulles.

[27] Pour la définition des emplois de ces deux particules pour les compléments de lieu, voir par exemple Alfonso (1966). Nous trouvons une explication claire de leurs différences avec des exemples frappants d'erreurs dans Suzuki (1978).

Toujours fidèles à notre hypothèse, nous tentons d'élucider la différence non pas entre des constructions différentes, mais entre les mêmes constructions apparaissant dans des places différentes.

Constructions sans particule de cas

Bien que l'on en parle peu, il existe des compléments de lieu constitués d'un SN sans particule. Nous avons trouvé dans nos extraits deux exemples de cette construction.

> 帰り道、**僕は**　車の中で　突然、初めて　デートした　女の子のことを　思い出した。 (MURAKAMI)
> (*kaeri michi - boku wa - kuruma no naka de - totsuzen - hajimete - dêtoshita - onna no ko no koto wo - omoidashita*)
> (chemin du retour - moi [wa] - intérieur de la voiture [de] - tout d'un coup - pour la première fois - sortir - jeune fille [wo] - se souvenir [passé])
> « Sur le chemin du retour, je me suis souvenu brusquement dans la voiture de la fille avec qui je suis sorti pour la première fois. »

Dans cette phrase, le SN disloqué en tête « *kaeri michi* (chemin du retour) » fonctionne, sans particule, comme un adverbe de lieu. Mikami considère ce mot « *michi* (chemin) » – qui, sans particule, ne peut généralement pas être élément de phrase, mais qui peut avoir un emploi adverbial lorsqu'il est déterminé par d'autres éléments, comme les *kyûchaku go* (mots agglutinants). Ce syntagme a comme tête le substantif de lieu, « chemin ». Mais, le complément de lieu du prédicat étant plutôt le SN introduit par la particule de cas *de*, « dans la voiture », il semble jouer une fonction proche de celle d'introducteur du cadre temporel.

> 元役員の求めで市関係部局の課長級幹部が参加した会議の席上、**元役員は**「いつまでこういう事業を続けるのか。公社としては、もうやめたい」と訴えた。 (YOMIURI)
> (*moto yakuin no motome de shi kankei bukyoku no kachôkyû kanbu ga sanka shita kaigi no sekijô - moto yakuin wa - itsumade kôiu jigyô wo tsuzukeru noka - kôsha toshite wa - mô yametai - to uttaeta*)
> (Sur la place de la réunion à laquelle assistaient les cadres du niveau directeur des sections concernées de la mairie suite à la demande de l'ancien administrateur - ancien administrateur [wa] - jusqu'à quand on continue une telle affaire [interrogation] - quant à la régie - vouloir le cesser - proclamer [passé])
> « Pendant la réunion à laquelle assistaient les cadres supérieurs des sections concernées de la mairie suite à sa demande, l'ancien administrateur proclama : "Jusqu'à quand continuons-nous une telle affaire ? Nous, la régie, nous aimerions la cesser". »

La phrase suivante décrit une scène se déroulant toujours devant le même arrière-plan « pendant la réunion », où l'ancien administrateur n'a pas réussi à obtenir le résultat souhaité devant les employés de la mairie qui lui demandaient la continuation. La scène change la phrase d'après par l'installation du nouveau cadre

avec l'introducteur « その後 » (*sonogo*, plus tard), placé et détaché en tête de la phrase.

Constructions avec particule de cas *de*

Nous n'avons pas trouvé d'exemple avec le syntagme en *de* jouant le rôle de complément de lieu proprement dit, c'est-à-dire le syntagme constitué d'un SN se référant à un lieu suivi de la particule *de*. Il existe en revanche beaucoup d'exemples avec le SN en *de* assurant le rôle du complément spatial dans un sens plus abstrait, désignant non pas un lieu physique mais plutôt un domaine abstrait. Ces éléments sont abondamment utilisés dans le corpus Yomiuri. Ils servent vraisemblablement d'introducteur d'un cadre spatial abstrait.

Considérons quelques exemples[28].

1. 耐震強度偽装事件で　**警視庁などの合同捜査本部は**　小嶋進容疑者を、神奈川県藤沢市の分譲マンションを巡る詐欺容疑で　逮捕した。 (YOMIURI)
   (*taishin kyôdo gisô jiken de - keishichô nado no gôdô sôsa honbu wa - kojima susumu yôgisha wo - kanagawaken fujisawashi no bunjô manshon wo meguru sagiyôgi de - taiho shita*)
   (affaire de la falsification de la résistance sismique [de] - quartier général chargé de l'enquête constitué entre autres de la Préfecture de police [wa] - suspect, Kojima [wo] - inculpation d'escroquerie pour l'appartement en vente de la ville de Fujisawa à Kanagawa [de] - arrêter [passé])
   « Dans l'affaire de la falsification de la résistance sismique, le quartier général chargé de l'enquête constitué entre autres de la Préfecture de police a arrêté le suspect, Kojima, au motif d'escroquerie pour l'appartement en vente dans la ville de Fujisawa à Kanagawa. »

Dans cet exemple, première phrase de l'article, le thème ne semble pas pouvoir remplir les conditions d'être thème sans être précédé par cet introducteur du cadre, et l'inversion de l'ordre « introducteur du cadre – thème » est probablement impossible, ce qui démontre l'extériorité encore plus forte que le thème de cet introducteur du cadre.

1. 専門家会合で　**米側は**　35の食肉処理施設を対象に行った再点検の結果などに関する報告書を　説明した。 (YOMIURI)
   (*senmonka kaigô de - beigawa wa - 35 no shokuniku shorishisetsu wo taishô ni okonatta saitenken no kekka nado ni kansuru hôkokusho wo - setsumei shita*)
   (réunion des spécialistes [de] - côté américain [wa] - rapport concernant le résultat du contrôle réalisé sur 35 établissements de traitement de la viande alimentaire [wo] - expliquer [passé])
   « Lors de la réunion des spécialistes, le côté américain a présenté le rapport concernant le résultat du contrôle réalisé sur 35 établissements de traitement de la viande alimentaire. »

[28] Les exemples sont légèrement simplifiés, pour éviter une complexité superflue, par la suppression des qualificatifs, jugée sans impact sur les relations entre les compléments primaires.

Dans cet exemple – phrase non initiale dans l'article –, les deux arguments, « réunion des spécialistes » et « côté américain », sont des informations déjà données dans le contexte. L'introducteur du cadre de cet exemple, *senmonka kaigô de* (lors de la réunion des spécialistes), a une portée plus large – s'étendant aux trois phrases suivantes –, que le thème, *beigawa wa* (le côté américain), qui, lui, perd sa fonction de thème à la fin de la phrase à laquelle il appartient.

Pour ces constructions « SN + *de/ni* (PC locative) », il existe des travaux réalisés par Klingler qui les considère également comme des circonstants cadratifs.

Travaux de Klingler

Contrairement à notre approche, dans les travaux de Klingler (2003), cette faculté d'ouverture du cadre est considérée comme donnée non par la place mais par la particule *wa*. En effet, Klingler considère que le SN nu introduit par *wa* fonctionne comme un topic du discours alors que le complément circonstanciel introduit par cette particule ouvre un cadre :

> « si l'on entend par topic l'élément sur lequel porte la prédication dans une mise en relation d'*aboutness*, on ne parlera de topic que pour le premier cas [i.e. SN+wa], un SNwa étant en plus susceptible de devenir un topic de discours [...]. On réservera au second cas [i.e. SN+ni/de+wa] le terme de cadre, dès lors que wa intervient après une particule de complément circonstanciel. On n'a pas dans ce cas de relation d'*aboutness* et de tels éléments ne sont pas susceptibles de devenir des topics de discours. »

Néanmoins, la frontière entre le cadre et le topic est souvent beaucoup moins nette. En effet, comme le signalent Le Goffic (1993a) et Charolles (2003), les circonstants initiaux fixent le cadre de la phrase qui peut constituer un véritable thème ou une partie du thème. Charolles (*Ibid.*) illustre cette observation par les exemples suivants.

1. *Lola sortit faire un tour.* ***En bas de l'immeuble****, elle croisa le facteur. Il portait un énorme sac et elle lui proposa de l'aider.*
2. *Lola sortit faire un tour.* ***En bas de l'immeuble****, un homme faisait les cents pas. Une fine couche de givre recouvrait le sol. Des enfants rentraient de l'école en se chamaillant.*

Dans l'exemple 1, « le circonstant spatial se contente de fixer le cadre où se déroule une série d'événements qui ont une cohérence indépendamment de l'endoit où ils se produisent », tandis que dans l'exemple 2, le texte est « à propos de ce qui se passe dans le lieu indiqué par le SP antéposé, les événements n'ont d'autre point commun que de se dérouler dans le même endroit ».

Cette compatibilité de ces deux notions semble révéler leur différence de plan. Nous préférons donc – au lieu d'associer le SN nu suivi de wa à la notion de thématisation et le SN suivi d'une particule circonstancielle terminé par wa à celle de cadre – défendre l'idée que l'installation du cadre est réalisée par la mise en tête

ou (le détachement) et que la particule *wa* sert à marquer non pas l'ouverture du cadre mais plutôt la thématisation (ou la mise en contraste).

Cette hypothèse est partiellement encouragée lorsqu'on réfléchit à la description en japonais des mêmes types de scènes que représentent les exemples de Charolles précédemment cités. Voici les traductions :

1. ローラは、少し散歩でもしようと家を出た。
   (*rôra wa - sukoshi snpo demo shiyô to ie wo deta*)
   « Lola sortit faire un tour. »
   アパートの下（　　）、ローラは郵便配達のおじさんに出くわした。
   (*apâto no shita ( ) - rôra wa haitatsu no ojisan ni dekuwashita*)
   « En bas de l'immeuble, elle croisa le facteur. »
   おじさんは、大きな包みを抱えていた。
   (*ojisan wa - ookina nimotsu wo kakaeteita*)
   « Il portait un énorme colis. »
   ローラは、おじさんに「お手伝いしましょうか」と声をかけた。
   (*rôra wa - ojisan ni "otetsudai shimashôka" to koe wo kaketa*)
   « Elle lui demanda : "Voulez-vous de l'aide ?". »

2. ローラは、少し散歩でもしようと家を出た。
   (*rôra wa - sukoshi snpo demo shiyô to ie wo deta*)
   « Lola sortit faire un tour. »
   アパートの下（　　）、男が道を行ったり来たりしていた。
   (*apâto no shita ( ) - otoko ga michi wo ittari kitari shiteita*)
   « En bas de l'immeuble, un homme faisait les cents pas. »
   薄い霜の層が地面を覆っている。
   (*usui shimo no sô ga jimen wo ootteiru*)
   « Une fine couche de givre recouvrait le sol. »
   子供たちは、じゃれあいながら家路についていた。
   (*kodomo-tachi wa - jareai nagara ieji ni tsuiteita*)
   « Des enfants rentraient de l'école en se chamaillant. »

Nous n'avons pas mis de particule après le circonstant spatial. Si nous nous appuyions sur la thèse de Klinger, dans l'exemple 1, nous remplirions l'espace entre parenthèses avec *dewa* car « le bas de l'immeuble » doit ouvrir un cadre, et dans l'exemple 2, on choisirait plutôt *wa* sans particule puisque « le bas de l'immeuble » est ce dont on parle. Or, nous choisissons, avec une très haute certitude, *de* sans *wa* pour l'exemple 1 et *dewa* pour l'exemple 2.

Ce choix correspond bien à notre hypothèse : dans les deux exemples, les circonstants, antéposés, ouvrent un cadre introduit ; dans l'exemple 1 où il est introduit par *de* sans *wa*, il ne devient pas le thème du discours contrairement à l'exemple 2 où il est introduit par *de* avec *wa* et où il constitue le thème.

6.5.10 Éléments ouvrant d'autres types de cadres

Il existe, notamment dans le corpus Yomiuri, beaucoup de locutions introduisant des syntagmes qui apparaissent avant le thème. Elles semblent servir à introduire différents types de cadres.

> 同署によると、**豪憲君は**　１７日午後３時ごろ、同級生４人と下校。 (YOMIURI)
> (*dôsho ni yoruto - gôken-kun wa - jûshichinichi gogo sanjigoro - dôkyûsei yonin to gekô*)
> (selon la police locale - petit Gôken [wa] - rentrer de l'école avec ses quatre camarades vers 3 heures de l'après-midi le 17 [passé])
> « Selon la police locale, le petit Gôken est rentré de l'école avec ses quatre camarades vers 3 heures de l'après-midi le 17. »

La locution *niyoruto* (ou *niyoreba*) peut être traduite en français par « selon », qui ouvre les cadres que Charolles appelle d'énonciation. Il crée un univers qui renvoie à un discours tenu par un énonciateur différent de l'auteur du texte. La fermeture du cadre n'est pas aussi claire que dans le cas du cadre spatial que nous venons de voir, mais il semble s'étendre sur les 6 phrases suivantes avant que le terme « police » ne soit repris en tant que thème, ou sur les 4 phrases avant l'apparition d'un nouveau thème « police départementale ».

> 新制度について、**内閣官房幹部は**　、「履修証明を制度化して、信頼性を高めることにより、企業が人材を採用する際の基準の一つにしたい」と期待している。 (YOMIURI)
> (*shinseido ni tsuite - naikaku kanbô kanbu wa - "rishûseido wo seidoka shite - shinreisei wo takameru koto niyori - kigyô ga jinzai wo saiyô suru sai no kijun no hitotsu ni shitai" - to kitaishiteiru*)
> (à propos du nouveau système - les dirigeants du Secrétariat du Cabinet [wa] - faire du système de l'attestation de l'obtention d'unités de valeur (UV) un système officiel - augmenter sa fiabilité - vouloir le faire devenir un critère lors de l'embauche de personnel par les entreprises" - [citation] espérer)
> « Pour ce nouveau système, les dirigeants du Secrétariat du Cabinet espèrent : "Nous souhaitons, en faisant du système d'attestation d'obtention d'UV un système officiel pour augmenter sa fiabilité, le faire devenir un critère lors de l'embauche de personnel par les entreprises". »

Cette locution *nitsuite*, traduite en français par « à propos de », fixe, semble-t-il, le cadre que Charolles désigne par le terme de cadre thématique. Non remplaçable par la particule *wa*, cette locution a une fonction différente de cette particule de thématisation. Mais nous reportons une analyse plus précise à nos travaux futurs.

6.5.11 Compléments avec particule de cas

Dans le corpus que nous avons étudié, nous avons trouvé plusieurs exemples de compléments essentiels avec particule de cas apparaissant avant le thème. Nous distinguons tout simplement les compléments avec particule de cas de ceux sans particule de cas, mais ces deux types de compléments semblent deux catégories contiguës à l'intersection floue et tous les compléments avec particule de cas n'entretiennent pas non plus de relation de même nature avec le prédicat.

D'après nos premières réflexions, il est difficile de considérer les compléments avec une particule de cas – en particulier *ga, wo* – comme extérieurs à la proposition. Leur place semble plus liée à l'effet de mise en focus, bien que nos études soient trop incomplètes pour pouvoir tirer une quelconque conclusion.

Nous n'avons pas trouvé d'exemple de complément en *ga* mis avant le thème, mais des constructions avec différentes autres particules, y compris la particule *de* indiquant non pas le lieu mais le moyen, la cause, etc.

1. 昔、あれほど忌み嫌い憎んだ宿題を　**この連中は**　要求しているのだ。(FUJIWARA)
   (*mukashi - arehodo imi kirai nikunda shukudai wo - konorenchû wa - yôkyû shiteirunoda*)
   (autrefois - devoir que (j'ai) si détesté et haï [wo] - ces gens-là [wa] - exiger, en effet)
   « En effet, ces étudiants-là exigent de moi des devoirs qu'autrefois je détestais et haïssais tellement. »
2. 女子大生だという七瀬の身分に　**彼は**　半信半疑であった。(TSUTSUI)
   (*joshidaisei da to iu nanase no mibun ni - kare wa - hanshinhangi de atta*)
   (identité de Nanase qui dit qu'elle est étudiante [ni] - lui [wa] - être incrédule [passé])
   « Il était incrédule quant à l'identité de Nanase qui disait qu'elle était étudiante. »
3. 左手の指が４本しかない女の子に　**僕は**　二度と会えなかった。(MURAKAMI)
   (*hidarite no yubi ga yonhon shika nai onnanoko ni - boku wa - nidoto aenakatta*)
   (fille qui n'a que quatre doigts à la main gauche [ni] - moi [wa] - ne jamais pouvoir rencontrer à nouveau [passé])
   « Je n'ai plus jamais pu revoir la fille avec seulement quatre doigts à la main gauche. »
4. ビーフシチューの湿っぽい熱気で　**部屋の中は**　ひどく蒸し暑かった。(MURAKAMI)
   (*bîfushichû no shimeppoi nekki de - heya no naka wa - hidoku mushiatsukatta*)
   (air chaud, humide, du ragoût de bœuf [de] - intérieur de la chambre [wa] - faire très chaud et humide [passé])
   « À cause de l'air chaud, humide, du ragoût de bœuf, dans la chambre, il faisait très chaud et humide. »

Dans ces exemples, tous tirés d'œuvres littéraires, l'interprétation de l'extériorité de ces compléments est très délicate. La détermination de l'effet exact né-

cessiterait des études sur un plus grand nombre d'exemples avec analyse de leur contexte.

6.5.12 Questions en suspens

Nous pouvons déduire de ce que nous venons de voir que la partie initiale de la phrase japonaise est une place privilégiée pour les éléments qui ne participent pas à la constitution du noyau syntaxique et qu'il existe beaucoup de types d'éléments externes, autres que les *dokuritsu-go* ou *setsuzoku-go* que les grammaires usuelles qualifient d'extérieurs, notamment l'ensemble des introducteurs du cadre. Toutefois, il existe encore beaucoup de questions en suspens :

- la différence entre le cadre thématique (introduit par les expressions) et le thème (introduit par *wa*), ainsi qu'entre le thème et le cadre-thème ;
- l'effet de mise en avant des compléments en particule de cas ;
- la catégorisation systématique de l'ensemble des éléments antéposés par rapport au thème favorisera sans doute la clarification du contour de la classe des éléments externes.

Il y a également des questions sur la frontière tracée par la particule *wa*. Existe-t-il d'autres indications de frontières dans les phrases sans syntagme en *wa* ? Tous les syntagmes en *wa* qui ont servi d'indicateur de frontière étaient-ils vraiment de même nature ?

Nous essayerons d'appliquer les hypothèses que nous avons défendues dans cette section à nos réalisations informatiques et d'évaluer leur justesse ou inconvénients dans les résultats obtenus et nous fournirons des données utiles aux études linguistiques futures sur le sujet.

CHAPITRE 7

ÉTUDE DE LA PHRASE COMPLEXE

Dans ce chapitre, nous tenterons de cerner les notions connexes à la phrase complexe, notamment la proposition. Le chapitre, ouvert par la définition des deux problèmes centraux de la proposition japonaise (§ 7.1), sera composé de quatre parties. La première sera dédiée au premier problème, que nous commencerons par décrire (§ 7.2), avant de nous intéresser aux travaux antérieurs sur la définition de la proposition (§ 7.3), qui nous permettront de découvrir les critères concrets pour déterminer les syntagmes à mot variable non-propositionnels (§ 7.4), et de réaliser la première définition de la proposition et de quelques autres unités (§ 7.5). La seconde partie, consacrée au second problème, sera constituée d'une description de celui-ci (§ 7.6) suivie d'un état de l'art (§ 7.7) et se terminera par la présentation de notre catégorisation des mots susceptibles de réaliser la connexion des propositions (§ 7.8). La troisième partie comportera les études sur la typologie des subordonnées, composées d'un état de l'art (§ 7.9) et de la présentation de notre typologie (§ 7.10). Enfin, la dernière partie commencera, pour une récapitulation, par une définition formelle de la phrase (§ 7.11), et se terminera par un débat sur les principaux problèmes de la phrase complexe liés à l'analyse syntaxique : les questions des relations entre le syntagme thématisé et les subordonnées (§ 7.12) et les problèmes liés au phénomène d'ellipse (§ 7.13).

7.1 Deux questions centrales pour une définition de la proposition

Nous avons défini jusqu'ici la phrase comme constituée d'une proposition et éventuellement d'un thème. Mais la proposition peut comporter une sous-

structure phrastique. Nous appelons les phrases contenant une sous-structure phrastique des phrases complexes. Néanmoins, nous n'avons pas encore défini exactement ces structures phrastiques que nous appelons proposition.

Suivant le modèle de la grammaire anglaise, Hashimoto a essayé de définir la proposition comme une unité comportant à la fois un sujet et un prédicat, ce qui a été largement critiqué par beaucoup de linguistes comme nous l'avons déjà vu dans la section 6.1.

Cependant, aucun de ces opposants n'a réussi à proposer de définition satisfaisante nous permettant de repérer réellement les propositions. En effet, ils semblent avoir réussi à distinguer la proposition de la phrase sur un plan conceptuel, mais ils ne sont pas arrivés (ou n'ont pas cherché) à trouver d'indications formelles de cette distinction.

Dans les travaux contemporains, Masuoka & Takubo (1992) définissent d'abord la phrase simple comme une phrase constituée d'un seul prédicat et la phrase complexe comme une phrase constituée de plusieurs prédicats. Ils appellent alors propositions les unités de la phrase complexe constituées d'un prédicat et de ses compléments.

Néanmoins, cette définition, très grossière, ne permet pas non plus la détermination exacte de la proposition, dans la mesure où la notion de prédicat est également floue. En effet, toutes les occurrences des mots variables ne peuvent pas être considérées comme des prédicats, noyaux de la proposition. Devant cette difficulté, certains comme Mikami et Garnier ont renoncé à la notion de proposition.

En plus de ce problème de distinction difficile entre les propositions et les syntagmes non-propositionnels à mot variable, il en existe un autre lié à la définition des éléments connecteurs. En effet, pour relier deux constructions à mot variable (y compris les propositions) en japonais, existent deux possibilités : d'une part marquer la connexion par une forme connective du mot variable subordonné et, de l'autre, réaliser la connexion à l'aide d'un élément suivant une forme autonome du mot variable de la subordonnée. Le second problème concerne donc les éléments assurant la connexion dans les structures de ce second type.

Nous nous appliquerons donc ici à élucider la notion de proposition en résolvant ces deux problèmes.

7.2 Premier problème : natures différentes des syntagmes à mot variable

Nous trouvons dans Mikami (1959) un passage dans lequel Mikami souligne l'impossible – ou même insensée – introduction de la notion de proposition dans l'analyse du japonais :

> Dans la grammaire anglaise, les séquences de mots sont distinguées en deux catégories :
> – *clause* (comprenant l'opposition sujet-prédicat, « proposition ») ;

> – *phrase* (sans opposition sujet-prédicat, « syntagme »).
> [...] Comme il n'existe pas d'opposition sujet-prédicat, la distinction entre la proposition et le syntagme est aussi impossible. Ainsi, l'unité « proposition » comprenant l'opposition sujet-prédicat ne concerne pas la grammaire japonaise, et l'unité « syntagme » est donc suffisante, ou plutôt nous sommes obligés de nous contenter de cette unité.

Il semble cependant rester une possibilité, celle de définir la proposition par la présence d'un prédicat conjugué. Mais, lorsque nous essayons d'extraire les propositions avec des mots variables comme repères, nous obtenons des unités paraissant avoir une nature assez différente, d'autant plus que le japonais n'a pas non plus d'opposition entre les formes conjugées et la forme infinitive pour ses mots variables. Il existe donc aussi bien des syntagmes contenant un mot variable – qui nous semblent très semblables à une phrase –, que d'autres qui sont presque des adverbes.

Si on refuse de renoncer complètement à la notion de proposition, cette absence de repère absolu pour la définition de la proposition pose un véritable problème pour distinguer une phrase complexe d'une phrase simple.

赤ん坊　が　歩く　ようになる。
(*akanbô - ga - aruku - yô ni naru*)
(bébé - [*ga*] - marcher - devenir faire q.c.)
« Le bébé commence à marcher. »

「古都」と　いう　本　を　読んだ。
(*koto - to - iu - hon - wo - yonda*)
("*koto*" - [*to* (citation)] - dire - livre - [*wo*] - lire [passé])
« (J')ai lu le livre intitulé "*koto*". »

Ces deux exemples contiennent chacun deux mots variables, bien qu'intuitivement nous les considérions plus facilement comme des phrases simples que complexes.

Dans un ouvrage récent sur la phrase complexe de Noda (2002), par exemple, plusieurs critères supplémentaires sont proposés pour définir la proposition : le nombre minimum d'arguments, la présence de la notion de temps, la possession du sens concret, etc. Mais comme les auteurs le soulignent, cela reste une question de probabilité : il existe beaucoup de cas difficiles à juger.

Ainsi, il est très difficile de tracer une frontière entre les phrases simple et complexe, c'est-à-dire de définir la proposition.

7.3 État de l'art des travaux visant à définir la proposition

Nous présentons dans cette section les travaux sur la définition de la proposition. Nous allons aborder tout d'abord l'abandon de cette notion par Mikami (§ 7.3.1) et Garnier (§ 7.3.2), qui ont cherché plutôt à déterminer les différences

de nature entre tous les types de syntagmes contenant un mot variable, puis les tentatives de définition de cette unité par les autres, Minami (§ 7.3.3), Teramura (§ 7.3.4) et Noda (§ 7.3.5). Pour terminer l'exposé, la dernière partie sera consacrée à notre analyse critique de ces travaux antérieurs (§ 7.3.6).

7.3.1 Capacités phrasogénératrices des prédicats selon Mikami

Au lieu de regrouper certains syntagmes contenant un mot variable sous une même étiquette de proposition, Mikami a cherché à définir différentes catégories pour les syntagmes à mot variable.

Études des formes des mots variables

Pour lui, chaque forme de mot variable a une capacité phrasogénératrice de niveau différent et plus la capacité phrasogénératrice du mot variable du syntagme est élevée, plus ce syntagme est autonome, c'est-à-dire proche de la phrase. Il a représenté la capacité phrasogénératrice de différentes formes de mots variables – c'est-à-dire la capacité à constituer en tant que prédicat un constituant aussi phrastique que possible – par une valeur, comme suit (Mikami, 1953) :

Forme neutre	1/4
Forme autonome (emploi déterminant)	1/2
Forme de condition	3/4
Formes autonome et impérative	1

TAB. 7.1 – Capacités phrasogénératrices des formes des mots variables

Comme nous l'avons déjà vu dans la section 6.2.2, Mikami distingue quatre types de connexions dans la phrase japonaise. Mikami considère la première, forme neutre, comme produisant une connexion de style simple, la forme de condition celle de style complexe souple et les formes autonome et impérative celles de style complexe dur. La forme autonome en emploi déterminant ne réalise pas de connexion avec le syntagme à mot variable, mais elle produit une connexion particulière appelée déterminante.

Comme l'affirme Mikami (1953), cette idée de classification des formes des mots prédicatifs est inspirée des travaux de Mio (1942) qui a présenté le résultat d'études, réalisées avec un corpus, sur la possibilité de transformation en forme polie des mots prédicatifs précédant une particule conjonctive.

Études des emplois des formes des mots variables

Nous constatons cependant une évolution de cette catégorisation au cours des années de recherche de l'auteur. En effet, chaque forme, en particulier la forme autonome, produit différents types de connexions selon son emploi effectif.

Reprenons les exemples de Mikami (1955). Les trois phrases suivantes sont toutes constituées de la partie exprimant la cause « je me suis réveillé trop tard » introduite par trois connecteurs introducteurs de cause différents (*tameni, node, kara*), suivie du prédicat principal « (j')étais en retard ».

寝坊した	ために	(*tame ni*)	遅刻した
(*nebôshita*)	ので	(*node*)	(*chikoku shita*)
(se réveiller trop tard [passé])	から	(*kara*)	(être en retard [passé])

En dépit de l'identité de la forme, la construction que constitue le premier verbe « se réveiller trop tard » avec son introducteur produit différents types de connexions selon le connecteur. Cette différence est claire avec un test d'enchâssement dans la structure déterminante. Quand on rajoute « *koto wa sukunai* (le fait - [wa] - peu) » à la fin des trois phrases ci-dessus, la phrase avec *tameni* est totalement enchâssée dans la structure déterminante :

nebôshita ***tameni*** *chikokushita* *koto* *wa sukunai*

« Le fait que je sois en retard parce que je me suis réveillé trop tard est rare »

En revanche, celle avec *kara* sort complètement de l'enchâssement et constitue une phrase sémantiquement incorrecte :

nebôshita* ***kara* *chikokushita* *koto* *wa sukunai*

« *Le fait que je sois en retard est rare, parce que je me suis réveillé trop tard »

La phrase avec *node* sort généralement en dehors de l'enchâssement comme la phrase avec *kara*, mais elle peut être interprétée comme la phrase avec *tameni*. Le résultat de ce test est donc comme suit :

- Verbe à la forme autonome (emploi déterminant) + *tame ni*
  = connexion de style simple ;
- Verbe à la forme autonome (emploi conclusif) + *node*
  = connexion de style complexe souple ;
- Verbe à la forme autonome (emploi conclusif) + *kara*
  = connexion de style complexe dur.

Dans Mikami (1959), il présente la catégorisation non pas des formes mais des emplois des formes (cf. tableau 7.2 (voir page suivante)), en cinq ensembles : emplois infinitif, neutre, de condition, conclusif non final et conclusif final. L'emploi conclusif non final désigne le cas où la forme conclusive ne se situe pas en fin de phrase mais qu'elle est suivie d'une particule conjonctive avec laquelle elle constitue une construction phrastique dépendant syntaxiquement du prédicat principal.

Emploi infinitif	0
Emploi neutre	1/4
Emploi de condition	1/2
Emploi conclusif (non final)	3/4
Emploi conclusif (final)	1

TAB. 7.2 – Capacités phrasogénératrices des emplois des mots variables

L'emploi de condition est le seul emploi de la forme de condition et c'est la seule forme ayant l'emploi de condition. Il en va de même pour la forme neutre et l'emploi neutre. L'infinitif est un des emplois de la forme autonome, forme qui a en outre les emplois conclusifs final et non final ainsi que l'emploi déterminant. Les formes volitive et impérative ont également l'emploi conclusif mais la forme impérative n'a pas d'emploi conclusif non final.

Lexicalisation des formes des mots variables

Par ailleurs, il parle également de la lexicalisation de certaines formes des mots variables. C'est le cas de l'exemple présenté dans la section précédente décrivant la difficulté de la définition (§ 7.2) :

「古都」と　いう　本　を　読んだ。
(*koto* - *to* - *iu* - *hon* - *wo* - *yonda*)
("*koto*" - [*to* (citation)] - dire - livre - [*wo*] - lire [passé])
« (J')ai lu le livre intitulé "*koto*". »

Dans cet exemple, le verbe « *iu* (dire) » a perdu la véritable fonction de prédicat et constitue avec la particule *to* qui le précède, un connecteur introduisant des éléments déterminants. Il cite quelques conditions empêchant cette lexicalisation, comme par exemple : possession d'un ou plusieurs compléments, avec particules de cas ou non, et ce notamment avec un acteur animé ; présence du complément en *ga*.

7.3.2 Les trois classes des syntagmes à mot variable de Garnier

Garnier (1982) cherche également à saisir la nature différente des syntagmes à mot variable.

Deux nouvelles unités

Renonçant à toutes les unités « classiques » mal définies, elle définit tout d'abord ses propres unités syntaxiques dont les deux principales, supérieures au syntagme minimal et inférieures à la phrase :

– séquence : unité formée par un mot prédicatif et ses compléments, y compris le thème en *wa* ;
– unité de discours : toute séquence qui n'a aucune contrainte pour les trois possibilités maximales : la réception de suffixes fonctionnels, la réception de segments en statut de complétisation et la factorisation (ou thématisation) de segments.

Elle étudie ensuite les phrases à deux (ou plusieurs) mots variables. En analysant combien de séquences et unités de discours elles contiennent, elle réalise une catégorisation de chaque forme des mots variables.

Les trois classes des syntagmes à mot variable

Ses travaux catégorisent finalement les syntagmes à mot variable en trois types[1] :

1. syntagme qui ne peut constituer une séquence fonctionnant comme un élément de phrase qu'avec un autre syntagme à mot variable en fin de phrase :

   *sôshite - sanbo wo - tabeyô **to** - shimashita*
   (et - Sanbo - manger - (faire))
   « Et il s'apprêta à avaler Sanbo. »

2. syntagme qui peut constituer à lui seul une séquence, mais qui ne peut constituer une unité de discours qu'avec une autre séquence qui le suit :

   *atarashii kôsha ga - dekita - **toki** - minna - taihen - yorokobimashita*
   (nouvelle école - être fini - moment - tous - très[2] - se réjouir)
   « Quand la nouvelle école fut finie, tout le monde se réjouit. »

3. syntagme qui peut constituer à lui seul, non seulement une séquence, mais aussi une unité de discours :

   *zepetto ga - tomemashita - **ga** - pinokio wa - kikimasen deshita*
   (Gepetto - arrêter - [ga] - Pinocchio - ne pas écouter)
   « Gepetto voulut l'arrêter mais Pinocchio ne l'écouta pas. »

Syntagmes à mot variable qui ne sont pas des séquences

Étant donné la définition, la différence entre la séquence et l'unité de discours est nette, mais celle entre la séquence et le segment à mot variable qui ne peut pas constituer tout seul un élément de phrase est plus difficile à saisir.

Il y a deux types de syntagmes à mot variable de ce genre :
– hyper-séquence : constituée de deux syntagmes à mot variable qui se complètent ;

[1] Exemples et traductions tirés de Garnier (*Ibid.*). La transcription de la particule *wo* est « o » dans l'original. Le mot variable concerné est souligné et le connecteur ou le substantif cheville est mis en gras par nous-mêmes, en modifiant le style de l'original.

[2] Absence de traduction dans l'original.

– semi-séquence incluse : ayant la possibilité de posséder ses propres compléments – avec tout de même quelques restrictions – qui ne peut fonctionner cependant qu'en tant qu'élément d'une séquence.

Hyper-séquence Ce sont deux syntagmes à mot variable contigus. Le premier comporte des compléments alors que le deuxième, verbe (ou mot prédicatif) support, sert à exprimer le temps, l'aspect et la modalité. Ces deux syntagmes se complètent et ne peuvent fonctionner en tant qu'élément de phrase qu'ensemble.

1. *sôshite - sanbo wo - tabeyô* ***to*** *- shimashita*
   (et - Sanbo - manger - (faire))
   « Et il s'apprêta à avaler Sanbo. »
2. *okane wo - ikura - haraeba - yoi deshô ka*
   (argent - combien - payer - bien c'est)
   « Combien dois-je payer ? »

Le deuxième mot variable restant toujours un syntagme minimal – donc aucune insertion possible entre les deux mots variables –, la détermination des hyper-séquences est relativement aisée.

Semi-séquence incluse Il s'agit de syntagmes à mot variable avec leurs compléments, qui sont enchâssés à l'intérieur d'une séquence et qui constituent un élément de cette dernière.

1. *yagate - morikara - dongajîsan ga -* [*taiko wo - narashi* ***nagara***] *- arawareta*[3]
   (bientôt - forêt - le vieux Donga - tambour - faire résonner - se montrer)
   « Bientôt, le vieux Donga, tapant sur son tam-tam sortit de la forêt. »
2. *beruto no ue de - jidôsha ga -* [*sukoshizutsu - kumitaterarete*] *- dandan - dekigatte ikimasu*
   (chaîne sur - voiture - morceau par morceau - être assemblé - peu à peu - se terminer)
   « Les voitures se construisent petit à petit, montées pièce par pièce sur la chaîne. »

Ce sont des syntagmes suivis d'un mot connecteur *nagara* (cf. ex. 1) ou *mama*, ou terminés par une forme adverbiale (cf. ex. 2).

L'analyse de Mikami pour ce type de phrase diffère. Il interprète les compléments antérieurs au premier mot variable comme dépendant de ce dernier. Mais la connexion que produit ce dernier est telle que les compléments franchissent ce premier mot prédicatif et dépendent à nouveau du deuxième mot variable. C'est la définition même de la connexion de style simple.

Malgré cette différence d'analyse, la conclusion de ces deux auteurs sur l'autonomie de ces mots variables semble très proche : la capacité phrasogénératrice de ces mots variables (ou les formes de ces mots variables) est très faible.

Pour l'analyse des compléments communs, nous suivons le modèle de Mikami. Ce choix provient cependant d'une raison pratique d'analyse automatique[4]

[3] L'encadrement est absent dans l'original, mais rajouté pour présenter clairement la structure enchâssée dont parle Garnier.

[4] Tous les compléments antérieurs au premier mot variable – sauf bien évidemment les éléments extérieurs tels que le thème ou les éléments indépendants en tête de phrase – dépendent soit de ce

et la détermination de leur appartenance du point de vue linguistique nécessiterait une étude beaucoup plus poussée que nous ne pouvons pas réaliser dans le cadre de la présente thèse.

7.3.3 Les propositions subordonnées de Minami

Dans ses travaux destinés à définir quatre couches constituant la phrase japonaise – par l'étude de la possibilité d'apparition ou non des éléments dans différentes structures subordonnées –, Minami définit tout d'abord les propositions subordonnées.

Définition des subordonnées

En expliquant que la subordonnée est un type d'unité ayant une structure proche de la phrase et qui apparaît dans une phrase, il énumère dans Minami (1993) les unités qu'il considère comme des subordonnées :

1. unité se terminant par un mot variable à la forme neutre :
   - 大臣は　その時には　Aさんであり、また、事務次官は　Bさんでありました。
     (*daijin wa - sonotoki ni wa - A san deari - mata - jimujikan wa - B san dearimashita*)
     (ministre [*wa*] - à cette époque [*wa*] - M./ Mme. A [copule] - de plus - vice-ministre [*wa*] - M./Mme. B [copule passé])
     Le ministre était, à cette époque, Monsieur A et le vice-ministre était Monsieur B.
2. unité se terminant par une particule conjonctive telle que :

   が (*ga*), から (*kara*), けれど（も） (*keredo(mo)*), し (*shi*), たら (*tara*), だら (*dara*), て (*te*), で (*de*), ては (*tewa*), では (*dewa*), ても (*temo*), でも (*demo*), と (*to*), ながら (*nagara*), なら (*nara*), ので (*node*), のに (*noni*), ば (*ba*).

   - よく　かきまぜながら、煮る。
     (*yoku - kakimaze - nagara - niru*)
     (bien - mélanger [simultanéité] - mijoter)
     (On le) mijote en mélangeant bien.
3. unité se terminant par certains substantifs formels[5] tels que :

   あげく (*ageku*), ため (*tame*), ところ (*tokoro*)

   - 部品を　取り替えて みた　ところ、エンジンが　作動する　ようになった。
     (*buhin wo - torikaete - mita - tokoro - enjin ga - sadôsuru - yôninatta*)
     (pièce détachée [*wo*] - changer - pour voir - [sub formel *tokoro*] - moteur - fonctionner -

dernier uniquement, soit également du mot variable postérieur. Mais ils ne dépendent pas forcément tous du deuxième mot prédicatif, ce qui complique considérablement l'analyse automatique.

[5]形式名詞 (*keishikimeishi*, substantifs formels) : il s'agit d'un type de substantif qui, « tout en ayant des caractéristiques nominales, n'a que très peu de sens et n'est donc utilisé qu'avec des éléments déterminants » (Masuoka & Takubo, 1992).

devenir [passé])
Quand j'ai changé la pièce détachée pour voir, le moteur s'est remis à fonctionner à nouveau.

Après avoir ainsi défini les propositions subordonnées, il les classe en trois groupes A, B et C définis sur la base de sa théorie de la structure en couches de la phrase japonaise, que nous avons déjà vue dans la section 6.2.2.

Exception des syntagmes à mot variable non-propositionnels

Bien qu'appartenant à l'une de ces classes, les éléments constituant une locution figée, ou lexicalisés comme s'ils étaient des adverbes, sont considérés comme des exceptions.

I. constituant une locution :
   a) précédant なる (*naru*, devenir) ou する (*suru*, faire) :
      – 手　が　動かなく（なる）
      (*te - ga - ugokanaku - (naru)*)
      (main - [*ga*] - bouger [négation] - (devenir))
      Les mains ne bougent plus.
   b) précédant d'autres expressions :
      – 〜て／ても（いい、かまわない、...）
      (X *te/temo - (ii, kamawanai, ...)*)
      (X [forme neutre en *te*] / même si X - (bien, pas génant, ...))
      Vous pouvez faire X / cela ne me gène pas que tu fasses X, ...

II. lexicalisés :
   – 飛んで（帰る）
   (*tonde - kaeru*)
   (voler - rentrer)
   Rentrer à toute vitesse.

D'autres constructions phrastiques non subordonnées

Il cite d'autres constructions phrastiques qu'il ne considère pas comme des subordonnées, telles que les constructions se terminant par une impérative, les phrases incidentes, les constructions constituant des éléments indépendants de la grammaire scolaire, les discours rapportés ou les constructions phrastiques déterminantes.

7.3.4 Les deux types de phrases simples de Teramura

Teramura (1982a) fait la remarque que, malgré une décision contraire vis-à-vis de la notion de proposition, Mikami et Minami sont arrivés finalement à une

conclusion semblable[6]. D'après Teramura, la catégorie de type simple de Mikami correspond aux subordonnées du groupe A de Minami, celle de style complexe souple aux subordonnées du groupe B et enfin celle de style complexe dur aux subordonnées du groupe C.

Teramura trace ensuite la limite de la proposition, *grosso modo* entre les groupes A et B de Minami et considère seulement les subordonnées appartenant aux groupes B et C comme des propositions subordonnées.

Par ailleurs, certaines propositions du groupe B de Minami sont également considérées comme non-propositionnelles lorsqu'elles partagent des compléments avec la proposition racine. En d'autres termes, les constructions se terminant par un mot variable à la forme neutre (adverbiale) doivent comporter plus d'un complément dépendant syntaxiquement directement de son prédicat et dans Teramura (1991), l'auteur parle plus précisément du partage, non pas de n'importe quel complément, mais du complément en *ga*. Cette condition supplémentaire montre qu'un des critères formels de Teramura pour reconnaître le statut de proposition était la présence explicite du complément en *ga*.

Ainsi, Teramura (1982a) définit deux types de phrases simples :

– Type 1 :
    1. phrase contenant un seul prédicat qui s'associe avec ses compléments et qui se termine par une forme conclusive ;
    2. phrase terminée par un verbe à la forme neutre suivi d'un second verbe exprimant l'aspect.
– Type 2 :
    1. phrase contenant, en plus du prédicat, un syntagme terminé par un verbe à la forme en *te* qui fonctionne comme un adverbe ;
    2. phrase contenant, en plus du prédicat, un syntagme terminé par un verbe à la forme neutre, qui partage son/ses complément(s) avec le prédicat principal (dans Teramura (1991), précisément le complément nominatif).

7.3.5 Les frontière entre phrase simple et phrase complexe selon Noda

Nous pouvons également trouver dans Noda (2002) une discussion très riche sur la proposition. Noda y présente trois typologies de la proposition, dont l'une est réalisée selon le niveau de ressemblance à la phrase de la structure de proposition.

La catégorisation est réalisée par un test examinant la présence possible ou non dans la proposition de six types d'éléments :

1. éléments de voix ;

[6]Minami (1993) fait lui-même la même remarque en disant que ses trois catégories des groupes A, B et C correspondent aux classes proposées par Mikami, respectivement les styles simple, complexe souple et complexe dur. Mais, il doute que leur classification soit tout à fait identique et présente quelques exemples pour lesquels leurs analyses divergent.

2. éléments d'aspect ;
3. éléments de négation ;
4. éléments de temps ;
5. éléments de modalité orientée vers le contenu ;
6. éléments de modalité orientée vers l'interlocuteur.

Les propositions comportant les éléments d'un niveau donné peuvent contenir tous les éléments des niveaux inférieurs. En d'autres termes, les propositions contenant les éléments du niveau 5, peuvent comporter également des éléments des niveaux 1 à 4.

Suite à ce test, les propositions sont distinguées en six types selon la possibilité d'apparition de ces éléments. Les syntagmes à mot variable ne pouvant comprendre aucun de ces éléments sont considérés comme des syntagmes non-propositionnels.

L'auteur compare son classement avec la catégorisation de Minami comme suit :

Propositions pouvant comporter	Catégorisation de Minami
des éléments 1	groupe A
des éléments 2	groupe B
des éléments 3	groupe B
des éléments 4	groupe B
des éléments 5	groupe C

7.3.6 Analyse critique

Tout comme Teramura a constaté une correspondance entre les travaux de Mikami et ceux de Minami, nous pouvons également reconnaître certains points communs dans les études de Garnier. Le tableau 7.3 page ci-contre est un comparatif des typologies des syntagmes à mot variable de ces trois chercheurs.

Frontière de la proposition

Nous pouvons maintenant tracer, sur le modèle de Teramura, la limite de la proposition entre la catégorie de style simple et celle de style complexe de Mikami. Cette frontière correspond à celle entre le groupe A et le groupe B de Minami et à celle entre la catégorie « 1 séquence - 1 unité de discours » et la catégorie « 2 séquences - 1 unité de discours » de Garnier.

Mais quels sont les critères concrets pour reconnaître les syntagmes non-propositionnels, qui n'appartiennent pas à la catégorie limitée par la frontière que nous venons de tracer ?

Minami se contente d'énumérer les exemples types de syntagmes à mot variable qu'il ne considère pas comme des propositions. Le critère de la possibilité

Comparaison des analyses des syntagmes à mot variable

Garnier	Mikami	Minami
1 séquence – 1 unité de discours · ～たりする · ～ようとする · （連用）に行く · （連用）ながら · （連体）ままV／にしておく · （連体）ようにする／なる · （仮定）ばよいでしょう（か）／ばならない · （連用）たら／～てよいでしょう（か） · （終止）といいです／いけないです · （終止）からです · ～ていけない／ならない · ～て$_1$／（連用-手段）$_1$V	**Capacité phrasogénératrice nulle** • Formes lexicalisées • Emploi infinitif (forme autonome)	**Non objet de l'examen（考察対象外）** · なる／するに続くもの · ～たら（いい、こまる、だめだ） · ～て（いい、かまわない） · ～ては（いけない、こまる、だめだ） · ～ても（いい、かまわない、だめだ） · ～と（いい、こまる） · ～なければ（ならない、だめだ） · ～ば（いい、さいわいだ） · 等
	Style simple（単式） • Emploi neutre (forme neutre)	**A** · （連用）ながら · ～つつ · ～て$_1$／連用$_1$ · 連用形反復（なめなめ）
2 séquences – 1 unité de discours · （連体）とき、あと、まで、まえ、ころ（等）、ほど · （談話）と／か · （連体）ため、ように、には · （仮定）ば／たら · （終止）と · （連体）なら、のに、ので · ～て$_{2,3}$／連用$_{2,3}$V · ～ても、てから	**Style complexe souple（複式・軟式）** • Emploi de condition (forme de condition) • Emploi conclusif non final (forme autonome)	**B** · ～て$_2$／連用$_2$ · ～と · ～ながら（逆接） · （仮定）ば／たら · （連体）なら、のに、ので · ～て$_3$／連用$_3$ · ～ず（に） · ～ないで · （？）形式名詞で終わる句
2 séquences – 2 unités de discours · （終止）し、けれども、が、から · ～て$_4$／連用$_4$V	**Style complexe dur（複式・硬式）** • Emploi conclusif final (forme autonome)	**C** · （終止）し、けれど、が、から · ～て$_4$／連用$_4$
	Connexions particulières • Style flottant（遊式） • Style fermé（ト式）	**Élément semblable à la phrase（文に似た部分）** · 命令形、已然形で終わるもの · 挿入的な性格のもの · 独立部 · 引用句 · 連体修飾句

TAB. 7.3 – Comparaison des typologies des syntagmes à mot variable

d'insertion d'un élément entre deux mots variables proposé par Garnier et le critère de la possibilité d'ajout au minimum d'un élément de voix proposé par Noda ne sont efficaces que pour certains syntagmes.

Quatre types de syntagmes à mot variable non-propositionnels

En effet, il existe quatre types de syntagmes à mot variable non-propositionnels à reconnaître :

1. mots variables supports ou auxiliaires tels que :
   Ｖ なけれ ば ならない
   (*V - nakere - ba - naranai*)
   (V - [négation] - [condition] - ça ne se fait pas)
   « il faut V »

2. syntagmes à mot variable avec un complément lexicalisés tels que :
   Ｎに 関して
   (*N ni - kanshite*)
   (N [ni] - concerner)
   « concernant N »

3. syntagmes avec le mot variable à une forme neutre non-propositionnels tels que :
   落ち着いて Ｎに 取り組む
   (*ochitsuite - N ni - torikumu*)
   (se calmer - N [ni] - s'attaquer)
   « s'attaquer calmement à N »

4. verbes composés tels que :
   Ｎ1に Ｎ2を 移し 替える
   (*N1 ni - N2 wo - utsushi - kaeru*)
   (N1 [ni] - N2 [wo] - déplacer - changer)
   « transférer N2 vers N1 ».

Pour le troisième type de syntagme, Mikami, Garnier et Teramura considèrent que le partage des compléments est un des critères. Mais, dans la phrase japonaise où tous les éléments peuvent être omis selon le contexte, le partage des compléments est un phénomène très courant et ne peut pas être un test efficace pour distinguer la nature d'un syntagme. Mikami affirme qu'il est important de distinguer les cas où les compléments sont partagés par les deux mots prédicatifs du fait du caractère grammatical de la connexion que produit le premier mot variable, d'une part, et les cas où les compléments sont partagés selon le contexte, de l'autre. Mais, il ne parle pas de la façon de les distinguer. Garnier explique que les deux actions décrites par les deux syntagmes à mot variable non-propositionnels (qui ne constituent pas tout seuls la séquence, selon sa terminologie) relèvent toujours d'un actant unique.

Afin de discriminer de manière automatique ces syntagmes à mot variable démunis du statut de proposition, il nous faudrait déterminer des critères concrets et précis.

7.4 Critères de détermination des syntagmes à mot variable non-propositionnels

Examinons maintenant pour ces quatre types de syntagmes à mot variable non-propositionnels les critères concrets de détermination possibles.

7.4.1 Mots variables supports ou auxiliaires

Ｖ なければ ならない
(*V - nakere - ba - naranai*)
(V - [négation] - [condition] - ça ne se fait pas)
« il faut V »

Ce sont des mots variables suivant, directement ou presque directement par l'intermédiaire d'un connecteur, un autre mot variable de manière à constituer un verbe composé ou une locution prédicative. Ce sont les hyper-séquences de Garnier. Nous pouvons donc employer le critère proposé par cette dernière et définir la règle permettant de constituer préalablement la liste des mots variables supports comme suit :

Règle 1 (mots variables supports) Lorsqu'un mot variable suit toujours un autre mot variable quelconque à une forme donnée et qu'il ne peut prendre aucun complément qui lui soit propre, il est considéré comme mot variable support ou auxiliaire et constitue le prédicat avec le mot variable qui le précède.

7.4.2 Syntagmes à mot variable avec un complément lexicalisés

Ｎに 関して
(*N ni - kanshite*)
(N [ni] - concerner)
« concernant N »

Comme Mikami le signale, il existe beaucoup de constructions comportant un mot variable – essentiellement un verbe – éventuellement suivi d'un connecteur, qui constituent une unité semblable à une particule. Mikami les appelle particules de cas composées. On ne trouve pas de discussion sur ces éléments dans les travaux de Garnier. C'est sans doute dû à la nature de son corpus, des manuels scolaires de l'école primaire. En effet, ce sont souvent des expressions utilisées plutôt à l'écrit et surtout dans un style académique ou journalistique.

La détermination de ces syntagmes à mot variable lexicalisés que nous traitons ici n'est en fait que la partie visible de l'iceberg : elle est liée à la définition d'un autre type d'unité, 連語 (*rengo*) que nous traduisons par « locution ».

Problèmes des *rengo*

La définition de ce terme est assez floue : dans un sens large, il désigne les constructions à plusieurs mots y compris les syntagmes et les propositions ; dans un sens strict, il désigne les unités constituées de plusieurs mots équivalentes à un mot. Dans ce dernier sens, se posent des problèmes de distinction avec le terme « mot composé », 複合語 (*fukugô-go*), désignant les mots constitués de plusieurs mots simples. Aujourd'hui, le terme *rengo* est également utilisé pour la traduction du terme « collocation ».

Nous utilisons dans la présente étude le terme locution pour désigner l'unité constituée de plusieurs mots ayant perdu plus ou moins leurs sens initiaux et ne constituant un sens stable équivalent à celui d'un mot qu'ensemble.

Dans les dictionnaires, figurent généralement les locutions. Étant donné qu'il est impossible d'énumérer toutes les locutions dans le sens large du terme, ce terme est utilisé sans doute dans un sens plus ou moins proche de ce que nous venons de définir. Mais, comme le montre Yazawa (1995), les entrées étiquetées comme des locutions varient extrêmement un dictionnaire à l'autre. De plus, même à l'intérieur d'un dictionnaire, on constate difficilement une cohérence entre les entrées étiquetées comme des locutions, les critères de cet étiquetage ne semblant pas être bien définis.

Critères de détermination des locutions à mot variable

Nous nous occupons dans la présente thèse uniquement des locutions qui nous concernent, les locutions à mot variable, et définissons les critères de leur détermination.

Le critère de Noda qui définit la possibilité d'inclusion d'éléments de voix comme condition minimale requise pour la proposition semble utile au premier abord. Mais, dans la mesure où l'acception des éléments de voix dépend également de la nature lexicale du verbe, il faut utiliser ce test avec prudence.

Nous n'avons pas trouvé jusqu'ici d'autres pistes sur ce sujet dans la littérature. Nous avons rencontré seulement un passage de Mikami que nous avons déjà présenté précédemment (cf. § 7.3.1) qui parle de quelques conditions empêchant la lexicalisation des constructions. Il y dit que les syntagmes verbaux avec le complément en *ga* ne subissent jamais de lexicalisation. C'est une condition nécessaire mais non suffisante. Pour trouver des conditions nécessaires et suffisantes, il faudrait plus d'études avec l'analyse de beaucoup d'exemples. Nous définissons de manière provisoire les règles suivantes pour reconnaître les locutions à mot variable :

Règle 2 (locutions à mot variable) Lorsqu'un mot variable à une forme non autonome – notamment une forme neutre – prend un seul complément (qui n'est pas complément en *ga*) dont le syntagme nominal joue une fonction syntaxique vis-à-vis du prédicat dont le mot variable dépend syntaxiquement, et que le mot variable ne partage pas le complément en *ga*, implicite ou explicite, de ce prédicat, la construction constituée de ce mot variable et la particule introduisant son complément sont considérées comme une locution assimilée à la particule.

Exemple d'application de la règle

Considérons l'exemple suivant :

Ａ社は、企業買収を　めぐり　虚偽の　発表を　した
(*A sha wa - kigyôbaishû wo - meguri - kyogi no - happyô wo - shita*)
(société A [wa] - rachat d'entreprises [wo] - faire le tour - faux - déclaration [wo] - faire [passé])
« La société A a fait de fausses déclarations concernant ses rachats d'entreprises »

Le syntagme « *kigyôbaishû wo meguri* » n'a pas de complément en *ga*, et le complément nominatif du prédicat principal, implicite car thématisé « *A sha (ga)* », n'est pas non plus son complément nominatif. On doit considérer que le verbe « *meguri* » ne possède pas, purement et simplement, de complément nominatif. En revanche, on peut analyser que le SN « *kigyôbaishû* » joue le rôle de complément accessoire vis-à-vis du prédicat principal. On considère donc « *wo meguri* » comme une locution équivalente à une particule.

Cas délicats

Pour certaines expressions, il est difficile de juger si elles ne partagent pas le complément en *ga* avec le prédicat. Dans ce cas, on applique la règle d'extension suivante.

Règle 2' (règle d'extension pour les locutions) Après avoir mis le mot variable à une forme conclusive, si la séquence créée ne peut jamais apparaître toute seule dans un texte quelconque, cette séquence est considérée comme une locution.

Par exemple,

Ａは、Ｂに対して　結果を　報告した
(*A wa - B ni - taishite - kekka wo - hôkoku shita*)
(A [wa] - B [ni] - faire face - résultat [wo] - rapporter [passé])
« A a rapporté le résultat à B »

Le complément nominatif du verbe « *taishite* » peut être « *A (ga)* » qui est le complément nominatif implicite du prédicat principal. On transforme alors le verbe en forme conclusive pour constituer une phrase autonome, i.e. *taisuru*. Mais, « *A ga B ni taisuru* » ne peut pas être considérée comme une phrase indépendante –

du moins dans le japonais moderne. On considère donc que « *ni taishite* » est une locution fonctionnant comme une particule.

Locutions variables

Certaines locutions, ainsi déterminées, sont moins figées que d'autres, conservant alors leurs propres variations.

Par exemple, supposons que l'expression Nに関して (*N ni kanshite*) vérifie les conditions présentées précédemment et qu'elle soit considérée comme une locution. Cette expression conserve sa propre variation et apparaît également sous forme de Nに関し (*N ni kanshi*) ou Nに関する (*N ni kansuru*) dans la structure déterminante. En revanche, l'expression Nについて (*N ni tsuite*) n'a que la forme Nにつき (*N ni tsuki*).

Règle 2" (locutions variables) Les expressions conservant leurs propres variations sont considérées comme des locutions sous l'ensemble de leurs formes.

L'ensemble des règles 2, 2' et 2", primitives et plutôt expérimentales, doit être bien évalué avec les résultats de l'analyse automatique afin d'être amélioré.

7.4.3 Syntagmes avec le mot variable à une forme neutre non-propositionnels et verbes composés

落ち着いて Nに 取り組む
(*ochitsuite - N ni - torikumu*)
(se calmer - N [ni] - s'attaquer)
« s'attaquer calmement à N »
N1に N2を 移し 替える
(*N1 ni - N2 wo - utsushi - kaeru*)
(N1 [ni] - N2 [wo] - déplacer - changer)
« transférer N2 vers N1 »

Contrairement aux deux types de syntagmes présentés précédemment, pour ceux que nous traitons ici, il est beaucoup plus difficile voire impossible de constituer des listes préalables car les possibilités sont beaucoup plus importantes. Il est donc nécessaire, pour ces syntagmes, de définir des règles permettant de les reconnaître de manière dynamique lors de l'analyse.

Examinons d'abord le cas des syntagmes avec le mot variable à une forme neutre. Pour la détermination de la nature des syntagmes à mot variable terminés par une forme neutre, il existe un critère proposé par Mikami et utilisé par Teramura : la présence du complément en *ga*. Nous adoptons une règle un peu plus large : la présence d'au moins un complément. En effet, aussi bien dans les travaux de Minami que dans ceux de Garnier, les emplois de forme neutre se distinguent en plusieurs emplois et chaque emploi appartient à une catégorie fort différente.

Ne connaissant pas pour l'instant les critères formels permettant de distinguer ces différents emplois, nous adoptons une règle plus large mais qui semble efficace.

Il faut tenir compte des deux cas :

1. V_1 (forme neutre) + Complément(s) + V_2
2. Complément(s) + V_1 (forme neutre) + V_2

Dans le premier cas, le mot variable V_1 n'a aucun complément, tandis que dans le second, c'est le mot variable V_2 qui apparaît sans aucun complément.

La règle pour le second cas permet également de reconnaître les verbes composés dans lesquels aucun complément n'apparaît entre les verbes simples constituants.

Par ailleurs, la règle pour le premier cas permet de détecter ceux qui sont lexicalisés du type, 続いて (*tsuzuite*, puis) ou 併せて (*awasete*, en même temps) – qui fonctionnent comme des mots de liaison – lorsqu'ils sont traités comme des verbes par l'analyseur morphologique.

Règle 3 (pour l'identification dynamique) Lorsqu'un syntagme terminé par un mot variable à une forme neutre ne comprend aucun complément, il est considéré comme un syntagme non-propositionnel dépendant du prédicat apparaissant à une position postérieure. Lorsqu'un syntagme terminé par un mot variable conclusif ne comprend aucun complément et qu'il est précédé directement par un mot variable à une forme neutre, il est considéré comme constituant un mot variable composé avec celui qui le précède directement.

7.5 Nos définitions des unités : proposition et sous-phrase

Nous définissons maintenant la proposition ainsi que quelques autres unités en tenant compte de tout ce que nous avons étudié jusqu'ici. Mais, avant de présenter nos définitions, nous allons passer brièvement en revue les problèmes liés à la notion de proposition principale.

Discussion sur la notion de proposition principale

Après avoir défini la phrase complexe comme une phrase comprenant plusieurs prédicats qui constituent les propositions, Masuoka & Takubo (1992) distinguent les propositions en deux grands types : proposition principale et proposition connectée. La proposition principale est celle qui régit, en se mettant en fin de phrase, l'ensemble de la phrase. La proposition connectée est celle qui n'est pas une proposition principale mais qui lui est reliée par une certaine relation.

Noda récuse tout d'abord la définition usuelle de la phrase complexe comme construction réalisée par assemblage de plus de deux propositions ayant une forme semblable à la phrase. Il considère comme phrase complexe la phrase dans laquelle une partie de la phrase simple devient une proposition lors d'un développement. Ainsi, il rejette la notion de proposition principale et appelle la struc-

ture constituée autour du prédicat principal, phrase principale. Les constructions phrastiques jouant le rôle d'élément dans une phrase principale sont appelées propositions.

Comme dans les études linguistiques du français, du fait de l'inexactitude de l'analyse classique totalement linéaire, nous évitons le terme de « proposition principale » et utilisons le terme de « proposition racine ».

Définitions

Proposition Nous appelons proposition toute construction à mot variable satisfaisant la condition de capacité régissant des compléments, à l'exception des constructions terminées par une forme neutre sans aucun complément. Nous appelons par ailleurs **prédicat** uniquement les mots variables capables de constituer autour d'eux une proposition.

Proposition racine Nous appelons les propositions se terminant par un mot variable à une forme conclusive qui ne dépendent syntaxiquement d'aucun élément de phrase, propositions racines.

Proposition subordonnée Les propositions subordonnées sont des propositions suivies d'un connecteur (particules conjonctives, substantifs formels) ou terminées par une forme connective.

Sous-phrase Par ailleurs, nous appelons sous-phrases les constructions – dépendant du prédicat principal de la phrase – constituées d'une proposition et d'un thème ainsi qu'éventuellement d'éléments externes précédant le thème et qui n'entrent pas dans la proposition.

Sous-structure subordonnée Nous appelons sous-structures subordonnées (ou plus simplement subordonnées) l'ensemble des propositions subordonnées et des sous-phrases. Les propositions subordonnées peuvent également être appelées subordonnées, tant que la confusion avec le sens plus large ne gêne pas l'interprétation.

Connecteur souple et connecteur dur Suivant la typologie de Mikami, nous posons comme hypothèse que les constructions produisant une connexion de type complexe souple peuvent comporter seulement des propositions, mais pas de sous-phrase. En revanche, les constructions produisant une connexion de type complexe dur peuvent comporter des sous-phrases. Nous appelons le connecteur susceptible de constituer les constructions produisant une connexion de type complexe souple, connecteur souple, et celui susceptible de constituer les constructions produisant une connexion de type complexe dur, connecteur dur.

7.6 Second problème : catégorisation imprécise des éléments suivant une forme conclusive du mot variable

7.6.1 Description du problème

Teramura (1982a) signale que la sous-structure dans une phrase peut être constituée par des mots variables, non seulement à une forme connective, mais aussi à une forme conclusive (voir aussi § 7.9.1). Il énumère quatre cas où le mot variable à une forme conclusive constitue le prédicat non principal de la phrase :

1. lorsqu'il constitue une subordonnée déterminant le substantif qu'il précède ;
2. lorsqu'il constitue une subordonnée déterminant un substantif formel et constituant ensemble une unité équivalente à un substantif ;
3. lorsqu'il constitue, en étant suivi par une particule conjonctive, une subordonnée conjonctive ;
4. lorsqu'il constitue, en étant suivi par la particule *to*, une subordonnée de citation.

Outre ces quatre possibilités, il existe en réalité une cinquième possibilité de construction des subordonnées conclusives : constitution avec une particule adverbiale. L'absence de description de cette dernière possibilité peut être un simple oubli, mais elle pourrait aussi être due à la définition généralement floue des catégories concernées : particules conjonctives, particules adverbiales, particule *ka*, substantifs formels, ou encore auxiliaires.

Comme nous l'avons vu dans la section 5.3.2, ces catégories constituent une zone totalement floue et désordonnée. Mais, le mot prédicatif étant incapable d'assurer lui-même la connexion par sa forme, dans le cas des subordonnées terminées par un mot prédicatif à une forme autonome conclusive, c'est l'élément suivant la forme autonome qui réalise le lien syntaxique. Aussi considérons-nous ces éléments suivant la forme autonome comme des connecteurs syntaxiques.

Afin de bien décrire tous les types de propositions japonaises, l'analyse correcte de ces connecteurs est indispensable. Nous allons donc maintenant essayer de réorganiser cette zone désordonnée afin de bien définir chacune de ces catégories.

7.6.2 Connecteurs syntaxiques des propositions

Comme nous venons de le voir, nous considérons comme des connecteurs syntaxiques les éléments suivant la forme autonome, qui assurent la connexion syntaxique des subordonnées terminées par un mot prédicatif à une forme autonome conclusive (subordonnées conclusives ci-après).

Mais les subordonnées terminées par un mot prédicatif à une forme non conclusive peuvent être elles-aussi suivies d'un autre élément – notamment les

particules de mise en relief. Ces éléments suivant une forme connective du mot prédicatif doivent-ils être considérés eux-aussi comme des connecteurs ?

Éléments suivant le mot prédicatif à une forme connective

Dans la mesure où, dans le cas des subordonnées terminées par un mot prédicatif à une forme connective, le mot prédicatif est à une forme lui permettant d'assurer lui-même la connexion, l'élément suivant le mot prédicatif est sans doute un élément non chargé de la réalisation du lien, c'est-à-dire un élément non obligatoire sur le plan syntaxique. Nous les considérons donc non pas comme des connecteurs syntaxiques mais comme des marqueurs du lien sémantique.

Différents connecteurs suivant le mot prédicatif à une forme autonome

Rappelons encore une fois les cinq possibilités de constitution des subordonnées terminées par un mot prédicatif à une forme autonome :

1. constitution d'une subordonnée déterminante ;
2. constitution d'une subordonnée déterminant un substantif formel (subordonnée assimilée substantive) ;
3. constitution d'une subordonnée conjonctive avec une particule conjonctive ;
4. constitution d'une subordonnée avec une particule adverbiale ;
5. constitution d'une subordonnée de citation avec la particule *to*.

Les subordonnées résultant de la première possibilité sont les seules subordonnées conclusives sans connecteur. Dans les autres types, la connexion est réalisée par le connecteur constitué respectivement par le substantif formel, la particule conjonctive, la particule adverbiale et la particule *to*.

Le problème que nous rencontrons ici est, encore une fois, que la définition des catégories concernant les mots constituant les trois premiers connecteurs étant assez floue, nous ne savons finalement pas déterminer exactement ces connecteurs.

Afin d'élucider ces notions obscures, nous allons maintenant passer en revue les travaux antérieurs sur la catégorisation de ces mots.

7.7 État de l'art des travaux sur la catégorisation des mots suivant une forme autonome

7.7.1 Les mots agglutinants de Sakuma

Sakuma (1940b) a regroupé les mots suivant le mot prédicatif à une forme autonome sous le nom de *kyûchakugo*, mots agglutinants, après avoir découvert leur point commun : être peu voire non autonome et capable de constituer un élément autonome et complet sémantiquement une fois déterminé.

Mais, Sakuma n'a pas défini de catégorie spécifique à part : *kyûchakugo* restait la simple appellation d'un ensemble de mots de catégories différentes regroupés selon une certaine caractéristique commune qu'ils possédaient.

Mikami définit en revanche une catégorie spécifique couvrant l'ensemble de ces mots qu'il appelle 準詞 (*junshi*, mot assimilé). Les *junshi* se distinguent eux-même en deux types : mots assimilés mots prédicatifs (準用詞, *jun'yô-shi*) et mots assimilés substantifs (準体詞, *juntai-shi*)

7.7.2 Les études comparatives de Teramura

Sans essayer de réorganiser les catégories existantes ni d'en définir une nouvelle, Teramura montre tout simplement la différence et la continuité de ces catégories. D'après l'auteur, cette continuité entre les substantifs constituant la base d'une subordonnée déterminante d'un côté et les particules conjonctives de l'autre, est formée par les substantifs constituant la base d'une subordonnée déterminante, plus ou moins lexicalisés, ayant ainsi un caractère formel de niveau différent.

Il propose plusieurs types de tests pour évaluer le caractère substantif de chaque mot. Il en existe deux types : tests pour la possibilité qu'un mot accepte la détermination d'un mot donné (déterminant démonstratif, qualificatif, juxtaposition d'un substantif, etc.) et tests pour la possibilité qu'un mot assume une fonction donnée (cas *ga*, cas *wo*, cas *no*, cas *kara*, cas *ni*, cas zéro).

Le tableau 7.4 (voir page suivante), repris de Teramura (1978), présente le résultat de ces tests réalisés par Teramura.

7.7.3 La réorganisation complète proposée par Okutsu et Numata

Okutsu propose dans l'introduction de Okutsu et al. (1986) une réorganisation complète des catégories existantes, notamment les différents types de particules. Il renonce à la catégorie de particule et les particules sont catégorisées dans différentes classes nouvellement définies, telles que les substantifs formels, les adverbes formels ou les *toritate-shi* (mots de mise en relief).

Substantifs formels et adverbes formels d'Okutsu

La catégorie des substantifs formels d'Okutsu est beaucoup plus limitée que celle définie antérieurement, alors que celle des adverbes formels (Okutsu, 1986) est relativement large incluant une grande partie des particules adverbiales et conjonctives. En effet, Okutsu détermine la catégorie équivalente selon les fonctions que peuvent jouer les syntagmes constitués par ces mots. Ainsi, beaucoup de substantifs formels et de particules adverbiales sont classés dans la catégorie d'adverbes formels par Okutsu.

連体修飾のシンタクスと意味　303

表　1

	承								接					
	①コレガ—ダ	②[名]ノ—	③コノ、ソノ—	④[形]イ—	⑤ソンナ、コンナ—	⑥[名]—	⑦コレ、ソレ—		⑧—ガ	⑨—ヲ	⑩—ノ	⑪—カラ	⑫—ニ	⑬—φ
キダロ	×	○	○	○	○	×	×		○	○	○	○	○	○
来	×	○	○	○	○	×	×		○	○	○	○	○	○
前	×	○	○	○	○	×	×		○	○	○	○	○	○
後	×	△	×	×	×	○	○		△	×	○	×	○	○
ラ	×	△	△	×	×	○	○		○	○	○	○	○	×
デ	×	×	×	×	×	○	○		○	○	○	×	○	×
ビ	×	×	×	×	×	○	○		○	○	○	×	○	○
毎	×	×	○	×	△	×	×		×	×	△	×	○	○
合	×	×	○	○	○	×	×		×	○	△	×	○	×
的	×	○	○	○	○	×	×		○	○	○	×	○	○
メ	×	×	○	○	○	×	×		△	×	×	×	×	×
イ	×	×	○	○	○	×	×		×	×	×	×	×	×
由	×	×	○	○	○	×	×		○	×	○	×	×	○
ラ	×	○	×	○	○	×	×		×	○	×	×	×	×
エ	×	×	○	○	○	×	×		×	×	○	×	×	×
ケ	×	×	○	○	○	×	×		×	×	○	×	×	×
因	×	×	○	○	○	×	×		×	×	○	×	×	×
果	×	×	○	○	○	×	×		×	×	○	×	×	×
クエリロ	×	×	○	○	○	×	×		×	×	○	×	×	×
度	×	×	○	○	○	×	×		×	×	○	×	×	×
ド	×	×	○	○	○	×	×		×	×	○	×	×	×
度	×	×	○	○	○	×	×		×	×	○	×	×	×
ケリイリリマセリリ子ウ	×	×	○	○	○	×	×		×	×	○	×	×	×
末	×	×	○	○	○	×	×		×	×	○	×	×	×
ゲ	×	×	○	○	○	×	×		×	×	○	×	×	×
マコ	×	×	○	○	○	×	×		×	×	○	×	×	×
ギラ	×	×	○	○	○	×	×		×	×	○	×	×	×
オ	×	×	○	○	○	×	×		×	×	○	×	×	×
イ	×	×	○	○	○	×	×		×	×	○	×	×	×
トアコ以以カマタ度場目タセ理カユワ原結	×	×	○	○	○	×	×		×	×	○	×	×	×
アウアト程ホ限ダカクキナマクトワ様ヨ	×	×	○	○	○	×	×		×	×	○	×	×	×

TAB. 7.4 – Tableau comparatif réalisé par Teramura (1978)

Mots de mise en relief de Numata

Les particules adverbiales et les *kakari-joshi* qui ne sont pas considérées comme des adverbes et des substantifs formels sont catégorisées pour la plupart dans la classe des とりたて詞 (*toritate-shi*), mots de mise en relief. Numata (1986) définit les mots de mise en relief comme des unités mettant en relief différents éléments de la phrase (appelés les internes) tout en représentant les relations logiques qu'entretiennent ces éléments avec d'autres éléments du même paradigme (appelés les externes). Ainsi, certaines particules adverbiales et d'autres particules *kakari-joshi* sont classées dans cette nouvelle catégorie.

Les mots regroupés ainsi sont caractérisés par quatre particularités syntaxiques :

1. grandes possibilités de positionnement ;
2. caractère facultatif ;
3. possibilité d'apparition dans les subordonnées déterminantes ;
4. caractère non substantif.

et ces quatre caractères sont considérés comme les conditions nécessaires et suffisantes pour la détermination des mots de mise en relief.

7.7.4 Analyse critique

Malgré tout l'intérêt que présentent ces travaux, la tentative de réorganisation des catégories d'Okutsu et la définition et les études des mots de mise en relief, *toritate-shi*, de Numata ne conviennent pas à nos travaux d'analyse automatique des phrases.

Problèmes de la définition des adverbes formels d'Okutsu

Comme nous l'avons vu, Okutsu détermine la catégorie équivalente selon les fonctions que peuvent jouer les syntagmes constitués par le mot en question. Mais, cette définition annulant la frontière entre les analyses en termes de catégories et en termes de fonctions, n'est pas la solution la plus économique.

Par exemple, Okutsu considère ため (*tame*) comme un adverbe formel, du fait notamment que ce mot, déterminé par une subordonnée, constitue une proposition de cause et de but. Mais la séquence terminée par *tame* peut former, en étant suivi de la particule *no*, le constituant déterminant. Si on définit le mot *tame* comme un adverbe formel, on doit également définir un autre substantif formel *tame* ou un qualificatif formel *tame no*.

Toutefois, il nous semble plus économique et même cohérent de considérer ce mot comme un substantif formel ayant un emploi adverbial, emploi possédé par beaucoup de substantifs japonais.

Comme l'affirme Le Goffic (1993a), nous considérons que les « étiquetage en termes de fonctions et étiquetage en termes de catégories doivent être parallèles et distincts » et nous ne pouvons pas approuver la position d'Okutsu.

Inconvénients de la définition des mots de mise en relief de Numata

La catégorie des mots de mise en relief définie par critères sémantiques comme décrit précédemment, regroupe des unités ayant des comportements syntaxiques si divers que Numata considère comme une des caractéristiques leur liberté de positionnement. Mais ce n'est que le résultat de l'utilisation de critères non syntaxiques mais sémantiques. Dans le cadre de nos travaux, nous préférons accorder plus d'importance aux aspects syntaxiques et distinguer au moins les unités capables de suivre une forme autonome et d'assurer la connexion syntaxique, des autres qui ne sont pas en mesure de remplir ces deux conditions.

Numata utilise tout de même également un critère syntaxique : le caractère facultatif. Mais ce critère peut remettre en cause la justesse de cette catégorisation. Par exemple, dans la phrase :

可能性が　ある　かどうか　すら　も　分からない。
(*kanôsei ga - aru - ka dôka - sura - mo - wakaranai*)
(possibilité [*ga*] - exister - [interrogation totale indirecte] - même - [*mo*] - savoir [négation])
« (Je) ne sais même pas s'il y a des possibilités. »

les mots *sura* et *mo*, classés dans les mots de mise en relief, sont tous les deux effectivement facultatifs syntaxiquement.

En revanche, dans la phrase (reprise de Numata (1986)) :

太郎は　働く　だけ　を　生き甲斐　としている。
(*tarô wa - hataraku- dake - wo - ikigai - to shiteiru*)
(Taro [*wa*] - travailler - [*dake*] - [*wo*] - raison d'être - considérer comme)
« Taro considère comme sa seule raison d'être le fait de travailler. »

la suppression de *dake* rend la phrase agrammaticale. Pour respecter le critère de caractère facultatif, il est possible de considérer que le mot *dake* suivant une forme autonome est un substantif formel et distinct du *dake* catégorisé dans les mots de mise en relief. Mais dans ce cas, on ne peut pas s'empêcher de se demander l'intérêt de distinguer deux unités ayant une forme et un sens identiques, pour n'en regrouper qu'une avec d'autres mots similaires.

Une autre caractéristique syntaxique des mots de mise en relief est leur caractère non substantif. Les mots de mise en relief comportent en fait des particules adverbiales considérées dans beaucoup de grammaires comme des mots assimilés substantifs. Mais Numata ne reconnaît pas le caractère substantif de ces particules. Afin de justifier sa position, elle utilise un test consistant à transformer la phrase en subordonnée déterminante de manière à faire du syntagme contenant la particule en question la base déterminée[7]. Par exemple, la phrase d'exemple précédente est transformée en :

*太郎が　生き甲斐　としている　働く　だけ
(*tarô ga - ikigai - to shiteiru - hataraku- dake*)
(Taro [*ga*] - raison d'être - considérer comme - travailler - [*dake*])

[7] Ce test est proposé par Okutsu (1974).

et comme la séquence résultant de cette transformation est agrammaticale, l'auteur considère que le mot *dake* n'a pas de caractère substantif.

Mais ce test nous laisse sceptiques quant à sa justesse. En effet, avec ce test, nous ne pouvons pas non plus reconnaître le caractère substantif du substantif formel *no*, dont personne ne nie la capacité nominalisatrice.

Le syntagme nominalisé par *no* (souligné dans l'exemple) :

太郎は　働く　の　を　生き甲斐　としている。
(*tarô wa - hataraku- no - wo - ikigai - to shiteiru*)
(Taro [*wa*] - travailler - [*no*] - [*wo*] - raison d'être - considérer comme)
« Taro considère comme sa raison d'être le fait de travailler. »

ne peut pas constituer la base d'une subordonnée déterminante, tout comme le syntagme terminé par *dake* :

*太郎が　生き甲斐　としている　働く　の
(*tarô ga - ikigai - to shiteiru - hataraku- no*)
(Taro [*ga*] - raison d'être - considérer comme - travailler - [*no*])

Le plus grand défaut de cette méthode d'évaluation du caractère substantif est sans doute qu'elle ne tient pas compte de l'existence des deux aspects différents du « caractère substantif ». Ce caractère doit, en effet, être évalué, comme Teramura l'a fait avec ses différents tests, sur les deux capacités distinctes : capacité à régir d'autres éléments (c'est-à-dire le nombre de manières de qualifier le mot en question) et capacité à être régi (c'est-à-dire le nombre de fonctions que le mot en question peut assumer).

Nous essayons maintenant de réaliser une catégorisation détaillée de l'ensemble des mots agglutinants en tenant bien compte de ces deux aspects différents du caractère substantif.

7.8 Notre catégorisation et ses critères

7.8.1 Méthodologie

Nous nous basons principalement sur le résultat de différents tests réalisés par Teramura et présenté dans le tableau 7.4 page 286. Pour les mots qui ne figurent pas dans le tableau de Teramura ou dont l'analyse nous semblait contestable, nous avons réalisé nos propres tests avec le corpus Shincho (représentant 66 899 phrases, cf. Liste des corpus utilisés p. 550 et suivantes) et de plusieurs articles du corpus Yomiuri (3 237 phrases) à l'aide d'un concordancier élémentaire[8] que nous avons développé spécifiquement (cf. fig. 7.5 (voir page suivante)).

[8] Il indique les occurrences d'une séquence saisie par l'utilisateur, avec indication de ses contextes gauche et droit de 20 caractères.

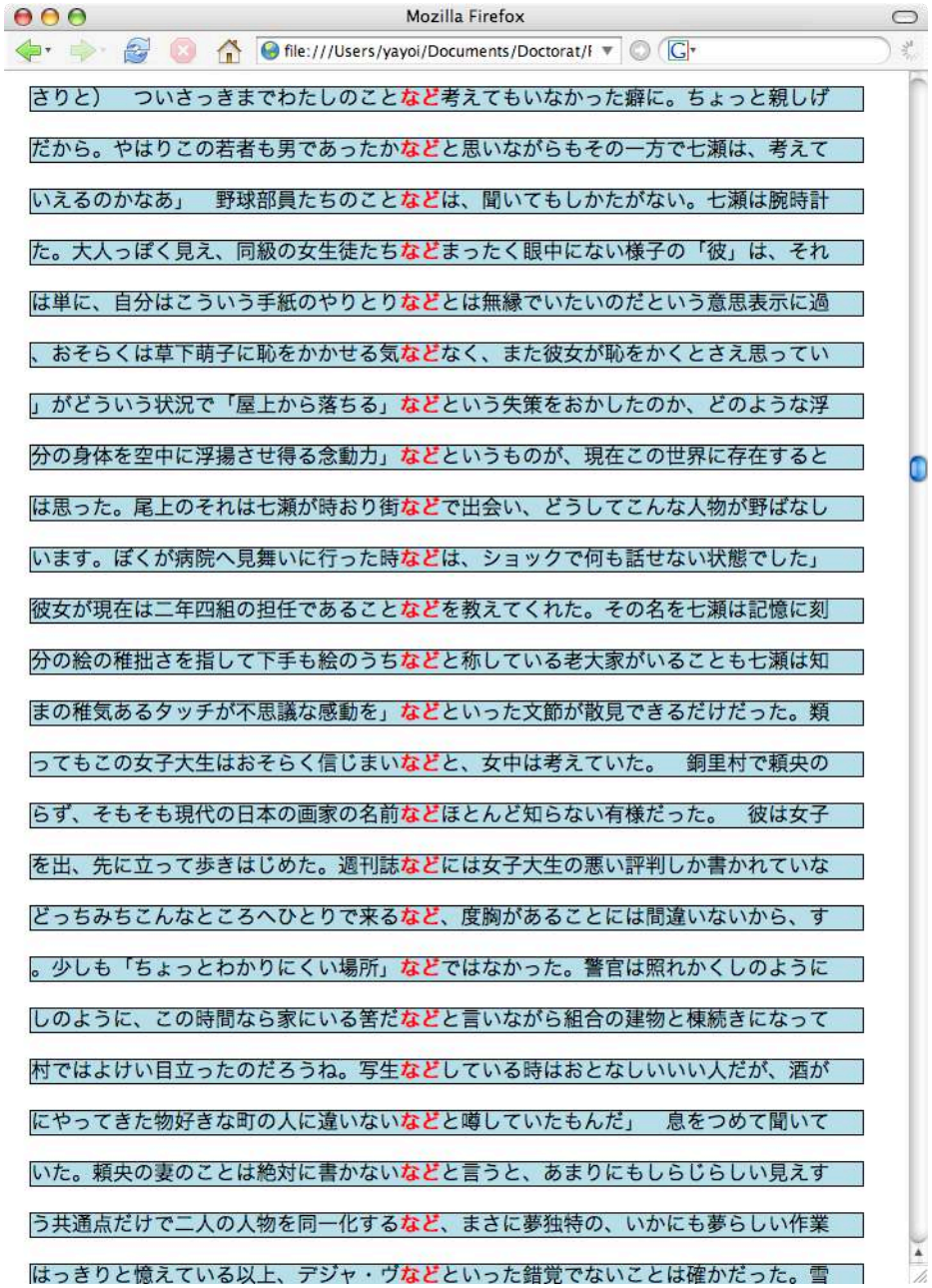


FIG. 7.5 – Résultat d'analyse par notre concordancier

7.8.2 Définition des connecteurs agglutinants

Parmi les mots qui peuvent suivre une forme autonome, nous considérons comme des connecteurs agglutinants, ceux qui ne sont ni substantifs autonomes, ni auxiliaires, ni particules conjonctives.

Nous éliminons d'abord avec des critères concrets les deux premières unités n'appartenant pas aux connecteurs agglutinants. Puis, nous examinons le reste des mots pour les catégoriser finalement selon leur caractère substantif en quatre classes – l'une regroupe les particules conjonctives, et les trois autres correspondent à des sous-catégories des connecteurs agglutinants.

La catégorisation se déroule comme suit :

- étape de catégorisation 0 : exclusion des éléments ne suivant pas une forme autonome ;
- étape de catégorisation 1 : exclusion des éléments autonomes et définition des mots agglutinants ;
- étape de catégorisation 2 : exclusion des auxiliaires ;
- étape de catégorisation 3 : classement en quatre catégories.
- étape de catégorisation 4 : définition des connecteurs agglutinants.

Catégorisation 0 : exclusion des éléments ne suivant pas une forme autonome

Nous distinguons tout d'abord les mots qui peuvent suivre une forme autonome de ceux qui ne le peuvent pas. Suite à ce test, les particules *koso, sae, shika, sura, wa, mo, zo,* définies comme *kakari-joshi* dans le dictionnaire ipadic[9], sont catégorisées quasiment toutes comme des éléments ne suivant pas une forme autonome.

Nous avons tout de même constaté quelques cas d'exception.

sae apparaît deux fois après la forme autonome comme « *omoidasu sae* » (se souvenir - sae) dans les œuvres de Mishima et de Tanizaki. Mais, ces exemples très peu fréquents (2 occurrences contre 33 pour la structure comportant un nominalisateur) ne peuvent pas être un motif de reconnaissance de l'emploi *sae* suivant directement la forme autonome. Nous n'avons donc pas retenu la particule *sae* comme un élément suivant une forme autonome.

Nous avons également constaté des emplois de *shika* (ne ... que) suivant une forme autonome. Mais ces occurrences précèdent presque toujours directement le prédicat principal *nai* (ne pas exister), ou avec au plus le syntagme nominatif, *hôhô ga/wa* ou *te ga/wa* (moyen), et/ou le syntagme locatif, *hoka ni* (ailleurs), telles que :

のぼりつめていく　しか　手は　ない。
(*nobori tsumete iku - shika - te wa - nai*)
(monter continuellement - [*shika*] - moyen [*wa*] - ne pas exister)
« (Je) n'ai aucun moyen (autre que) de monter continuellement. »

[9]Dictionnaire électronique pour le TAL utilisé par l'analyseur morphologique ChaSen que nous utilisons dans notre réalisation informatique.

Cette construction avec *shika* étant assez figée, nous ne traitons pas dans la présente étude cette particule comme un élément assurant le lien syntaxique, position contestable nécessitant peut-être des réexamens futurs.

Catégorisation 1 : exclusion des éléments autonomes et définition des mots agglutinants

Nous utilisons un test de Teramura : si le mot *N* peut constituer la phrase « *kore ga N desu* » (c'est N), le mot *N* est considéré comme un substantif autonome.

Les autres mots (c'est-à-dire les mots capables de suivre une forme autonome mais qui ne sont pas autonomes) sont des *kyûchakugo*, mots agglutinants.

Catégorisation 2 : exclusion des auxiliaires

Les mots agglutinants qui peuvent constituer le prédicat principal éventuellement avec la copule et qui ne peuvent constituer un élément de phrase (y compris une proposition subordonnée) qu'à l'aide d'une particule conjonctive ou par la variation de leur forme, sont considérés comme des auxiliaires.

Catégorisation 3 : classement en quatre catégories

La classement est réalisé suite à l'évaluation du caractère substantif de chaque mot sur les deux aspects dont nous avons parlé précédemment.

Les mots sont examinés du point de vue, d'une part de leur capacité à régir d'autres éléments (c'est-à-dire le nombre de manières de qualifier le mot en question), et d'autre part, de leur capacité à être régi (c'est-à-dire le nombre de fonctions que le mot en question peut assumer).

Pour le premier aspect, nous recourons essentiellement au test vérifiant si le mot *N* en question est capable de constituer le syntagme nominal « (substantif) *no N* ». Par exemple, dans la mesure où il est tout à fait possible de dire « *kodomo no toki* » (enfant - [*no*] - quand ; temps), on considère que le mot *toki* a un caractère substantif élevé du point de vue de la capacité à être déterminé. En revanche, le mot *dake* pouvant être déterminé non pas avec la particule *no* « **kodomo no dake* » (enfant - [*no*] - seulement) mais seulement par juxtaposition « *kodomo dake* » (enfant - seulement), on considère qu'il a un caractère substantif faible du point de vue de la capacité à être déterminé.

Pour le second aspect, nous évaluons si le mot *N* en question est capable ou non de constituer – une fois déterminé – les constituants de la phrase du cas *ga*, cas *wo*, cas *no*, cas *kara*, cas *ni* et cas zéro. Par exemple, le mot *dake* pouvant constituer tous les constituants sauf celui du cas *kara* (d'après l'analyse de Teramura), on considère qu'il a un caractère substantif élevé du point de vue de la capacité à assumer des fonctions. En revanche, le mot *kiri* (depuis) ne pouvant constituer que le syntagme nu (cas zéro), on considère qu'il a un caractère substantif très faible du point de vue de la capacité à assumer des fonctions.

En combinant ces deux analyses, nous pouvons établir quatre classes comme présenté dans le tableau 7.6.

capacité à assumer des fonctions / capacité à être déterminé		caractère substantif : fort	caractère substantif : faible
caractère substantif	fort	***substantif formel***	***adverbe / qualificatif substantifs***
caractère substantif	faible	***particule nominalisatrice***	***particule conjonctive***

TAB. 7.6 – Catégorisation des éléments suivant une forme autonome du mot variable

Nous appelons substantifs formels les mots ayant un caractère substantif élevé pour les deux aspects tels que *toki* (quand ; temps) ou *mama* (tel quel).

Sont appelés particules nominalisatrices les mots assurant plusieurs types de fonctions comme les substantifs, mais pour lesquels les types de déterminants qui peuvent les déterminer sont limités, tels que *dake* (seulement), *ka* ([marqueur d'interrogation]) ou *no* ([nominalisateur]).

Les mots acceptant plusieurs types de déterminants mais qui sont figés quant au choix des fonctions qu'ils assurent, se distinguent en deux types : qualificatifs substantifs tels que *yô* (manière) et adverbes substantifs comme *wari (ni)* (relativement) ou *kuse (ni)* (malgré)[10].

Enfin, les mots n'ayant qu'un faible caractère substantif sur les deux aspects sont des purs connecteurs de propositions : des particules conjonctives.

Catégorisation 4 : définition des connecteurs agglutinants

Nous considérons comme connecteurs agglutinants les substantifs formels, les particules nominalisatrices et les adverbes et les qualificatifs substantifs (sur fond coloré dans le tableau 7.6), lorsqu'ils réalisent la nominalisation de la proposition qui les précède. La connexion syntaxique à proprement parler est généralement assurée par la particule de cas qui les suit (ou par son absence, ou encore par la terminaison) mais nous considérons la construction constituée de ces mots et éventuellement de la particule comme des connecteurs reliant la proposition qu'ils introduisent au prédicat postérieur.

Les particules nominalisatrices apparaissant également après différentes unités autres que les formes autonomes du mot variable (sauf la particule *ka*) sont en fait des particules usuellement classées dans les particules de mise en relief.

[10] Ces adverbes substantifs sont si figés qu'ils ne peuvent pas non plus constituer le prédicat principal de la phrase contrairement aux autres mots agglutinants qui peuvent, eux, former le prédicat de la phrase en étant suivi de la copule.

Les particules nominalisatrices peuvent être considérées comme constituant une sous-catégorie des particules de mise en relief.

7.8.3 Résultat général de notre catégorisation

Nous avons proposé ici une manière de tracer des frontières entre les unités, frontières assez floues jusqu'aujourd'hui, avec des critères tout à fait concrets.

Sur un plan pratique, la catégorisation a été réalisée sur la base du résultat des analyses réalisées par Teramura, que nous avons complétées par l'analyse de quelques mots supplémentaires.

Cette première expérimentation nous a fourni un résultat qui confirme dans les grandes lignes la justesse de la catégorisation traditionnelle.

Cette analyse devrait cependant être réalisée, dans des travaux futurs, de manière plus systématique et pour un plus grand nombre de mots par l'analyse d'un corpus de taille importante.

7.8.4 Caractéristiques et problèmes des *kyûchakugo*

Fonction des mots agglutinants assimilés à des auxiliaires

Outre des connecteurs semblables aux particules conjonctives, ces mots agglutinants peuvent également constituer, comme le montre Teramura (1978), des prédicats, devenant ainsi des éléments semblables aux auxiliaires, *jodôshi*.

Dans la grammaire de Masuoka & Takubo (1992), il existe une sous-catégorie des auxiliaires dite « contenant un substantif formel ».

Nous considérons les mots agglutinants suivis directement de la copule non comme des connecteurs, mais comme constituant avec la copule un auxiliaire.

Problèmes liés aux mots agglutinants : polysémie

Teramura (1978) signale que ces mots n'ont pas toujours le même caractère formel et qu'ils expriment un sens plus ou moins stable et autonome selon le contexte. La polysémie pose toujours des problèmes difficiles lors de l'analyse automatique. Dans le cadre de la présente thèse, nous ne prétendons pas proposer de solution à ce problème de polysémie des mots agglutinants, mais nous observerons l'influence de cette polysémie afin d'obtenir des données bénéfiques à nos futurs travaux.

7.9 État de l'art sur les typologies des subordonnées

Nous allons maintenant passer en revue les typologies des subordonnées proposées par différents linguistes.

7.9.1 Typologie selon la forme de connexion

Teramura (1982a) distingue, comme nous l'avons abordé dans la section 7.6.1, les subordonnées constituées du mot variable à une forme connective de celles se terminant par une forme conclusive et il définit d'abord quatre types de subordonnées terminées par un mot prédicatif à une forme conclusive :

1. subordonnée déterminant le substantif qui la suit, appelée **subordonnée déterminante** ;
2. subordonnée déterminant un substantif formel (tel que *no*, *koto* ou *ka*) et constituant ensemble une unité équivalente à un substantif, qu'il appelle **subordonnée assimilée à un substantif** (et qu'il considère comme un type de subordonnée déterminante) ;
3. subordonnée terminée par une particule conjonctive, appelée **subordonnée conjonctive** ;
4. subordonnée terminée par la particule *to*, appelée **subordonnée de citation**.

En tenant compte de ces quatre types de subordonnées terminées par une forme conclusive, ainsi que des subordonnées terminées par une forme connective, il définit finalement cinq types de phrases complexes selon la subordonnée qu'elles contiennent :

- Type 1 :
  phrase contenant une subordonnée coordonnée terminée par un verbe à la forme neutre, ayant ses propres compléments différents de ceux du prédicat principal ;
- Type 2 :
  1. phrase contenant une subordonnée terminée par un mot prédicatif à une forme de condition ;
  2. phrase contenant une subordonnée terminée par un mot prédicatif à une forme autonome suivie de la particule *to* produisant l'expression de condition ;
- Type 3 :
  phrase contenant une subordonnée déterminante (y compris subordonnée assimilée à un substantif) ;
- Type 4 :
  phrase contenant une subordonnée conjonctive ;
- Type 5 :
  phrase contenant une subordonnée de citation.

Plus le chiffre est grand, plus le caractère phrasogénérateur de la proposition subordonnée est élevé, c'est-à-dire plus elle est proche de la phrase autonome.

7.9.2 Typologies selon les fonctions des subordonnées dans la phrase

Nous constatons des typologies selon les fonctions des subordonnées dans la phrase dans les travaux de Noda et dans ceux de Masuoka et Takubo. Les travaux de Noda (2002) étant très brefs et ne faisant pas ressortir de point de vue particulier, nous ne présentons ici que les travaux de ces derniers, la grammaire de Masuoka & Takubo (1992) et les deux ouvrages plus récents de Masuoka (1997, 2002).

Masuoka & Takubo (1992) catégorisent l'ensemble des propositions comme suit :

1. Principale ;
2. Connectée :
    a) Proposition subordonnée ;
        i. Proposition complétive ;
        ii. Proposition adverbiale ;
        iii. Proposition déterminante ;
    b) Proposition coordonnée.

Les propositions connectées (cf. § 7.5) sont catégorisées en deux classes : subordonnée et coordonnée. La première a elle-même trois sous-classes : propositions adverbiales, propositions complétives et propositions déterminantes (i.e. relatives).

Cette catégorisation est modifiée dans les travaux postérieurs de Masuoka (1997, 2002).

Dans ces nouveaux travaux, les propositions sont à nouveau catégorisées comme suit :

1. Principale ;
2. Subordonnée :
    a) substantive ;
    b) adverbiale ;
    c) déterminante ;
    d) coordonnée ;
    e) flottante.

La première grande différence est la disparition de l'opposition subordonnée-coordonnée. Dans la nouvelle catégorisation, les coordonnées sont un type de subordonnée.

La deuxième différence est que les propositions appelées complétives sont regroupées selon la nouvelle catégorisation sous le nom de substantives. Les propositions dites interrogatives appartiennent toujours à cette classe, mais les propositions de citation sont classées maintenant dans les subordonnées adverbiales.

Le troisième changement est la définition des subordonnées flottantes. Ce sont des propositions fonctionnant comme des éléments externes à la proposition tels que les éléments indépendants de la grammaire scolaire.

Propositions subordonnées complétives

鍵　を　忘れた　こと　に　気がついた。
(*kagi - wo - wasureta - koto - ni - kigatsuita*)
(clés - [*wo*] - oublier [passé] - le fait de [substantif formel] - [*ni*] - se rendre compte [passé])
« (Je) me suis aperçu que (j')avais oublié mes clés. »

Ce sont des propositions fonctionnant comme des compléments du prédicat. Il en existe trois types.

Le premier est la proposition terminée par un substantif dit formel tel que *koto* dans l'exemple. Cette structure est identique à celle d'une subordonnée déterminant un substantif normal et l'ensemble formé par la proposition déterminante et par le substantif formel déterminé est donc équivalent à un SN. Si bien qu'il est généralement suivi d'une particule de cas qui marque sa fonction syntaxique comme le fait la particule de cas *ni* dans l'exemple.

Le deuxième type est la proposition dite interrogative telle que :

何を　していた　か（が）　知りたい。
(*nani wo - shiteita - ka - (ga) - shiritai*)
(quoi [*wo*] - faire [état, passé] - [interrogation] - [*ga*] - vouloir savoir)
« (Je) veux savoir ce qu'(il) faisait. »

Les auteurs notent que les particules *ga* – marquée entre parenthèses dans l'exemple – et *wo* suivant la subordonnée sont souvent omises.

Le troisième type est la proposition introduite par l'expression de citation telle que :

誰も　いない　と　思った。
(*dare mo - inai - to - omotta*)
(personne - se trouver [négation] - [citation] - penser [passé])
« (J')ai pensé qu'il n'y avait personne. »

Dans les travaux plus récents de Masuoka, ce dernier type est classé dans les subordonnées adverbiales, du fait notamment de l'existence de l'emploi tel que :

こんなことを　していては　いけない　と　ワープロの　スイッチを　入れた。
(*kon'na koto wo - shiteite wa - ikenai - to - wâpuro no - suicchi wo - ireta*)
(telle chose [*wo*] - faire [état] - il ne faut pas - [citation] - machine de traitement de textes [no] - bouton [*wo*] - appuyer pour allumer [passé])
« (J'ai) allumé le traitement de textes (en me disant) que je n'avais pas que ça à faire. »

Propositions subordonnées adverbiales

時間　が　あれば　出席します。
(*jikan - ga - areba - shusseki shimasu*)

(temps - [*ga*] - se trouver [condition] - assister [non passé])
« Si (j')ai le temps, (j')y assisterai. »

Il existe six formes :

1. proposition terminée par le prédicat à une forme non conclusive telle qu'une forme neutre ou de condition ;
2. proposition suivie d'une particule de mise en relief ;
3. proposition déterminant un substantif formel suivie d'une particule de cas ;
4. proposition suivie d'une particule conjonctive ;
5. proposition suivie d'un suffixe ;
6. proposition suivie d'une locution conjonctive.

Les auteurs proposent également des sous-catégories selon le sens (temporel, causal, etc.).

Propositions subordonnées déterminantes

きのう　借りた　本　を　読んだ。
(*kinô - karita - hon - wo - yonda*)
(hier - emprunter [passé] - livre - [*wo*] - lire [passé])
« (J)'ai lu le livre que (j')avais emprunté hier. »

Les subordonnées déterminantes sont distinguées en trois classes selon la relation syntaxique existant entre elles et leur base.

Lorsque la base a une fonction syntaxique de complément vis-à-vis du prédicat de la subordonnée déterminante, cette dernière est appelée 補足語修飾節 (*hosokugo shûshoku setsu*, proposition qualifiant le complément).

Les auteurs appellent substantifs relatifs les substantifs tels que 前日 (*zenjitsu*, la veille) ou あと (*ato*, après) dont le sens a une valeur relative. Lorsque la base est un substantif relatif, la subordonnée qui la détermine est appelée 相対名詞修飾節 (*sôtai meishi shûshoku setsu*, proposition qualifiante avec un substantif relatif). La subordonnée suivante en est un exemple :

フランスへ　発つ　前日
(*furansu e - tatsu - zenjitsu*)
(France [*e*] - partir - la veille)
« la veille (du jour) (où il) partira/est parti pour la France ».

Le complément supprimé de la subordonnée est « le jour », adverbe de temps pour le prédicat « partir ». La base déterminée par la subordonnée, « la veille », est un substantif relatif ayant une valeur relative par rapport à ce complément supprimé de la subordonnée, « le jour ».

Le troisième type est la proposition appelée 内容節 (*naiyô setsu*, proposition de contenu). Dans cette proposition, la base n'a pas de fonction syntaxique. La proposition exprime le contenu de l'élément auquel se réfère la base.

実験が　失敗した　報告
(*jikken ga - shippai shita - hôkoku*)
(expérience [*ga*] - échouer [passé] - rapport)
« rapport (informant) (que) l'expérience a échoué ».

La subordonnée exprime le contenu « l'expérience a échoué » du « rapport ». Les auteurs expliquent que ce dernier type de proposition est souvent réalisé avec des expressions de citation telles que « *to iu* ». Dans les travaux postérieurs de Masuoka, l'ensemble des déterminantes est d'abord distingué en deux types selon la nature de connexion : sans connecteur et avec connecteur « *to iu* » ou « *yôna* ».

Masuoka semble avoir retravaillé ce classement, et dans ses ouvrages plus récents, l'auteur définit une nouvelle catégorisation comme suit :

1. déterminantes en relation interne (内の関係, *uchi no kankei*) ;
2. propositions de contenu ;
3. déterminantes réduites (縮約連体節, *shukuyaku rentai setsu*).

Le premier type concerne les subordonnées qu'il a appelées dans ses travaux antérieurs *hosokugo shûshoku setsu*, proposition qualifiant le complément. La relation interne[11] est celle constatée entre cette déterminante et sa base : cette dernière a une fonction syntaxique de complément vis-à-vis du prédicat de cette première.

Le troisième type, le nouveau, désigne des subordonnées déterminantes dont la relation avec leur base est implicite. Par exemple, dans la structure :

復習しなかった　報い
(*fukushû shinakatta - mukui*)
(réviser [négation, passé] - punition)
« punition qui résulte de mon absence de révision »

la relation entre « punition » et « ne pas avoir révisé » qui est « A comme résultat de B » est implicite et l'interprétation par « punition qui résulte de mon absence de révision » nécessite des données extra-linguistiques.

Propositions coordonnées

父　が　フランス人　で、母　が　日本人　です。
(*chichi - ga - furansu jin - de - haha - ga - nihon jin - desu*)
(père - [*ga*] - français - [copule : neutre] - mère - [*ga*] - japonais - [copule : conclusive, polie])
« Mon père est français, ma mère japonaise. »

Ce sont des propositions qui sont sur un pied d'égalité avec la principale.

[11] La notion de relation interne, opposée à la relation externe (外の関係, *soto no kankei*), provient des travaux sur les subordonnées déterminantes de Teramura, très captivants. L'analyse précise des subordonnées déterminantes ne concernant pas directement nos travaux actuels, nous ne décrivons pas ici les détails de ses travaux, quoiqu'intéressants et bénéfiques. Ses principaux articles sur la subordonnée déterminante sont repris dans Teramura (1993).

Il existe deux types de formes : propositions terminées par le prédicat à une forme neutre et propositions avec un connecteur, en particulier une particule conjonctive.

Les auteurs signalent que la première forme sert également à constituer la subordonnée adverbiale.

Enrichissement avec les travaux réalisés dans le cadre d'une application au TAL

Dans le cadre de la conception du système CBAP (*Clause Boundaries Annotation Program*) (Maruyama et al., 2004) – système de détection automatique des frontières de proposition –, les auteurs complètent la liste des propositions de Masuoka et Takubo afin d'obtenir une liste exhaustive permettant de détecter les propositions de manière automatique[12].

Du fait de leur caractère appliqué et surtout automatique, ces travaux présentent le grand avantage d'être complets. Les auteurs ont d'abord défini dix sous-classes de propositions connectées en sous-catégorisant les quatre classes de subordonnée définies dans Masuoka & Takubo (1992) :

- Proposition complétive :
  1. proposition complétive (2) ;
  2. proposition de citation (5) ;
  3. proposition interrogative indirecte (3) ;
- Proposition adverbiale :
  4. proposition de condition et de concession (23) ;
  5. proposition de cause (8) ;
  6. proposition de temps (21) ;
  7. proposition de manière (12) ;
  8. autres (38) ;
- Proposition déterminante :
  9. proposition déterminante (15) ;
- Proposition coordonnée :
  10. proposition coordonnée (12).

Les valeurs entre parenthèses indiquent le nombre de patrons définis pour chaque classe.

Ainsi, sont finalement définis 139 types de propositions à reconnaître automatiquement.

Outre ces propositions, les compléments extra-prédicatifs sont aussi extraits à part et le nombre de types de frontières de proposition s'élève à 147.

[12] La description plus détaillée du système CBAP sera présentée dans § 10.2.1 et nous aborderons également le résultat de notre propre évaluation de ce système dans § 11.1.

7.9.3 Autres typologies

Les typologies des subordonnées selon les couches auxquelles elles appartiennent – du type les travaux de Minami (cf. 7.3.3) – sont abondantes, et nous ne citerons que Masuoka (1997) et Noda (2002). Nous ne réalisons pas un état de l'art plus détaillé de ces travaux dont nous ne pouvons profiter dans nos travaux. Masuoka (1997) présente également une typologie selon le niveau de dépendance de la subordonnée vis-à-vis de la principale.

7.9.4 Récapitulation et analyse critique

Le tableau 7.7 (voir page suivante) est une synthèse comparative des études présentées.

Tableau comparatif

La colonne « Teramura » indique la classe à laquelle appartient la subordonnée de l'exemple selon la catégorisation de Teramura (1982a), la colonne « Masuoka & Takubo (92) » celle selon la grammaire de Masuoka & Takubo (1992), et la colonne « Masuoka (1997, 2002) » celle selon les travaux de Masuoka (1997, 2002).

Le point d'interrogation signale que la subordonnée équivalente à l'exemple n'est pas traitée dans les travaux concernés et que la classe indiquée résulte de notre interprétation réalisée d'après la définition des auteurs.

Analyse des typologies

La catégorisation de Masuoka et Takubo, que nous avons qualifiée de catégorisation selon la fonction, est en réalité beaucoup moins cohérente que ne le laisse entendre cette qualification.

En effet, la définition même de coordination est plutôt sémantique que syntaxique. Masuoka (1997) dit lui-même que les subordonnées coordonnées peuvent être interprétées comme adverbiales, mais qu'étant donné que du point de vue sémantique, les propositions coordonnées sont sur un pied d'égalité avec leurs principales, il est incorrect de les traiter comme des adverbiales.

De plus, le statut des subordonnées substantives est également discutable car ce regroupement est réalisé avec des critères non pas du niveau fonctionnel mais du niveau lexical, à savoir les catégories lexicales équivalentes. Dans la mesure où un mot d'une catégorie donnée peut assumer différentes fonctions, la classe substantive peut coïncider avec les autres classes définies par des critères fonctionnels. C'est justement la raison pour laquelle les propositions de citation posent autant de problèmes dans leur catégorisation. Les auteurs n'abordent pas ce point dans leurs ouvrages, mais certaines subordonnées classées dans les adverbiales sont également des substantives (ou complétives selon la terminologie de leurs anciens travaux).

	Teramura	Masuoka & Takubo (92)	Masuoka (1997, 2002)
花子が　詩を　書き、太郎が　作曲した。 *hanako ga - shi wo - kaki - tarô ga - sakkyokushita* Hanako [*ga*] - parole [*wo*] - écrire - Tarô [*ga*] - composer **« Haeako a écrit les paroles et Tarô a composé »**	**T1**	**Crd**	**Crd**
雨が　降って、試合が　中止に　なった。 *ame ga - futte - shiai ga - chûshi ni - natta* pluie [*ga*] - tomber - match [*ga*] - cessation [*ni*] - devenir **« Il a plu et le match a été aeeulé (le match a été aeeulé à cause de la pluie) »**	**T1**	**Crd** ou **Adv**	**Crd** ou **Adv**
時間が　あれば、出席します。 *jikan ga - areba - shussekishimasu* temps [*ga*] - exister [condition] - assister [politesse] **« Si j'ai le temps, j'y assisterai »**	**T2**	**Adv**	**Adv**
鍵を 忘れた　こと　に　気がついた。 *kagi wo - wasureta - koto - ni - kigatsuita* clés [*wo*] - oublier [passé] - le fait [substantif formel] - [*ni*] - se rendre compte [passé] **« (Je) me suis aperçu que (j')avais oublié mes clés »**	**T3**	**Cmp**	**Sub**
きのう　借りた　本を　読んだ。 *kinô - karita - hon wo - yonda* hier - emprunter [passé] - livre [*wo*] - lire [passé] **« (J)'ai lu le livre que (j')avais emprueté hier »**	**T3**	**Dét**	**Dét**
太郎は　英語が　上手だ　し、花子は　スペイン語が　話せる。 *tarô wa - eigo ga - jôzuda - shi - hanako wa - supeingo ga - hanaseru* Tarô [*wo*] - anglais [*ga*] - fort - [addition] - Hanako [*wa*] - espagnol [*ga*] - savoir parler **« Tarô est fort ee aeglais, et ee plus Haeako sait parler espageol »**	**T4**	**Crd**	**Crd**
何度も　説明した　のに　理解してもらえなかった。 *nando mo - setsumeishita - noni - rikaishite morae nakatta* plusieurs fois - expliquer [passé] - [concession] - pouvoir faire comprendre [négation, passé] **« (J'ai eu) beau expliquer à plusieurs reprises, je e'ai pas pu (lui) faire compreedre »**	**T4**	**Adv**	**Adv**
何を　していた　か　が　知りたい。 *nani wo - shiteita - ka - ga - shiritai* quoi [*wo*] - faire [état, passé] - [interrogation] - [*ga*] - vouloir savoir **« (Je) veux savoir ce qu'(il) faisait »**	**T5 ?**	**Cmp**	**Cmp**
誰も　いない　と　思った。 *dare mo - inai - to - omotta* personne - se trouver [négation] - [citation] - penser [passé]) **« (J')ai peesé qu'il e'y avait persoeee »**	**T5**	**Cmp**	**Adv ?**
こんなことを　していては　いけない　と　ワープロの　スイッチを　入れた。 *kon'na koto wo - shiteite wa - ikenai - to - wâpuro no - suicchi wo - ireta* telle chose [*wo*] - faire [état] - il ne faut pas - [citation] - machine de traitement de textes [*no*] - bouton [*wo*] - appuyer pour allumer [passé] **« (J'ai) allumé le traitemeet de textes (ee me disaet) que je e'avais pas que ça à faire »**	**T5**	**Cmp** ou **Adv ?**	**Adv ?**
話が　まとまった　かどうか　早く　結果が　知りたい。 *hanashi ga - matomatta - ka dôka - hayaku - kekka ga - shiritai* propos [*ga*] - conclure [passé] - [interrogation totale] - vite - résultat [*ga*] - vouloir savoir **« (Je) veux vite coeaître le résultat (pour savoir) s'ils se soet mis d'accord »**	**T5 ?**	**Cmp** ou **Adv ?**	**Flt**

Crd = coordonnée, **Cmp** = complétive, **Sub** = substantive, **Adv** = adverbiale, **Dét** = déterminante, **Flt** = flottante

TAB. 7.7 – Comparaison des typologies des subordonnées

En revanche, la typologie de Teramura basée uniquement sur des critères formels est tout à fait cohérente. Nous adopterons donc pour nos travaux la typologie proposée par Teramura.

7.10 Notre typologie des subordonnées

Nous définissons donc les différentes subordonnées comme suit :

– Subordonnées sans connecteur :
  1. **subordonnées neutres** : propositions ou sous-phrases terminées par un mot variable à une forme neutre, éventuellement suivi de particules de mise en relief ;
  2. **subordonnées de condition** : propositions terminées par un mot variable à une forme de condition, éventuellement suivi de particules de mise en relief ;
  3. **subordonnées déterminantes sans connecteur** : propositions terminées par un mot variable à une forme conclusive.

– Subordonnées avec connecteur :
  1. **subordonnées avec particule conjonctive** : propositions ou sous-phrases terminées par un mot variable à une forme conclusive suivi d'une particule connective ;
  2. **subordonnées avec connecteur agglutinant** : propositions constituées d'une déterminante et d'un connecteur agglutinant, éventuellement suivi de particules de cas et/ou de mise en relief. Ces propositions se distinguent elles-mêmes en deux types :
     a) à fonction de complément essentiel : suivies d'une particule de cas ;
     b) à fonction de complément accessoire : utilisées seules sans particule de cas ;
  3. **subordonnées de citation** : propositions ou sous-phrases terminées par un mot variable à une forme conclusive terminées par la particule *to* ;
  4. **subordonnées déterminantes avec connecteur** : propositions ou sous-phrases terminées par un mot variable à une forme conclusive suivi d'un connecteur déterminant.

Distinction des connecteurs durs et souples

Notre définition ne tient pas encore compte de la différence entre les connecteurs durs et souples.

Leur distinction exacte – considérée comme non indispensable pour notre sujet actuel – ne sera pas réalisée dans le cadre de la présente thèse, mais l'état de l'art présenté dans ce chapitre pourrait constituer la base de futurs travaux.

Cas de la double apparition des connecteurs

Il existe des cas où plusieurs connecteurs se succèdent les uns après les autres, comme par exemple la phrase interrogative terminée par la particule *ka* (connecteur agglutinant) introduite dans son ensemble par la particule *to* de citation. Dans ce cas, nous considérons comme connecteur celui apparaissant en dernier.

7.11 Récapitulation : définition formelle de la phrase

D'après tout ce que nous avons vu jusqu'ici, nous pouvons écrire formellement la structure de la phrase japonaise comme suit[13] :

phrase → sous-phrase ? phrase
phrase → élément-externe* thème ? proposition
proposition → subordonnée ? proposition
proposition → racine
racine → complément* prédicat-forme-conclusive
sous-phrase → élément-externe* thème ? subordonnée-neutre particule-mise-en-relief ?
sous-phrase → phrase particule-conjonctive
subordonnée → (subordonnée-condition | subordonnée-connecteur-agglutinant | subordonnée-neutre) particule-mise-en-relief ?
subordonnée → proposition particule-conjonctive
subordonnée-déterminante → (subordonnée-déterminante particule-mise-en-relief ?) | (phrase connecteur-déterminant)
subordonnée-citation → phrase particule-*to*.
complément → subordonnée-citation.
complément → subordonnée-connecteur-agglutinant particule-de-cas.
…
SN → déterminant N.
…
déterminant → subordonnée-déterminante.
…
élément-externe → subordonnée.
…

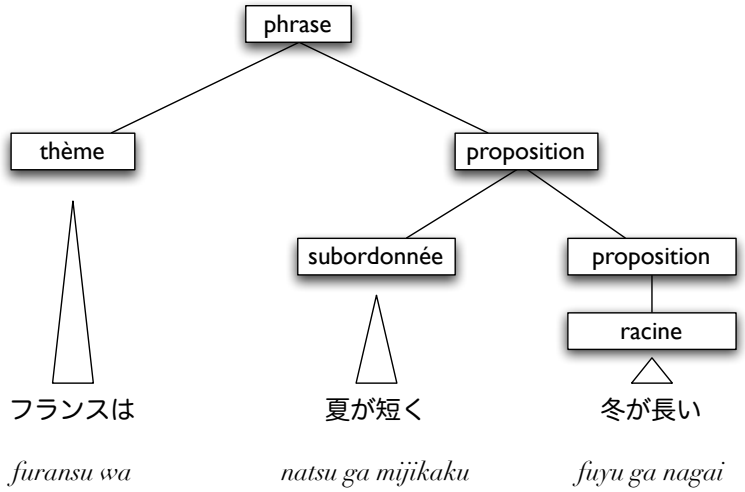
Considérons maintenant deux exemples d'analyse de phrases selon cette définition.

Exemple d'analyse 1 : phrase contenant une subordonnée

フランス　は　{夏　が　短く、}{冬　が　長い。}
(*furansu - wa -* *natsu - ga - mijikaku - fuyu - ga - nagai*)
(la France - [*wa*] - été - [*ga*] - être court - hiver - [*ga*] - être long)

[13] Du fait du sujet de nos travaux, seules les règles concernant les structures propositionnelles sont décrites. Cette définition formelle a pour but d'obtenir cohérence et précision dans notre définition des subordonnées et nous ne cherchons nullement à construire une grammaire pour l'analyse syntaxique complète. Il est évident que beaucoup d'autres règles seraient à définir pour une analyse syntaxique complète.

« En France, l'été est court et l'hiver est long. »



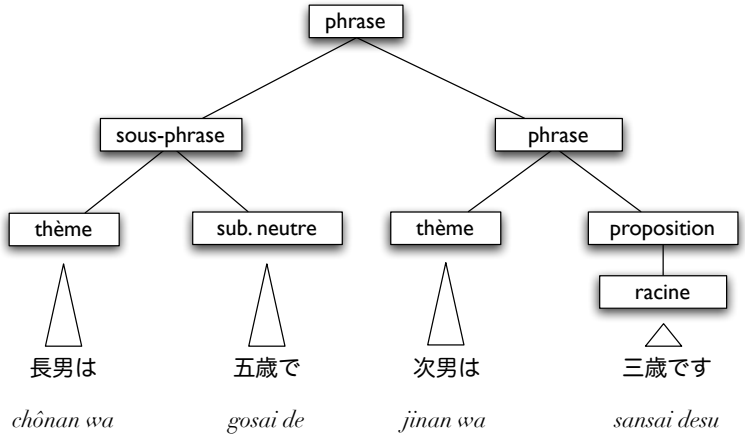
Exemple d'analyse 2 : phrase comportant une sous-phrase

長男　は　五歳　で、　次男　は　三歳　です。

(*chônan - wa - go sai - de* - *jinan - wa - san sai - desu*)

(fils aîné - [*wa*] - 5 ans - [copule] - deuxième fils - [*wa*] - [copule : poli])

« Notre fils aîné a cinq ans et notre deuxième a trois ans. »



7.12 Relations entre le syntagme thématisé et les subordonnées

Comme nous l'avons vu dans la section 6.4.3, le syntagme thématisé a généralement une double fonction. Il entre en relation, en tant que thème de la phrase,

avec l'ensemble du prédicat principal régissant tous ses compléments. Le syntagme thématisé joue, en plus, vis-à-vis du radical du mot prédicatif, un rôle de complément.

Nous allons maintenant étudier les relations que le syntagme thématisé entretient avec le prédicat de chaque subordonnée constituant la phrase.

7.12.1 Mécanisme général

中国は [国力が　伸びて] [{自分たちの　国際的地位を　高める}戦略を　とっている]。

(*chûgoku - wa -* [*kokuryoku ga - nobite*] -

[*jibuntachi no - kokusaiteki chii wo - takameru - senryaku wo - totteiru*])

(Chine [*wa*] - pouvoir national [*ga*] - grandir -

leurs - statut sur le plan international [*wo*] - élever - stratégie [*wo*] - adopter [progressif])

« Son pouvoir national ayant grandi, la Chine adopte une stratégie permettant d'élever son statut sur le plan international »

La proposition s'opposant au thème de cette phrase est constituée de la subordonnée neutre et de la proposition racine (encadrées).

Le syntagme thématisé « *chûgoku wa* (la Chine [*wa*]) » assure, en plus de celui de thème de la phrase, le rôle de complément vis-à-vis des mots prédicatifs de ces sous-propositions. Il assure le rôle de complément secondaire du nominatif « *kokuryoku* (pouvoir national) » du premier prédicat « *nobite* (grandir) » et le cas de *ga* vis-à-vis du prédicat de la racine « *totteiru* (adopter) ».

La proposition racine contient elle-même une subordonnée déterminant son complément en *wo*. Dans la phrase d'exemple, le syntagme thématisé n'assure aucune fonction cumulative vis-à-vis du prédicat de la déterminante contenue dans la racine, mais il est tout à fait possible qu'il joue un rôle de complément vis-à-vis du prédicat situé au plan secondaire.

Détermination de la fonction cumulative du thème

Afin de déterminer la fonction cumulative du syntagme thématisé, on cherche la place du cas vide dans chaque proposition constituant la phrase, comme représenté figure 7.8 page suivante.

Les informations sur les cas nécessaires à chaque prédicat sont prédéfinies. Lors de l'analyse, on prévoit les emplacements des cas nécessaires à chaque prédicat. Les compléments explicites sont ensuite associés a ces emplacements.

Dans l'exemple étudié (cf. figure 7.8 page ci-contre), le prédicat principal « *totteiru* » a entraîné la génération des emplacements pour le cas *ga* et pour le le cas *wo* (encadrés sur la figure). L'emplacement ouvert par le prédicat de la subordonnée « *nobite* » se limite au cas *ga*, mais le syntagme nominal a également ouvert un emplacement pour son complément secondaire.

Pour automatiser cette opération, différentes informations seraient nécessaires et devraient être stockées dans des bases de données du type diction-

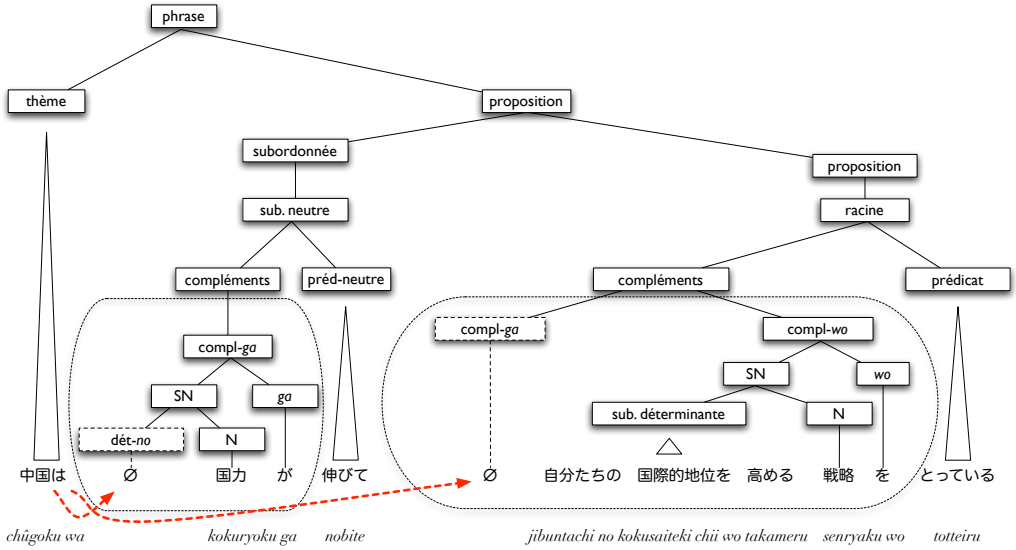


FIG. 7.8 – Détermination de la fonction cumulative du thème

naire : types de compléments nécessaires pour chaque mot variable, caractères de chaque substantif constituant chaque complément, etc.

De plus, beaucoup de travaux restent à réaliser afin de déterminer précisément les différentes règles : conditions d'ouverture des emplacements de complément secondaire, accessibilité du syntagme thématisé, etc.

7.12.2 Problème lié à la portée du thème dans la structure introduite par la particule *to*

Il existe des phrases où le syntagme thématisé est considéré comme appartenant à la subordonnée. Comme nous pouvons le voir dans notre définition formelle de la phrase, certains connecteurs, notamment celui de citation, peuvent introduire une sous-phrase ou même une phrase contenant un thème.

Masuoka & Takubo (1992) analysent le thème des deux phrases suivantes par celui appartenant à la subordonnée de citation introduite par la particule *to* (ou le connecteur déterminant comportant cette particule)[14] :

P1 計画　は　中止する　べきだ　という　意見　が　多かった。

(*keikaku - wa - chûshi suru - beki da - to iu - iken - ga - ôkatta*)

(projet - [*wa*] - annuler - il faut que - [connecteur déterminant] - opinion - [*ga*] - être nombreux [passé])

« Les opinions considérant qu'il fallait annuler le projet étaient nombreuses. »

[14] L'encadrement n'est pas effectué dans l'ouvrage, mais ajouté ici pour mieux comprendre la structure.

P2 鈴木さん　は　まだ　学校　に　いる　と　思う。

(*suzuki san - wa - mada - gakkô - ni - iru* - *to - omou*)

(M. Suzuki - [*wa*] - encore - école - [locatif] - se trouver - [citation] - je pense)

« Je pense que M. Suzuki est encore à l'école. »

Ambiguïté d'interprétation

La figure 7.9 représente cette interprétation – par laquelle le thème appartient à la sous-phrase – du premier exemple sous forme d'un arbre syntaxique selon notre définition.

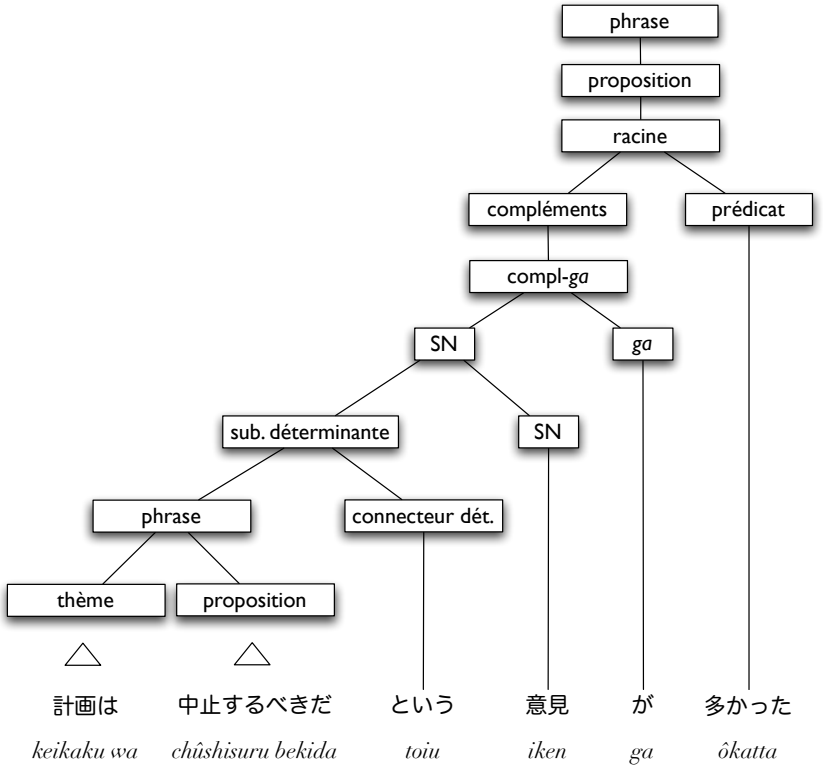


FIG. 7.9 – Interprétation de P1 - 1

Mais l'appartenance du thème à la subordonnée peut-elle vraiment être déterminée selon le type de subordonnée ? En effet, il est tout à fait possible d'imaginer une situation où l'on parle d'un projet et que quelqu'un explique qu'il y avait beaucoup d'opinions défavorables vis-à-vis de ce projet lors d'un référendum en disant : « Ce projet, il y avait plus de gens qui pensaient qu'il fallait l'annuler. » Dans ce cas, le thème « *keikaku - wa* » est bien celui de la phrase et non pas seulement de la sous-phrase de citation. La figure 7.10 page ci-contre représente cette seconde interprétation.

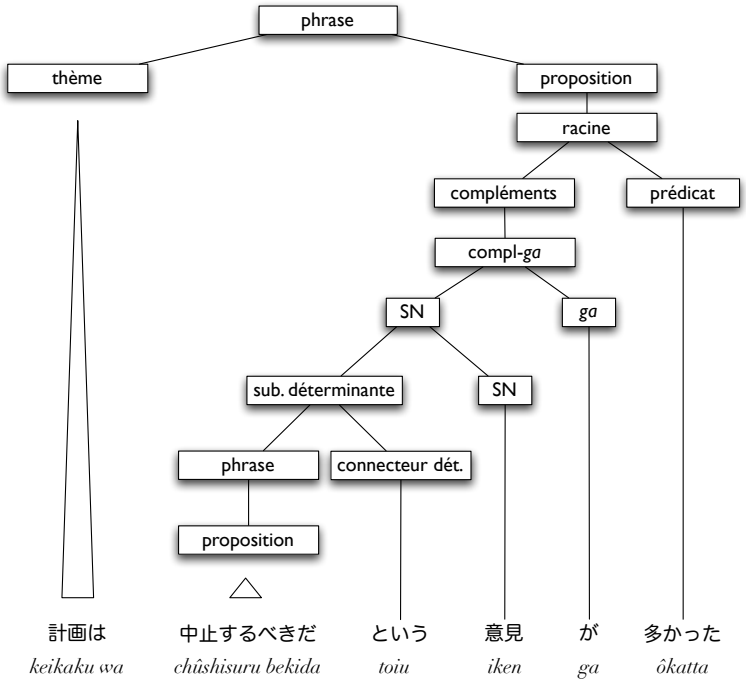


FIG. 7.10 – Interprétation de P1 - 2

Analyse du syntagme thématisé

L'appartenance du thème à une subordonnée ne peut pas être déterminée par le type de cette subordonnée, qui fournit seulement une information sur la possibilité ou non d'appartenance du thème. En d'autres termes, si le type de connecteur (soit le connecteur proprement dit, soit la forme du mot variable réalisant la connexion) peut introduire une sous-structure contenant le thème, on installe le nœud « sous-phrase » (ou « phrase »), et dans le cas contraire, on ne peut installer que le nœud « subordonnée » duquel ne peut pas dériver la branche « thème ».

Toutefois, la détermination de l'appartenance effective du thème – dans le cas où la subordonnée peut constituer la sous-phrase contenant le thème – dépendant totalement du contexte ou de la situation, il est extrêmement difficile voire impossible de résoudre cette ambiguïté de manière automatique.

7.12.3 Notre position pour la réalisation

La détermination de la fonction cumulative du thème semble indispensable à l'analyse syntaxique de la phrase japonaise, en particulier dans un cadre multilingue avec traitement de plusieurs langues ayant des structures complètement différentes.

Néanmoins, une implémentation complète demanderait des recherches ex-

trêmement poussées. Ce point constitue un sujet à part entière, dépassant le cadre de la présente thèse. Aussi, nous contentons-nous d'examiner les résultats d'alignement sans traitement de résolution de fonction cumulative, pour analyser sa réelle influence et évaluer correctement son utilité.

7.13 Problèmes liés au phénomène d'ellipse

Tout comme pour la phrase française, le phénomène d'ellipse est une source de problèmes lors de la définition de la proposition du japonais.

7.13.1 Omission du prédicat

Étudions un exemple cité dans Mikami (1959) :

三月　に　梅　が、四月　に　桜　が　咲きます。
(*sangatsu - ni - ume - ga - shigatsu - ni - sakura - ga - sakimasu*)
(mars - [*ni*] - prunier - [*ga*] - avril - [*ni*] - cerisier - [*ga*] - fleurir [poli])
« Les pruniers fleurissent en mars, et les cerisiers en avril. »

Si on détermine les propositions en fonction de la présence d'un prédicat, cette phrase est constituée d'une seule proposition à un prédicat auquel sont attachés deux ensembles coordonnés de compléments en *ga* et en *ni*. Mais, comme le dit Mikami, il est tout à fait possible de considérer qu'elle est constituée de deux propositions et que dans la première le prédicat est omis.

Il existe également une structure similaire avec coordination d'autres compléments. Comparons les deux exemples suivants :

私　が　朝日を、夫　が　読売　を　読んでいます。
(*watashi - ga - asahi - wo - otto - ga - yomiuri - wo - yonde imasu*)
(moi - [*ga*] - Asahi - [*wo*] - mon mari - [*ga*] - Yomiuri - [*wo*] - lire [habitude])
« Je lis l'Asahi et mon mari, le Yomiuri. »

ワイン　を　二杯、ビール　を　三本　飲んだ。
(*wain - wo - ni hai - bîru - wo - san bon - nonda*)
(vin - [*wo*] - deux verres - bière - [*wo*] - trois bouteilles - boire [passé])
« J'ai bu du vin, deux verres, et de la bière, trois bouteilles. »

Le premier exemple dans lequel les sujets sont coordonnés pourrait être considéré comme constitué de deux propositions dans sa traduction française, alors que le second exemple ne peut être défini que comme une phrase d'une seule proposition.

Toutefois, dans le cas du japonais, il est impossible – du moins difficilement justifiable – de reconnaître ce type de différence entre ces deux structures. Les auteurs du système CBAP considèrent les structures de coordination partielle, telles que les exemples ci-dessus, comme des phrases complexes aussi bien pour

la coordination des compléments en *ga* que pour celle d'autres compléments. Néanmoins, les concepteurs de ce système reconnaissent également l'extrême difficulté de déterminer automatiquement les frontières de propositions pour ces structures.

7.13.2 Omission de la partie variable du prédicat

Il existe également une structure dans laquelle la partie variable du premier prédicat n'apparaît pas dans la représentation du niveau de surface. Cette construction est extrêmement courante dans l'ensemble du style écrit mais plus particulièrement dans les textes journalistiques.

二階西側　の　部屋　から　出火、二階建て住宅　を　全焼　した。
(*nikai nishigawa - no - heya - kara - shukka - nikai date jûtaku - wo - zenshô - shita*)
(le coté ouest du premier étage - de - appartement - [*kara* : point de départ] - apparition du feu - immeuble de deux étages - [*wo*] - destruction totale par l'incendie - [verbe support : passé])
« Le feu a pris dans l'appartement du premier étage du coté ouest, et l'immeuble de deux étages a été entièrement détruit (par cet incendie). »

高速道路　で　トラック　が　衝突、三人　が　死亡　した。
(*kôsokudôro - de - torakku - ga - shôtotsu - san'nin - ga - shibô - shita*)
(autoroute - [*de*] - camion - [*ga*] - collision - trois personnes - [*ga*] - décès - [verbe support : passé])
« (Deux) camions se sont heurtés sur l'autoroute et trois personnes ont été tuées. »

Les prédicats de cette structure sont, la plupart du temps, des substantifs ayant un sens représentant une action[15]. Ces substantifs constituent un verbe en s'attachant au verbe support, *suru* (する, faire)[16].

Dans cette structure, l'ensemble des éléments de la proposition, y compris le mot prédicatif, est coordonné sauf la partie variable du prédicat, *suru*. Contrairement aux exemples de la section précédente que nous hésitions à définir comme des phrases complexes du fait de l'absence de prédicat sur la représentation du niveau de surface, ces phrases conservant le radical du prédicat dans la première proposition peuvent être facilement considérées comme des phrases complexes.

Mais la détection automatique de cette structure est très difficile. Les systèmes tels que CBAP, basés uniquement sur les informations locales d'une zone assez limitée, ne peuvent pas reconnaître les frontières de ces propositions.

[15] Il s'agit souvent de mots constitués en *kanji* (*kango*) que Matsushita considérait comme des verbes et appelait verbes invariables. Mais Minami (1993) fait remarquer qu'il existe également des mots purement japonais (*wago*) de ce type et qu'aujourd'hui on rencontre beaucoup d'exemples avec des mots emprunts constitués en *katakana*.

[16] Ces substantifs peuvent également constituer un autre type de prédicat en s'associant avec la copule. Minami appelle la phrase ainsi construite phrase pseudo-substantive (voir la section 6.3.2) et présente la particularité de sa structure, semblable à celle de la phrase verbale.

7.13.3 Notre position pour la réalisation

Étant donné que nous utilisons la présence d'un mot variable comme condition minimum pour la détection des propositions, la reconnaissance des propositions dans lesquelles le mot variable constituant le prédicat est omis est impossible.

Nous reportons la réalisation de la détection des subordonnées terminées par le radical du mot prédicatif, en réalité indispensable, à de prochains travaux.

7.14 De l'arbre des constituants à la représentation en graphe des relations de dépendance des propositions

Jusqu'ici, nous avons utilisé les arbres des constituants pour représenter l'analyse structurale de la phrase. Mais comme nous exploiterons les relations de dépendance entre les propositions lors de l'opération postérieure d'alignement, nous allons présenter dans cette section comment une structure peut être représentée sous forme d'un graphe des relations de dépendance.

7.14.1 Arbre des constituants et relations de dépendance

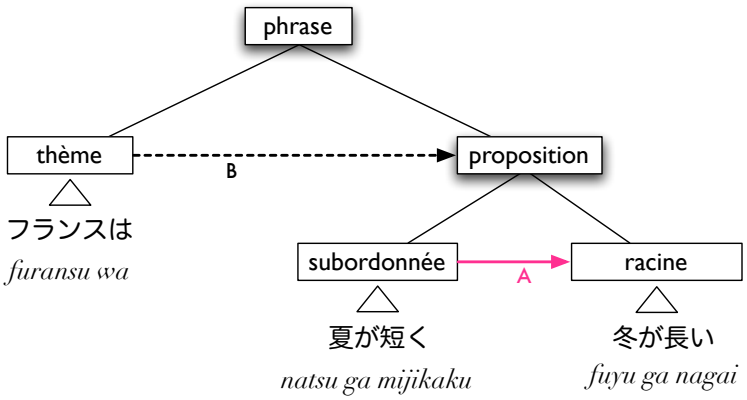
Reprenons l'arbre déjà présenté dans la section 7.11 représentant la phrase :

フランス　は　{夏　が　短く、}{冬　が　長い。}

(*furansu - wa - natsu - ga - mijikaku - fuyu - ga - nagai*)

(la France - [*wa*] - été - [*ga*] - être court - hiver - [*ga*] - être long)

« En France, l'été est court et l'hiver est long. »



Dépendance entre les constituants d'un nœud « proposition »

Nous appellerons relations de dépendance les relations qu'entretiennent deux constituants dérivant d'un même nœud « proposition », telles que celle représentée par la flèche étiquetée A dans la figure. Nous dirons que le constituant apparaissant sur la feuille gauche dépend syntaxiquement du constituant apparaissant

sur la feuille droite et l'arc orienté représente cette relation en se dirigeant de l'élément dépendant vers l'élément régissant.

Relation entre les constituants d'un nœud « phrase »

En revanche, deux constituants dérivant d'un même nœud « phrase » ou « sous-phrase » tels que ceux reliés par la flèche discontinue étiquetée B dans la figure, n'entretiennent pas de relation de dépendance mais ils entrent en relation sur un pied d'égalité afin de constituer une unité supérieure. Mais, pour des raisons pratiques et afin de minimiser la complexité de la représentation, nous représentons cette relation de la même manière que la relation de dépendance entre deux constituants d'un nœud proposition, à la différence près que l'arc est représenté par une ligne discontinue.

Représentation des constituants intermédiaires

Pour la même raison, les constituants intermédiaires (i.e. les constituants représentés par un nœud intermédiaire qui n'est pas une feuille tels que β dans le graphe A de la figure 7.11, ou encore α et β dans le graphe B) ne seront pas représentés dans les graphes de dépendance.

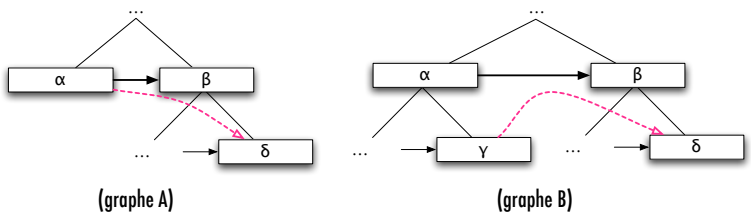


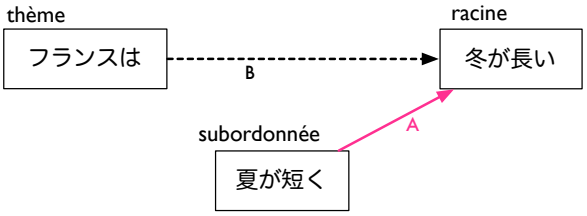
FIG. 7.11 – Relation de dépendance avec des constituants intermédiaires

Ainsi, lorsqu'un élément α entretient une relation avec un constituant intermédiaire β (cf. graphe A de la figure 7.11), nous la représenterons comme si l'élément α était en relation avec le constituant apparaissant sur la feuille tout à droite δ dérivant du nœud du constituant intermédiaire en question β. Dans le cas de la phrase d'exemple représentée dans la figure précédente, l'arc reliant le thème au nœud proposition doit alors être prolongé jusqu'à la feuille « racine ».

De même, les relations entre les deux constituants intermédiaires, α et β (cf. graphe B de la figure 7.11), seront représentées par l'arc reliant les deux constituants γ et δ, apparaissant sur la feuille tout à droite des nœuds intermédiaires α et β.

7.14.2 Graphe des relations de dépendance

En tenant compte des définitions présentées précédemment, la phrase d'exemple peut être présentée en relations de dépendance comme suit :



7.14.3 Exemple

Considérons pour mieux illustrer la définition précédente encore deux autres exemples.

Exemple 1 :

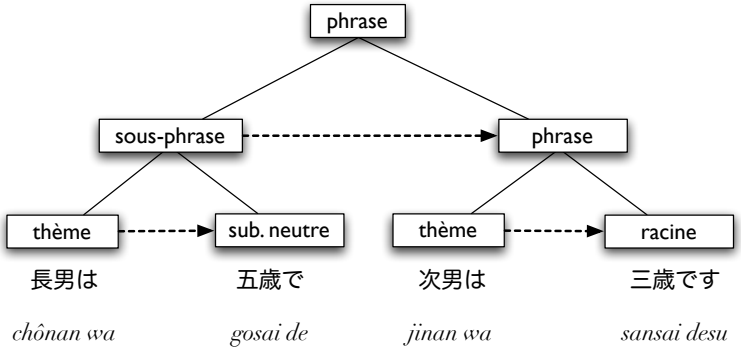
Reprenons d'abord l'exemple également présenté dans la section 7.11 :

長男　は　五歳　で、 | 次男　は　三歳　です。

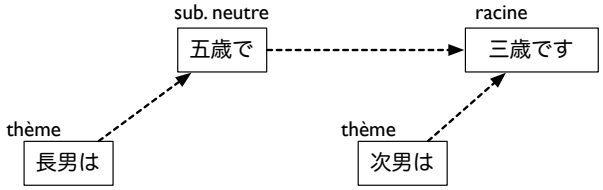
(*chônan - wa - go sai - de* - *jinan - wa - san sai - desu*)

(fils aîné - [*wa*] - 5 ans - [copule] - deuxième fils - [*wa*] - [copule : poli])

« Notre fils aîné a cinq ans et notre deuxième a trois ans. »



La représentation en graphe des relations de dépendance (au niveau des propositions) pour cette phrase est :



Exemple 2 :

Reprenons maintenant l'exemple utilisé dans la section 7.12.1 :

中国は 国力が　伸びて {自分たちの　国際的地位を　高める}戦略を　とっている。

(*chûgoku - wa -* *kokuryoku ga - nobite* -

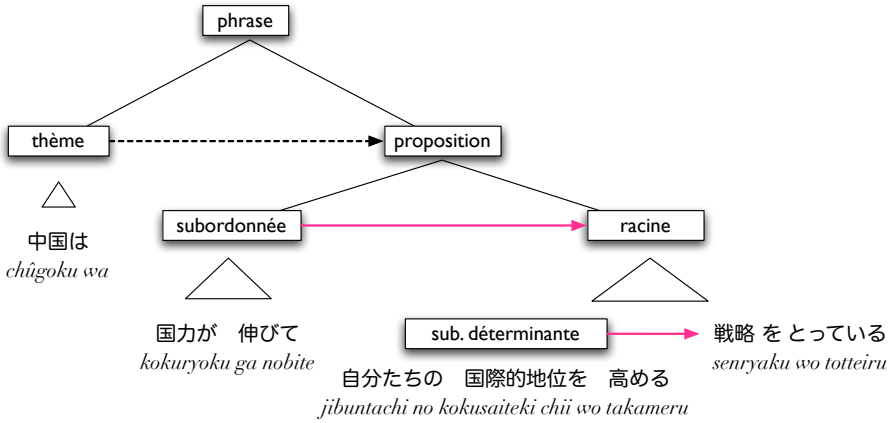
jibuntachi no - kokusaiteki chii wo - takameru - senryaku wo - totteiru)

(Chine [*wa*] - pouvoir national [*ga*] - grandir -

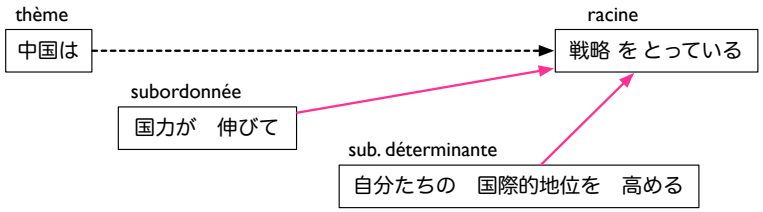
leurs - statut sur le plan international [*wo*] - élever - stratégie [*wo*] - adopter [progressif])

« Son pouvoir national ayant grandi, la Chine adopte une stratégie permettant d'élever son statut sur le plan international »

La représentation arborescente en constituants (simplifiée aux niveaux inférieurs aux propositions) est :



La représentation en graphe des relations de dépendance (au niveau des propositions) pour cette phrase est :



Troisième partie

Réalisations informatiques pour l'alignement des propositions

PLAN DE LA PARTIE

La présente partie décrit nos réalisations informatiques pour l'alignement des propositions. Elles s'articulent autour de trois tâches : reconnaissance des propositions françaises (**ch. 8** et **ch. 9**), identification des propositions japonaises (**ch. 10** et **ch. 11**) et alignement des propositions (**ch. 12** et **ch. 13**).

Chacune des descriptions de nos propres réalisations (ch. 9, ch. 11 et ch. 13) est précédée par un état de l'art sur les travaux liés existants (ch. 8, ch. 10 et ch. 12).

CHAPITRE 8

RECONNAISSANCE DES PROPOSITIONS FRANÇAISES : ÉTAT DE L'ART

Le présent chapitre est consacré à la description des travaux existants relatifs à la détection des propositions. Les méthodes proposées – principalement pour l'anglais – sont classées en deux types : celles qui recourent à un apprentissage automatique (§ 8.1) et celles avec une grammaire écrite manuellement (§ 8.2). Par ailleurs, parallèlement à cette classification, nous rencontrons également des méthodes syntaxiques partielles (conçues non spécialement pour la détection des propositions) non classiques qui ne recourent à aucune grammaire formelle (§ 8.3).

8.1 Méthodes avec apprentissage automatique

8.1.1 Ejerhed

Ejerhed (1988) propose une méthode stochastique qu'il compare avec sa méthode utilisant une expression régulière, méthode que nous abordons dans la section 8.2.1 (voir page suivante). Son système stochastique utilise pour son entraînement le résultat obtenu avec la méthode à expression régulière, et corrigé à la main, l'objectif étant d'observer la localisation des débuts et fins des propositions ainsi que celle des verbes conjugués.

Les résultats sont meilleurs avec cette méthode stochastique mais la nature de leurs erreurs diffère. Alors que la méthode à expression régulière souffre de la sous-reconnaissance des propositions, la méthode stochastique pose surtout le problème de la sur-reconnaissance.

8.1.2 *Share task* de CoNLL 2001

Lors du *ACL 2001 Workshop on Computational Natural Language Learning* (CoNLL), la détection automatique des propositions basée sur l'apprentissage machine a été choisie comme sujet de « *Share task* » (Tjong et al., 2001). Six systèmes y ont participé et différentes approches ont été proposées. L'évaluation a mis en avant le système de Carreras & Lluís (2001) utilisant l'algorithme AdaBoost.

8.2 Approche avec une grammaire régulière

Les méthodes basées sur des grammaires régulières reposent essentiellement sur les caractéristiques observables en début de proposition, le début d'une proposition étant considéré également comme la fin de la proposition précédente.

8.2.1 Ejerhed

Comme nous venons de le voir, Ejerhed (1988) propose une méthode utilisant une expression régulière, dans le même article décrit précédemment où il la compare avec la méthode stochastique.

L'expression régulière est composée de trois parties. La première comporte l'ensemble définissant les syntagmes nominaux, créé dans les travaux antérieurs. La seconde est constituée de petits ensembles de définitions supplémentaires : la ponctuation, les complémenteurs et les temps grammaticaux. La troisième partie, la plus grande, contient l'ensemble des définitions dédiées à la détection des propositions.

La définition de la proposition élémentaire utilisée dans ses travaux a trois particularités : elle se base essentiellement sur les caractéristiques observables en début de proposition ; elle considère que le commencement d'une nouvelle proposition entraîne automatiquement la fin de la proposition précédente ; enfin, elle ne traite pas plus d'une proposition à la fois. Comme la dernière caractéristique le sous-entend, dans une structure enchâssée, la principale est segmentée en deux unités, séparées par la subordonnée emboîtée. L'expression régulière ainsi définie est transformée en automate fini déterministe (AFD).

Les erreurs sont toutes dues à une sous-reconnaissance : le système n'a pas détecté beaucoup de frontières de proposition, mais celles effectivement détectées étaient toutes correctes. Ce résultat est dû à la nature incomplète de la grammaire. Par exemple, dans une structure où la subordonnée est insérée devant le verbe principal, telle que :

> *The announcement [that the President was late] was made late in the afternoon.*

la limite finale de la subordonnée n'étant pas détectée, la seconde partie de la proposition principale est incluse dans la subordonnée.

8.2.2 Abney

Abney (1990) inclut une étape de détection des propositions dans son analyseur syntaxique CASS (*Cascaded Analysis of Syntactic Structure*). En effet, le principe de ce *parser*, rapide et fiable, est de reconnaître les constituants majeurs avant d'analyser les détails de leur structure interne. Ce système présuppose comme entrée le résultat du programme de Church (1988) – réalisant l'étiquetage d'une partie de discours et de syntagme nominal –, et est constitué de trois grandes étapes : reconnaissance des *chunks*, détection des propositions et construction des arbres syntaxiques complets.

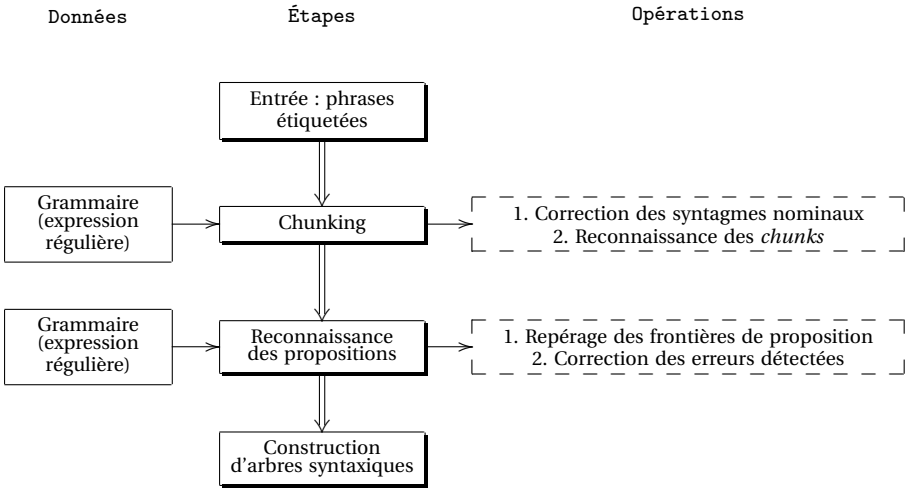


FIG. 8.1 – Analyseur syntaxique CASS

Contrairement au système d'Ejerhed qui ne remet pas en cause l'étiquetage morphologique réalisé sur le texte entré par un système extérieur, l'algorithme d'Abney étudie à chaque étape d'éventuelles erreurs produites lors des étapes antérieures.

À l'étape de *chunking*, le système vérifie les syntagmes nominaux reconnus par le programme de Church pour corriger les éventuelles erreurs. Il réalise ensuite un regroupement des mots en *chunks* à l'aide d'une grammaire écrite sous la forme d'une expression régulière.

À l'étape de reconnaissance des propositions, le *parser* détermine le début et la fin des propositions simples et identifie leur sujet et leur prédicat dans une première phase. Si aucun sujet ou prédicat n'est identifié ou si plusieurs ont été reconnus, le système indique le type d'erreur et il retouche dans une seconde phase le regroupement des *chunks* en modifiant la catégorie de certains éléments de manière à corriger les erreurs détectées.

8.2.3 Papageorgiou

Papageorgiou (1997), qui propose une méthode de détection de propositions destinée à l'alignement, se base sur cette méthode d'Abney. Sa méthode est constituée de quatre étapes. La première étape de pré-traitement consiste en la décomposition en mots graphiques du texte entré. La deuxième est dédiée à l'étiquetage réalisé par un *tagger* extérieur. À la troisième étape, le système complète le résultat du *tagging* en identifiant certaines locutions et en déterminant, en particulier, les connecteurs. La dernière étape consiste en la reconnaissance des propositions à l'aide de règles prédéfinies.

8.2.4 Leffa

Leffa (1998) propose un algorithme de détection des propositions utilisant une grammaire régulière et les informations des valences verbales. L'étiquetage de la catégorie grammaticale avec désambiguïsation ainsi que l'identification de syntagmes nominaux sont des pré-traitements requis. L'algorithme est constitué de trois étapes : identification des indicateurs de début de proposition, détection des indicateurs de fin de proposition et reconnaissance et catégorisation des propositions. Trois indicateurs de début sont définis : conjonction, verbe sans sujet et sujet d'un verbe. L'indicateur de fin est défini comme l'unité précédant un indicateur de début.

8.2.5 Maegaard et Spang-Hanssen

Pour le français, Maegaard & Spang-Hanssen (1973) présentent un programme de segmentation en propositions à l'aide de trois types d'indicateurs : les verbes finis, les conjonctions et les pronoms relatifs, ainsi que la ponctuation.

Le texte entré est d'abord converti en séquences de chiffres indiquant une des huit classes de mots définies. La seconde partie du programme réalise la segmentation par analyse des chaînes de chiffres obtenues. La grammaire est de type « *transition network grammar* » et est constituée de diagrammes à deux états, un pour les propositions principales et l'autre pour les autres subordonnées.

Ces travaux précurseurs ont eu des résultats satisfaisants, à savoir plus de 90%. Mais la grammaire étant assez limitée, le système n'est capable de détecter la fin d'une proposition subordonnée enchâssée que dans le cas où celle-ci est marquée par un élément indicateur quelconque tel qu'une virgule ou un verbe fini.

8.3 Nouvelles méthodes d'analyse syntaxique partielle

Par ailleurs, lorsque nous regardons le domaine plus large de l'analyse syntaxique partielle, beaucoup de méthodes qui ne recourent à aucune grammaire et surtout à aucun algorithme classique sont proposées. Vergne (1990) signale la

nécessité d'utiliser d'autres techniques plus robustes qu'une grammaire formelle pour représenter une langue.

Ces nouvelles analyses présentent toutes comme caractéristiques principales une grande robustesse et une rapidité, qui permettent de réaliser l'analyse de corpus de grande taille en un temps assez limité.

Nous allons maintenant étudier l'analyseur du GREYC de l'université de Caen, puis de manière détaillée, le système Syntex de l'université de Toulouse, développement le plus récent de ce type de système.

8.3.1 L'analyseur du GREYC

Nous pouvons citer comme étude plus récente pour cette tâche l'analyseur du GREYC (Vergne et al., 1999) de l'université de Caen, descendant de l'analyseur de Vergne (1989). Il est caractérisé par sa capacité de mise en relation des syntagmes, et est réalisé sans grammaire formelle. Le *tagging*, le *chunking*, et même la mise en relation des syntagmes sont réalisés en temps linéaire grâce à un algorithme conçu par Vergne et modélisé par Giguet (1998), reposant sur un calcul sur graphe.

Leur algorithme est basé sur les hypothèses psycholinguistiques d'attente et d'oubli. La construction du graphe est réalisée par interaction permanente des différentes relations, au travers de la création et la suppression d'attentes, par oubli ou satisfaction.

La délimitation des propositions dans ce système a comme fonction de définir les domaines propositionnels pour empêcher des relations inter-propositionnelles de se constituer entre deux syntagmes appartenant à des propositions différentes.

La délimitation des propositions commence par la détection de marques formelles telles que le mot « que ». Et lorsqu'on rencontre un marqueur de fin potentiel et qu'on réussit la mise en relation des syntagmes du domaine, ce rattachement déclenche la fermeture du domaine. Ainsi, la reconnaissance des propositions est réalisée en parallèle avec la mise en relation basée sur des contraintes relationnelles.

8.3.2 Syntex

Syntex (Bourigault et al., 2005), développé par une équipe de l'université Toulouse Le Mirail, réalise une analyse syntaxique en dépendance et fournit une annotation des relations de dépendance entre les mots. Tout comme l'analyseur du GREYC, c'est un système d'analyse syntaxique de corpus, robuste et opérationnel, qui ne recourt à aucune théorie syntaxique et qui n'utilise aucune grammaire formelle.

Le système reçoit comme entrée un texte étiqueté par *TreeTager*, développé à l'université de Stuttgart, duquel certaines erreurs connues sont corrigées par un module de post-étiquetage réalisé par l'équipe de Syntex.

L'analyse syntaxique que réalise ensuite Syntex est la résolution du problème :

> soit m un mot de catégorie C dans la phrase étiquetée S, quel est le recteur syntaxique de m dans S ?

et s'effectue alors via un enchaînement en cascade d'une suite de modules qui prennent chacun en charge une relation syntaxique. Les modules sont constitués d'un ensemble de règles heuristiques développées à la main par des linguistes informaticiens, selon une méthode qui met en œuvre le recours à des connaissances grammaticales et à des tests nombreux et variés sur des corpus diversifiés.

Les principaux modules sont :

1. module résolvant les relations AUX, qui cherche un dépendant (participe passé) pour un auxiliaire ;
2. module résolvant les relations ADV, qui cherche un gouverneur (verbe, nom ou adjectif) pour un adverbe ;
3. module résolvant les relations DET, qui cherche un gouverneur (nom ou pronom) pour un déterminant ;
4. module résolvant les relations XPREP, qui cherche un dépendant (nom, pronom ou verbe) pour une préposition ;
5. module résolvant les relations DE, qui cherche un gouverneur (verbe, nom ou adjectif) pour la préposition « de » ;
6. module résolvant les relations ADJ, qui cherche un gouverneur (nom) pour un adjectif et un participe passé ;
7. module résolvant les relations PREP, qui cherche un gouverneur (verbe, nom ou adjectif) pour une préposition ;
8. module résolvant les relations OBJ, qui cherche un dépendant (nom, pronom, conjonction ou verbe) pour un verbe ;
9. module résolvant les relations SUJ, qui cherche un dépendant (nom, pronom) pour un verbe.

La recherche s'effectue sous certaines contraintes telles que l'unicité du gouverneur pour chaque mot ou la projectivité interdisant le croisement des relations de dépendance.

Chaque module prend en entrée la sortie du module précédent. L'ordre des modules a une influence sur leur programmation et représente un choix très difficile à prendre pour les auteurs. Les retours en arrière sont également possibles : des relations posées par des modules antérieurs peuvent être détruites et remplacées par un module postérieur.

Syntex exploite deux sortes de ressources lexicales, exogènes et surtout endogènes. Pour le premier type, il utilise un lexique de probabilités de sous-catégorisation construit à partir d'un corpus de 200 millions de mots. Ayant constaté que certains mots du corpus spécialisé ont des comportements syntaxiques spécifiques et imprédictibles, les auteurs ont développé des procédures d'apprentissage endogène sur corpus permettant à l'analyseur d'acquérir des informations de sous-catégorisation spécifiques au corpus au cours du traitement.

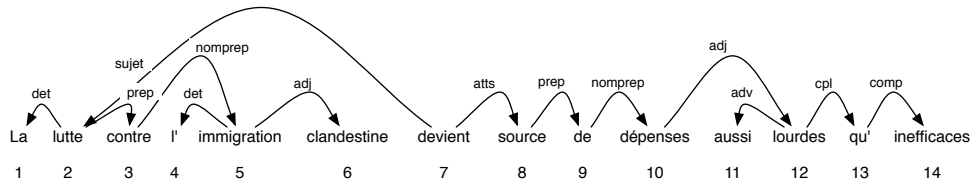
L'analyse de la phrase :

> La lutte contre l'immigration clandestine [devient] source de dépenses aussi lourdes qu'inefficaces

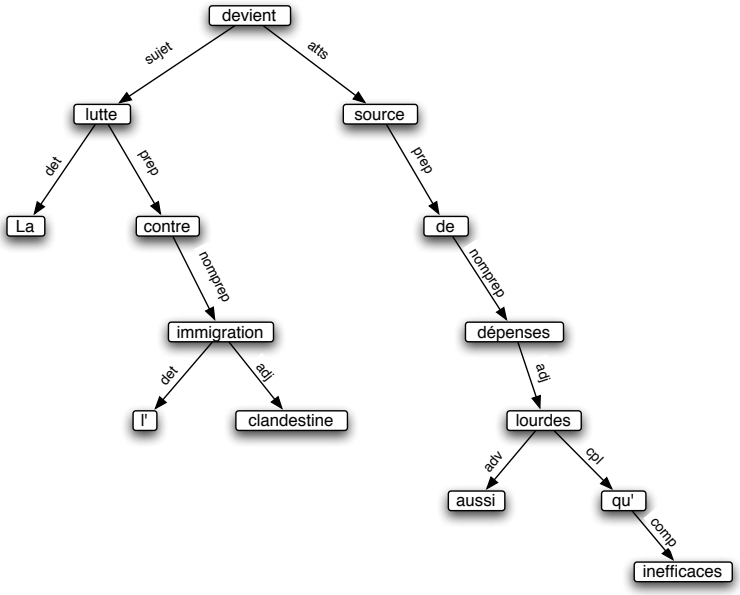
retourne :

```
<SEQ id=_1703; analyse=1;>
<TXT>La lutte contre l'immigration clandestine devient source de dépenses
aussi lourdes qu'inefficaces.
<ETIQ>DetFS|le|La|1|DET;2|
NomFS|lutte|lutte|2|SUJ;7|DET;1,PREP;3
Prep|contre|contre|3|PREP;2|NOMPREP;5
DetMS|le|l'|4|DET;5|
NomFS|immigration|immigration|5|NOMPREP;3|DET;4,ADJ;6
AdjFS|clandestin|clandestine|6|ADJ;5|
VCONJS|devenir|devient|7||SUJ;2,ATTS;8
NomFS|source|source|8|ATTS;7|PREP;9
Prep|de|de|9|PREP;8|NOMPREP;10
NomFP|dépense|dépenses|10|NOMPREP;9|ADJ;12
Adv|aussi|aussi|11|ADV;12|
AdjFP|lourd|lourdes|12|ADJ;10|ADV;11,CPL;13
CSub|que|qu'|13|CPL;12|COMP;14
AdjMP|inefficace|inefficaces|14|COMP;13|
Typo|.|.|15||
```

On peut représenter ce résultat d'analyse des relations de dépendance sous un format plus facile à lire comme suit :



L'arbre suivant est la représentation arborescente équivalente :



En extrayant de l'arbre du résultat de l'analyse d'une phrase complexe, des sous-arbres ayant comme racine le nœud verbal (ou peut-être d'un autre type selon la définition de la proposition adoptée), il serait théoriquement possible de réaliser la détection des propositions avec les résultats de Syntex.

Pour vérifier cette hypothèse, nous avons travaillé sur quelques résultats d'analyse de Syntex[1] de manière à reconnaître les propositions dans les phrases.

Pour cela, nous avons d'abord construit des arbres de dépendance à partir des résultats des phrases afin de pouvoir extraire des sous-abres dont la racine est un verbe fini, qui correspondraient à l'unité que nous avons définie comme proposition.

Analyseur robuste, Syntex fournit toujours un résultat, contrairement à notre système dont le rappel d'analyse n'atteint jamais 100%. Toutefois, l'analyse de Syntex est très détaillée au niveau des constituants plus petits que la proposition, mais très partielle au niveau de l'ensemble d'une phrase : le résultat d'une phrase correspond souvent non pas à un arbre mais à une forêt dont certains arbres sont constitués d'un seul élément. La figure 8.2 page suivante montre la forêt représentant le résultat d'analyse par Syntex de la phrase :

> J'ai souhaité rappeler que les gens qui semblent n'en pas disposer (ouvriers, gens de couleur, femmes), sitôt qu'ils s'organisent et protestent à l'échelle d'une nation, se donnent un pouvoir qu'aucun gouvernement ne peut aisément réprimer.

Avec cette forêt, sont difficiles – sinon impossibles – non seulement l'identification des relations entre les propositions, mais aussi la reconnaissance même des propositions constituant la phrase.

[1] Nous remercions D. Bourigault qui a eu la gentillesse de nous fournir les résultats d'analyse de nos corpus.

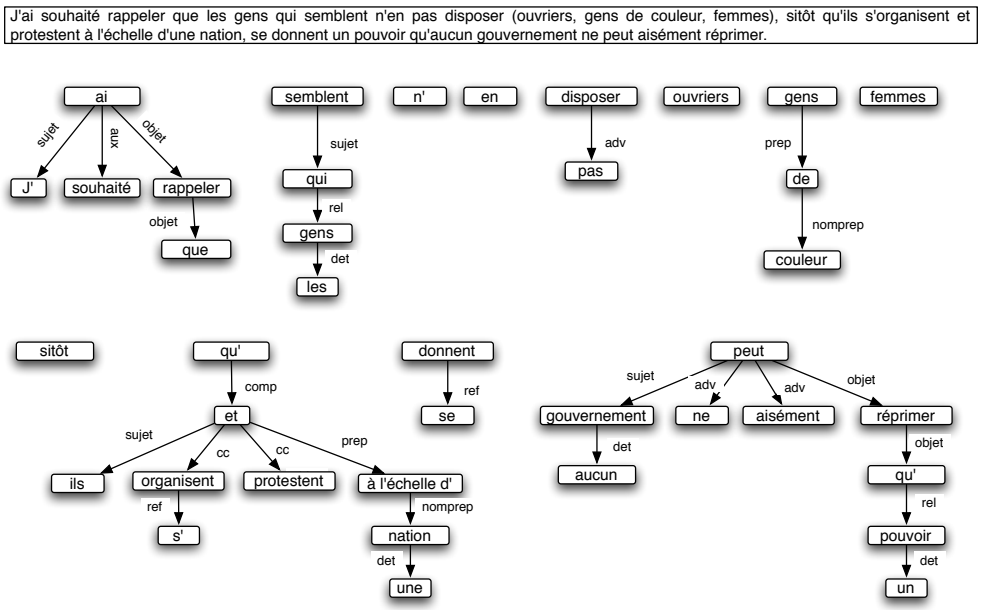


FIG. 8.2 – Forêt représentant le résultat d'analyse d'une phrase par Syntex

Ce constat montre que différentes sortes de partialité existent selon les besoins de l'application visée.

CHAPITRE 9

NOTRE SYSTÈME DE DÉTECTION AUTOMATIQUE DES PROPOSITIONS FRANÇAISES : SIGLé

しぐれ　時雨【ʃigɯre】n.

1. Brève averse. **2.** INFORM. ***SIGLé*** *(**S**ystème d'**I**dentification de propositions avec **G**rammaire **Lé**gère)* système réalisaet la détectioe des propositioes fraeçaises caractérisé par l'utilisatioe d'uee grammaire hors coetexte écrite daes ue formalisme DCG et par uee implémeetatioe ee laegage PROLOG.

Nous présentons dans ce chapitre le système SIGLé, que nous avons conçu, réalisant la détection des propositions françaises à partir du résultat du chunker *développé à l'université Paris 7. Nous allons d'abord aborder brièvement trois notions (§ 9.1), grammaires hors contexte, formalisme DCG et langage PROLOG, qui caractérisent le système. Nous présenterons ensuite le fonctionnement général de SIGLé (§ 9.2), suivi d'une évaluation du système (§ 9.3). L'exposé se terminera enfin par une discussion sur les perspectives de nos travaux (§ 9.4).*

9.1 Caractéristiques du système

Pour reconnaître les propositions à l'intérieur d'une phrase constituée d'une simple suite de caractères (i.e. texte brut sans analyse morphologique préalable), il existe deux approches possibles :

- en réalisant une analyse syntaxique complète de la phrase entrée ;
- en se contentant d'une analyse morphologique et d'une analyse syntaxique partielle, telle qu'un *chunking*.

La première approche dépend directement du résultat de l'analyse syntaxique. La seconde, au contraire, permet une certaine désambiguïsation via les résultats de l'analyse morphologique, facilitant ainsi une analyse syntaxique complète postérieure si besoin est.

La détection des propositions et l'analyse syntaxique sont en fait dans une relation circulaire. La détection des propositions serait en effet une opération aisée si les résultats de l'analyse syntaxique automatique étaient extrêmement fiables, mais elle peut également améliorer ces derniers en permettant certaines désambiguïsations à l'aide d'une grammaire locale.

Étant donné la difficulté extrême de création d'une grammaire complète permettant une analyse syntaxique intégrale, l'approche reposant sur une analyse morphologique suivie d'une analyse syntaxique partielle semble, du moins à l'heure actuelle, plus raisonnable et surtout plus bénéfique au domaine du TAL.

Nous avons donc conçu le système SIGLé, réalisant la détection des propositions françaises à partir du résultat du *chunker* développé à l'université Paris 7. SIGLé fournit comme résultat les phrases segmentées en propositions avec indication de leur nature syntaxique et de leurs relations. Il est caractérisé par l'utilisation d'une grammaire hors contexte écrite selon le formalisme des grammaires à clauses définies et par une implémentation en PROLOG.

Dans cette section, nous présentons de manière brève trois notions caractérisant notre système SIGLé : la grammaire hors contexte ou non contextuelle (*Context-Free Grammar* – CFG) – catégorie à laquelle notre grammaire appartient –, le formalisme appelé Grammaire à clauses définies (*Definite Clause Grammars* – DCG) et le langage de programmation PROLOG.

9.1.1 CFG et DCG

Nous allons d'abord aborder les avantages de la grammaire hors contexte et de la DCG avant de discuter de leurs points faibles par rapport à d'autres moyens, et de nos motivations pour ce choix en dépit de ces défauts.

Avantages

Grammaire hors contexte Contrairement aux travaux présentés dans le chapitre 8 qui recourent à des expressions régulières, notre système utilise une grammaire hors contexte. Les expressions régulières sont des grammaires dites de

type 3, sous-ensemble des grammaires de type 2 auquel appartiennent les grammaires hors contexte.

Une grammaire hors contexte est constituée des productions de la forme : soient N, l'ensemble des symboles non terminaux et Σ, l'ensemble des symboles terminaux,

$$A \rightarrow \beta \qquad \text{où } A \in N \text{ et } \beta \in (N \cup \Sigma)^*$$

alors que les grammaires de type 3 ne permettent que celles de la forme :

$$A \rightarrow x \quad \text{ou} \quad A \rightarrow xB \qquad \text{où } x \in \Sigma \text{ et } B \in N$$

Beaucoup d'unités syntaxiques ayant une structure récursive ne peuvent donc pas être représentées par des grammaires de type 3. Si bien que les systèmes de reconnaissance des propositions existants, utilisant des expressions régulières, reconnaissent en deux temps les propositions dans une phrase à structure enchâssée. On segmente d'abord la phrase en trois parties : la partie antérieure de la proposition emboîtante, la proposition enchâssée et la partie postérieure de la proposition emboîtante. Puis, par l'étude de la nature des constituants de chaque partie, on reconstitue la proposition (souvent principale) discontinue.

Par ailleurs, comme il est expliqué dans la littérature telle que Tanaka (1989), la CFG augmentée[1] possède des performances équivalentes à celles des grammaires de type 0. De plus, si une grammaire est augmentée par des règles sémantiques, il est possible de fusionner l'analyse syntaxique avec l'analyse sémantique.

Grammaires à clauses définies La DCG est justement un formalisme permettant d'écrire ces grammaires hors contexte augmentées, qui possèdent une grande capacité de description d'une langue.

Mais, le plus grand atout des grammaires à clauses définies est leur possibilité d'être presque directement compilées en tant que code PROLOG et de produire un analyseur rapide. Dans l'article de Miller & Torris (1990), l'auteur mentionne également comme un de leurs avantages la facilité d'intégration dans un autre programme écrit en PROLOG en raison de la relation étroite que ce formalisme entretient avec ce langage de programmation.

Avantages des autres grammaires ou méthodes, et nos motivations

Les travaux existants sur la détection des propositions utilisent, comme nous l'avons vu dans le chapitre 8, des expressions régulières plutôt qu'une CFG. En effet, les grammaires régulières sont préférées aux CFG pour leur simplicité de conception, d'implémentation et de calcul.

[1] Une grammaire augmentée est une grammaire à laquelle on associe un ensemble d'actions. Quand la règle de la grammaire est utilisée, ces actions sont exécutées. L'inconvénient principal des grammaires augmentées est qu'elles exigent des concepteurs de la grammaire une connaissance d'un langage de programmation pour écrire les actions à associer. De plus, il est important de conserver une bonne lisibilité après l'ajout de ces programmes, ce qui est très difficile quand il s'agit d'une grammaire de taille importante.

Par ailleurs, comme on l'a vu également dans le chapitre 8, beaucoup de méthodes d'analyse syntaxique partielle qui ne recourent à aucune grammaire et surtout à aucun algorithme classique sont proposées, et ces nouvelles analyses partielles basées sur la résolution des relations de dépendance entre les unités, sont caractérisées par leur grande robustesse.

Malgré ces avantages que d'autres méthodes présentent, nous avons choisi l'utilisation d'une CFG pour deux raisons. D'une part, une CFG écrite en formalisme DCG permet de créer facilement un *parser* en profitant du mécanisme du langage PROLOG, sans concevoir aucun moteur particulier pour l'analyse syntaxique.

Et d'autre part, une CFG, même si nous ne pouvons pas profiter dès maintenant de toutes les possibilités que présente cette grammaire, conserve toujours de grandes possibilités d'amélioration et d'évolution, et ce sans développer et rajouter à chaque fois de nouveaux modules supplémentaires.

9.1.2 Langage PROLOG

L'analyse syntaxique avec une grammaire en DCG réalisée selon le mécanisme du PROLOG est une méthode descendante en profondeur d'abord, non parallèle avec rebroussement.

Les analyseurs PROLOG avec une grammaire en DCG possèdent donc deux grands inconvénients, typiques de ce type d'algorithme et généralement objets de critiques : le problème de récursivité à gauche et celui de répétition des mêmes calculs due aux retours en arrière.

Problème de récursivité à gauche

Le problème de récursivité à gauche concerne les règles ayant en tête de leur partie droite le même non-terminal que celui de leur partie gauche, telles que :

$$A \rightarrow A\beta \qquad A \in N, \beta \in (N \cup T)^*$$

où N est l'ensemble des non-terminaux et T, l'ensemble des terminaux.

Lorsqu'on réalise une analyse avec un algorithme d'analyse descendante, ces règles provoquent une boucle infinie comme :

$$A \Rightarrow A\beta \Rightarrow A\beta\beta \Rightarrow A\beta\beta\beta \Rightarrow \ldots$$

Toutefois, les boucles ne sont pas produites seulement par ces règles : il existe des règles de récursivité à gauche « indirecte » créant le même type d'effet. Par exemple, les règles :

$A \rightarrow B\alpha$
$B \rightarrow C\beta$
$C \rightarrow A\sigma$

peuvent également provoquer des boucles infinies comme :

$$A \Rightarrow B\alpha \Rightarrow C\beta\alpha \Rightarrow A\sigma\beta\alpha \Rightarrow \cdots \Rightarrow A\sigma\beta\alpha\sigma\beta\alpha \Rightarrow \ldots$$

Néanmoins, Greibach (1965) a montré que pour toutes les grammaires hors contexte, il existait une équivalence en forme, dite forme normale Greinbach, permettant d'éviter le problème de la récursivité à gauche.

Cette forme est définie comme :

> Une grammaire hors contexte est dans une forme normale si et seulement si toutes les règles sont sous la forme
>
> $$A \rightarrow a\beta \qquad A \in N, a \in T, \beta \in (N \cup T)^*$$
>
> où N est l'ensemble des non-terminaux et T, l'ensemble des terminaux.

La forme normale de Greibach n'autorisant que les terminaux en tête de la partie droite des règles, supprime tout risque de récursivité à gauche.

Répétition des mêmes calculs

Au problème du calcul redondant dû aux retours en arrière ont déjà été proposées plusieurs solutions par différentes techniques de compilation, telles que le système BUP proposé par Matsumoto et al. (1983). Ce système transforme une grammaire écrite en DCG en un programme PROLOG d'analyse syntaxique à base de la méthode du coin gauche – qui est un algorithme d'analyse bi-directionnel[2] –, permettant ainsi non seulement d'éviter la répétition des mêmes calculs, mais aussi de résoudre le problème de récursivité à gauche. D'autres techniques ont également été proposées telles que le système SAX[3] Matsumoto & Sugimura (1986) utilisant un algorithme de *Chart Parsing* ou le système GLP (Numazaki et al., 1989) utilisant un algorithme LR.

Nos solutions adoptées

Dans la présente réalisation, nous adoptons une solution relativement simple, utilisant l'analyse tabulaire descendante, qui consiste à développer un interpréteur basé sur la méthode proposée par Pereira & Shieber (1987). Cet algorithme permet de réutiliser des résultats de calculs déjà réalisés, par consultation des résultats stockés avant de commencer un nouveau calcul, qui s'avère identique.

Néanmoins, cette solution, utilisant une stratégie descendante ne résout pas le problème de la récursivité à gauche. Afin d'éviter le problème lié à la récursivité à gauche, nous recourons à l'utilisation de la forme normale Greibach. Toutefois, pour des raisons pratiques de lisibilité, nous nous permettons parfois d'utiliser des non-terminaux en tête de la partie droite des règles lorsqu'il n'existe aucun risque de récursivité à gauche indirecte.

[2] Pour plus d'informations sur les algorithmes d'analyse syntaxique, voir Nakamura-Delloye (2003a).

[3] Le mécanisme de SAX est également étudié de façon détaillée dans Nakamura-Delloye (2003a).

9.2 Fonctionnement de SIGLé

Cette section est consacrée à la présentation du système SIGLé. Nous allons tout d'abord aborder la chaîne de traitement au cours de laquelle un texte brut entré subit différentes opérations et est segmenté à la fin en propositions. Nous parlerons ensuite de l'architecture du système SIGLé lui-même et présenterons l'ensemble des procédures réalisées par chaque module constituant le système.

9.2.1 Chaîne de traitement : du texte brut au résultat de la segmentation en propositions

L'ensemble du processus de détection des propositions à partir de la phrase brute se déroule comme représenté figure 9.1 page ci-contre.

Le texte entré est d'abord étiqueté par le *tagger* conçu à Paris 7, puis il est segmenté en *chunks* à l'aide du *chunker* développé également à l'université Paris 7 par une équipe du laboratoire LATTICE. Enfin, SIGLé reçoit comme entrée ce résultat de *chunking* et réalise la détection des propositions.

Tagger et *Chunker* de Paris 7

Il s'agit d'un *tagger* probabiliste utilisant des chaînes de Markov, dont le corpus d'entraînement est celui du « Monde » de Paris 7. Il peut fournir comme résultat non seulement celui avec l'étiquette la plus probable de chaque mot, mais aussi ceux avec les n meilleures étiquettes.

Le *chunker* de Paris 7 utilise le meilleur résultat du *tagger* et réalise un *chunking* à l'aide d'expressions régulières écrites manuellement. La grammaire est constituée d'un ensemble de règles préparatoires et de cinq types de règles, chacun destiné à identifier une catégorie de *chunks* : adverbiaux, adjectifs, nominaux, prépositionnels et verbaux. Ces cinq types de règles sont appliqués en cascade avec une stratégie de *Longest Match Method*. Le système est implémenté sous forme d'un automate déterministe à état fini.

9.2.2 Architecture du système SIGLé

La figure 9.2 page 338 présente l'architecture générale du système SIGLé. Il est constitué d'un module principal, de trois petits modules de pré-traitement, et d'un module de post-traitement :

1. Module principal :
   - gramProp
2. Modules de pré-traitement :
   - postTagging
   - postChunking
   - chu2pl
3. Module de post-traitement :

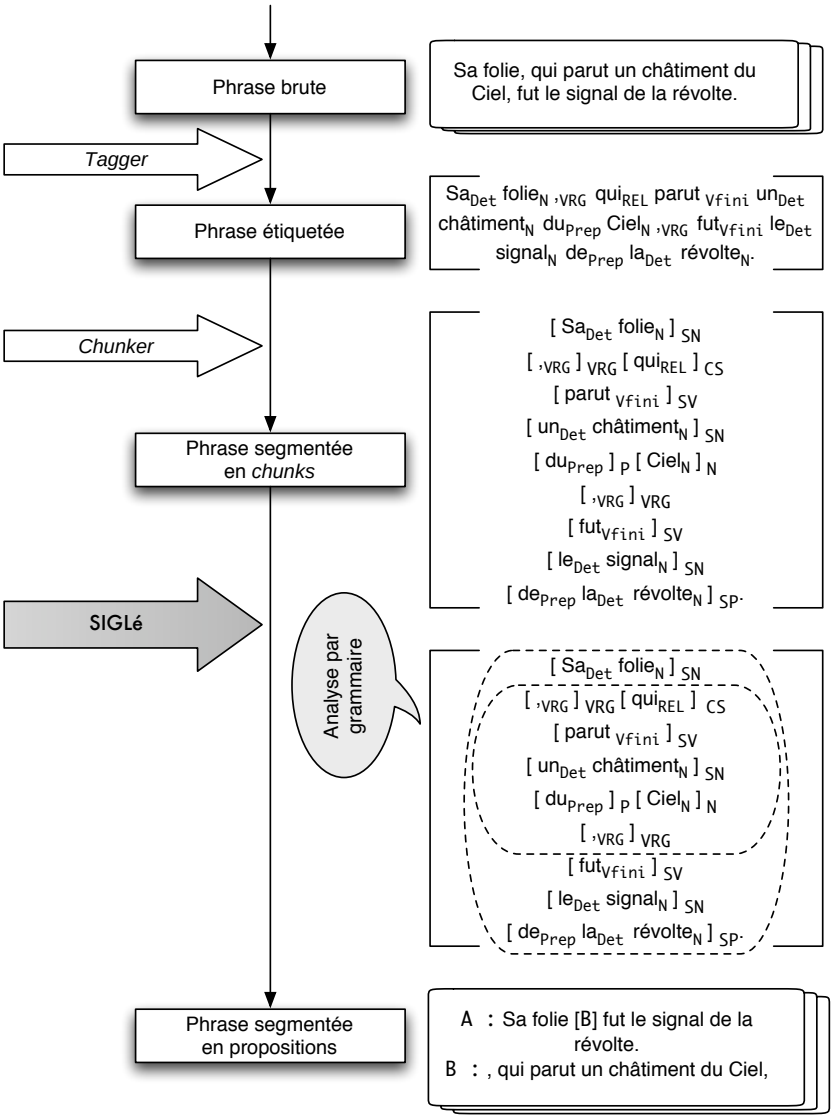


FIG. 9.1 – Procédure de détection

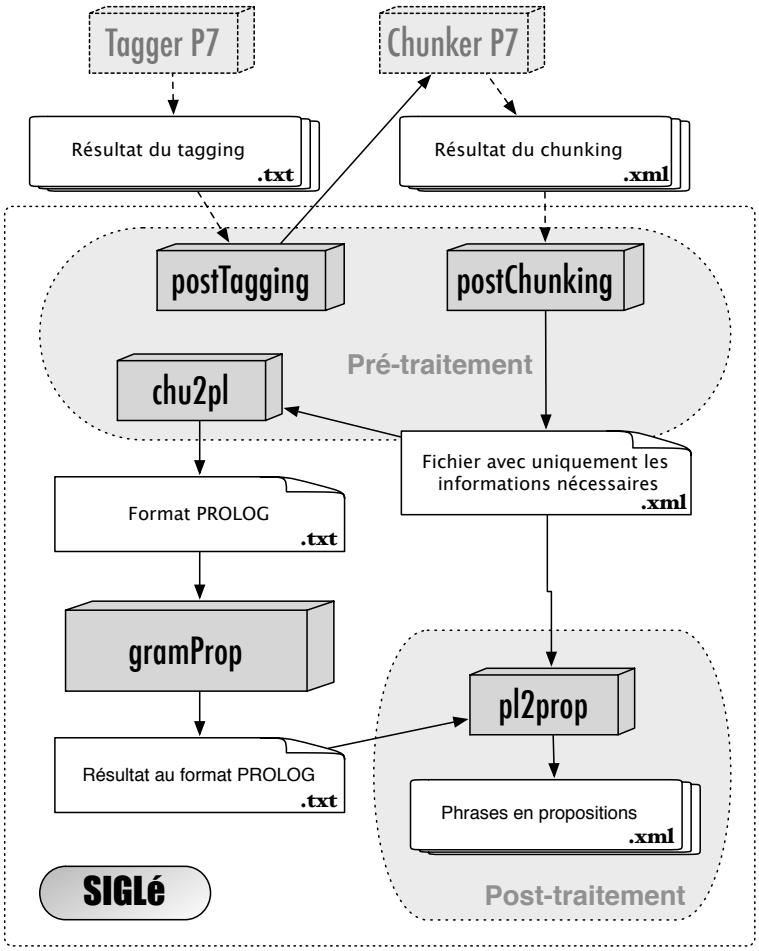


FIG. 9.2 – Schéma du système SIGLé

– pl2prop

Le module principal est un programme PROLOG. Les quatre petits modules sont tous des programmes écrits en langage Perl. Nous allons présenter maintenant les opérations particulières auxquelles est dédié chaque module.

9.2.3 Module principal : gramProp

Le module principal est développé en langage PROLOG[4]. Il réalise la détection des propositions avec une grammaire CFG, écrite sous la forme de DCG. Afin d'éviter les problèmes de répétition des calculs, l'exécution de l'analyse ne se fait pas par une simple compilation de cette grammaire DCG comme code PROLOG,

[4]Dans notre réalisation, le programme fonctionne sous l'environnement prolog SWI-Prolog (`http://www.swi-prolog.org/`).

mais à l'aide de l'interpréteur en analyse tabulaire développé également en langage PROLOG et inséré dans un niveau supérieur du programme.

En effet, le module est constitué de trois niveaux supérieurs constituant le moteur d'analyse et d'un niveau de grammaire.

Premier niveau : interface

Le premier niveau du programme du module principal correspond au prédicat analyser/2 qui s'occupe essentiellement de l'entrée des données. Il traite des buts PROLOG du fichier entré un par un : il les lit et les passe au prédicat du niveau inférieur pour qu'ils soient analysés.

Deuxième niveau : noyau du moteur

Le deuxième niveau est celui du prédicat traiter/2 représentant le noyau du moteur qui réalise principalement trois opérations : préparation de la table (*chart*), gestion du temps de calcul et insertion dynamique de certaines règles de grammaire.

La préparation de la table consiste en l'indexation des positions dans la phrase entrée. Par exemple, la phrase entrée « `np vfin np` » est indexée comme :

$$_0\ \texttt{np}\ _1\ \texttt{vfin}\ _2\ \texttt{np}\ _3$$

Cette table est représentée, comme dans Pereira & Shieber (1987), à l'aide du prédicat connects/3 sous forme de :

```
connects(np, 0, 1).
connects(vfin, 1, 2).
connects(np, 2, 3).
```

Après cette indexation, la véritable procédure d'analyse syntaxique commence enfin par l'appel du prédicat du niveau inférieur, unité principale de l'interpréteur en analyse tabulaire.

Une instruction destinée à gérer le temps de calcul est également insérée dans ce niveau afin d'interrompre l'analyse dans le cas où le temps d'analyse dépasserait un seuil défini. En effet, la conception d'un interpréteur en analyse tabulaire permet de réduire considérablement le temps de calcul par suppression des répétitions des mêmes calculs, mais les phrases fort longues (du type, ayant plus de 100 *chunks*) risquent d'entraîner de très importants calculs – surtout lorsqu'elles contiennent des erreurs d'étiquetage – pour essayer toutes les possibilités avant d'aboutir à la conclusion qu'il n'existe aucune analyse possible. Les résultats de l'introduction de ce mécanisme de *time out* seront examinés dans l'évaluation du système.

La dernière fonction principale de ce niveau, l'insertion dynamique de certaines règles de grammaire, concerne les règles traitant les subordonnées de fréquence rare (cf. études linguistiques sur la typologie des propositions § 4.6). Afin

de favoriser une analyse avec des subordonnées fréquentes plutôt qu'avec des subordonnées rarement utilisées, les règles définissant les subordonnées rares ne sont insérées dynamiquement qu'après l'échec de toute analyse avec seulement les règles des subordonnées fréquentes. L'insertion de ces règles est réalisée à l'aide du prédicat prédéfini assert/1.

Troisième niveau : interpréteur en analyse tabulaire

Le troisième niveau correspond au prédicat parse/3, unité principale de l'interpréteur en analyse tabulaire. L'instruction de la gestion du temps de calcul est insérée également à ce niveau de l'interpréteur.

La principe de l'interpréteur est basé sur la méthode proposée par Pereira & Shieber (1987). Lorsqu'il est appelé, il commence par la consultation de la table afin de chercher une/des analyses déjà abouties pour la séquence entrée. Quand il n'y a pas de résultat d'analyse, il commence une nouvelle analyse avec les règles de grammaire en appelant les clauses DCG. Le résultat d'analyse par grammaire nouvellement obtenu est représenté sous forme du prédicat known_phrase/3 et stocké à l'aide du prédicat prédéfini assert/1 afin de pouvoir être utilisé en cas de retour en arrière.

Niveau de grammaire : règles en formalisme DCG

Le niveau de grammaire est constitué d'un ensemble de clauses correspondant aux règles de grammaire, appelées par l'interpréteur du niveau supérieur pour réaliser l'analyse syntaxique.

Les règles de la grammaire présentée dans l'annexe § B sont réécrites selon le formalisme DCG avec des arguments sous forme de liste Prolog pour chaque prédicat, permettant ainsi de fournir comme résultat non seulement l'analyse structurale de la phrase entrée, mais aussi les étiquettes de chaque constituant reconnu par la grammaire.

Par exemple, la règle définissant la proposition prop → sujet, predicat. est réécrite en[5] :

```
prop([proposition, SUJ, PRED]) --> sujet(SUJ), predicat(PRED).
```

et lorsque le système réussit l'analyse d'une séquence de terminaux par application de cette règle, nous pouvons obtenir le résultat indiquant que le système a reconnu un constituant, étiqueté comme proposition, composé d'une séquence SUJ et d'une séquence PRED. Ces variables SUJ et PRED, qui représentent la structure de deux séquences constituant cette proposition, doivent d'abord être instanciées elles-mêmes par l'unification des prédicats correspondants, respectivement sujet et predicat avec des règles, comme par exemple :

[5] Par convention PROLOG, les chaînes de caractères commençant par une lettre minuscule sont des constantes et celles par une majuscule sont des variables.

```
sujet([sujet, SN]) --> sn(SN).
predicat([predicat, SV) --> sv(SV).
```

puis :

```
sn([sn, np]) --> [np].
sv([sv, vfin]) --> [vfin].
```

Ce mécanisme d'étiquetage est appliqué à tous les prédicats, et tous les constituants de phrase reconnus par une règle de grammaire sont étiquetés. Ainsi, en tant que système d'identification des propositions, le système peut fournir non seulement les propositions détectées, mais aussi les étiquettes indiquant le caractère syntaxique des propositions détectées.

9.2.4 Module de pré-traitement 1 : postTagging

Le module postTagging réalise la modification de certaines étiquettes du résultat de *tagging*. Il effectue deux types de modification : adaptation et correction.

Afin de mieux adapter les résultats de *tagging* à l'opération de reconnaissance des propositions, nous avons défini certaines étiquettes propres à notre système.

Pour la correction des erreurs évidentes, nous nous sommes concentrés notamment sur celles concernant les clitiques et les verbes qui ont des conséquences cruciales pour notre traitement.

Adaptation

Le module réalise des modifications de type « adaptation ». En effet, certaines étiquettes, bien que correctement attribuées selon la théorie adoptée par le *tagger* et le *chunker*, ne conviennent pas à notre traitement de détection des propositions.

Étiquette NE Le mot « ne » est étiqueté comme adverbe par le *tagger*, mais nous préférons l'étiquette ne propre au système. En effet, le *tag* adverbe n'apporte aucune indication directe pour la détection des frontières de proposition, alors que le mot « ne » qui apparaît en début de syntagme verbal peut indiquer une fin de proposition enchâssée, permettant donc de reconnaître correctement la proposition qui le précède.

ex. ne [adv → ne]

Étiquettes propres au système pour les connecteurs Tous les connecteurs, mots en « qu- », reçoivent, quelle que soit l'étiquette précise classique attribuée par le *tagger*, des étiquettes propres à notre système, présentées dans la figure 9.3 (voir page suivante) (voir aussi dans la section 4.7).

L'intérêt de l'utilisation de nos étiquettes non précises est considérable. Elle nous libère du risque de blocage dû à un étiquetage erroné de ces mots, certains

1. Qui, Que, Dont, Où : connecteurs isolés

   (respectivement) **qui, que, dont, où**
   comportement particulier ;

2. Camb : connecteurs ambigus

   quand, comme, si
   apparaissant en position post-verbale, en positions initiale/finale et en position post-nominale ;

3. IP : indicateurs de propositions

   quel (et ses formes fléchies), **combien, comment, pourquoi**
   apparaissant seulement en position post-verbale ;

4. Rel : connecteurs relatifs

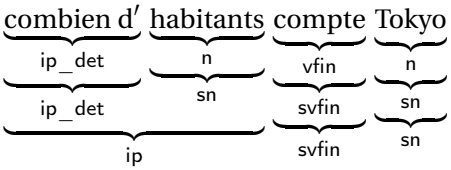
   quoi, lequel (et ses formes fléchies)
   apparaissant en position post-verbale et en position post-nominale.

FIG. 9.3 – Étiquettes des connecteurs

très polysémiques, difficiles à réaliser sans une analyse syntaxique plus large que celle avec le simple contexte immédiat. C'est typiquement le cas du mot « que ». La difficulté de son étiquetage correcte est telle que son amélioration constitue même un sujet de recherche à part entière (Jacques, 2005).

Nos propres étiquettes permettent également, sans compliquer la grammaire, d'examiner toujours les deux possibilités syntaxiques que possèdent ces connecteurs : introductions d'un syntagme et d'une proposition[6].

Comme nous l'avons déjà vu dans les études linguistiques, les connecteurs « quel(les) », « combien (de) » et « lequel (de) » (et ses formes fléchies) qui constituent parfois un syntagme avec un nom doivent être traités différemment des autres. Dans, notre réalisation, ils sont d'abord étiquetés comme ip_det (ip déterminant) et après le *chunking*, ils sont regroupé avec le chunk nominal qui les suit de manière à constituer ensemble un *chunk* ip. Par exemple, l'analyse de la phrase : « combien d'habitants compte Tokyo » se réalise comme suit :



[6]On entend ici par « proposition » et « syntagme », des unités purement de surface. Nous n'entrons pas dans la discussion sur la véritable nature de ces unités introduites par ces connecteurs, que certaines théories linguistiques traitent comme un phénomène d'ellipse.

La première ligne est le résultat du module postTagging dans lequel « combien d' » est étiqueté comme ip_det, la deuxième, celui du *chunking*, et enfin la dernière, celui après le module postChunking où ip_det constitue avec le syntagme nominal qui le suit un *chunk* ip (voir aussi la section 9.2.5).

Correction des erreurs

Comme nous l'avons déjà dit, nous nous sommes concentrés sur les corrections concernant les clitiques et les verbes ayant une influence importante pour notre traitement, et dont l'ordre linéaire est bien défini dans la phrase française. L'ordre des clitiques peut se résumer comme présenté dans le tableau 9.4 (tiré de Gardes-Tamine (1998)) :

I	II	III	IV	V	VI
je	me	le	lui	y	en
tu	te	la	leur		
il	se	les			
elle	nous				
on	vous				
nous					
vous					
ils					
elles					

TAB. 9.4 – Ordre des clitiques

Avant de présenter les corrections réalisées par le module, étudions de plus près l'étiquetage des clitiques réalisé par le *tagger*.

Étiquetage des clitiques du *tagger* de Paris 7 Le *tagger* de Paris 7 que nous utilisons adopte l'ensemble des étiquettes utilisées dans le corpus de Paris 7. L'article de Abeillé & Clement (2003) présente le principe d'étiquetage avec lequel a été créé ce corpus. Le tableau 9.5 (voir page suivante) est la reproduction du tableau « Récapitulatif Pronoms personnels (et Clitiques) »[7] présenté dans cet article.

Les mots susceptibles d'être étiquetés comme clitiques compléments sont particulièrement ambigus. Le tableau 9.6 (voir page suivante) présente tous les mots susceptibles d'être étiquetés comme clitiques compléments et différentes étiquettes qu'ils peuvent recevoir : il est créé à partir du tableau 9.4, dans lequel sont rajoutées certaines informations manquantes :

Règles pour la correction Les règles se distinguent en huit types :

[7] CL : pronoms clitiques, S : sujet, O : objet, R : réfléchi / PRO : autres pronoms / D : déterminant / V : verbe, K : participe passé / N : nom

Forme	**Étiquette 1**	**Étiquette 2**	**Étiquette 3**	**Étiquette 4**
c'	CL3ms			
ce, -ce	CLS3ms	PRO3ms	Ddefms	
elle	CLS3fs	PRO3fs		
-elle, -t-elle	CLS3fs			
en	CLO3fs	CLO3ms	CLO3mp	CLO3fp
eux	PRO3mp			
ils, -ils, -t-ils	CLS3mp			
je, -je, j'	CLS1fs	CLS1ms		
me, m'	CLO1fs	CLO1ms	CLR1ms	CLR1fs
-leur	CLO3fp	CLO3mp		
leur	CLO3fp	CLO3mp	PRO3mp	PRO3fp
lui	CLO3ms	CLO3fs	PRO3ms	VKms
-lui	CLO3ms	CLO3fs		
moi	CLO1fs	CLO1ms	PRO1ms	PRO1fs
-moi	CLO1fs	CLO1ms	NCms	
nous	CLS1fp ou mp	CLO1fp ou mp	CLR1fp ou mp	PRO1mp ou fp
-nous	CLS1fp ou mp	CLO1fp ou mp	CLR1fp ou mp	
l'on, on, (-t)-on	CLS3ms	CLS3fs		
s'	CLR3ms	CLR3fs	CLR3fp	CLR3mp
se	CLR3fs	CLR3ms	CLR3mp	CLR3fp
soi	PRO3ms	PRO3fs		
toi	PRO2ms	PRO2fs		
-toi	CLR2fs	CLR2ms		
te, t'	CLO2fs	CLO2ms	CLR2fs	CLR2ms
tu	CLS2fs	CLS2ms	VKms	
-tu	CLS2fs	CLS2ms		
vous	CLS2fp ou mp	CLO2fp ou mp	CLR2fp ou mp	PRO2mp ou fp
vous	CLS2fp ou mp	CLO2fp ou mp	CLR2fp ou mp	
y	CLO3ms	CLO3fs	CLO3fp	CLO3mp

TAB. 9.5 – Récapitulatif pronoms personnels et clitiques (reproduction de Abeillé & Clement (2003))

	cl. comp.	cl. sujet	pronom	dét.	prép.	connecteur	symbole
me, m', te, t', se	√						
s'	√					√	
lui	√		√				
le, la, l', les, leur	√			√			
nous, vous	√	√	√				
en	√				√		
y	√						√

TAB. 9.6 – Autres étiquettes de clitiques

1. concernant la préposition « en » ;
2. concernant le mot « ne » ;
3. concernant le clitique sujet ;
4. concernant le clitique complément ;
5. concernant le pronom ;
6. concernant le déterminant possessif ;
7. concernant le déterminant défini ;
8. concernant le déterminant indéfini.

Chaque règle relève des contraintes d'ordre imposées par ces catégories de mots.

La fenêtre de comparaison est limitée à deux mots consécutifs, chaque mot peut donc être examiné deux fois, avec son contexte gauche et avec son contexte droit.

Les règles de correction sont présentées dans l'annexe C.1.

Mécanisme du score Certaines corrections portent sur les erreurs pour lesquelles il est possible de choisir la correction de manière sûre. D'autres sont moins évidentes : leur possibilité de correction n'est pas unique. Nous avons tout de même choisi la solution qui nous paraissait la plus probable pour corriger le résultat de manière adéquate dans la plupart des cas, mais qui risque parfois de transformer l'erreur.

Il arrive que la correction d'un mot puisse être non ambiguë si l'on considère un contexte plus large. Pour profiter du caractère non ambigu de certains mots, nous leur attribuons un score particulier, ce qui permet de transmettre certaines informations aux mots plus éloignés que les voisins immédiats.

Par exemple, pour l'analyse erronée :

$$\text{(n') en}_{\texttt{[prép]}}\ \text{compte}_{\texttt{[n]}}\ \text{(plus que...)}$$

l'étiquette du mot « en » est d'abord modifiée en clitique objet du fait de la présence de l'élément non ambigu « ne » qui n'autorise pas de préposition à son contexte droit, et ensuite, grâce au mécanisme de l'attribution du score transmettant la sûreté de l'étiquetage, l'étiquette du mot « compte » est modifiée en verbe fini.

9.2.5 Module de pré-traitement 2 : postChunking

Le résultat de *chunking* contient tous les détails des résultats, non seulement ceux du *chunker* mais aussi ceux du *tagger* (des exemples de résultat sont présentés dans l'annexe C).

Le module postChunking extrait de ces résultats de *chunking* uniquement les informations nécessaires et crée un nouveau fichier qui sera utilisé non seulement à l'étape suivante par le module chu2pl, mais aussi par le module de post-traitement pl2prop, après la détection des propositions, pour constituer le fichier résultat final au format xml.

Le module postChunking est également chargé d'une petite modification des résultats de *chunking*, destinée à mieux les adapter à notre traitement de détection des propositions.

Les modifications effectuées sont les suivantes :

1. les *chunks* syntagmes verbaux infinitifs, vp-inf, contenant une préposition précédant un verbe à l'infinitif sont étiquetés comme pp-vinf, syntagme verbal infinitif prépositionnel :
   ex. (menaçais) d'envahir [vp-inf → pp-vinf] vs. (pouvait) donner [vp-inf] ;
2. les *chunks* syntagmes verbaux infinitifs, vp-inf, constitués d'une préposition suivie non pas d'un verbe à l'infinitif mais d'un verbe au participe présent sont étiquetés comme vger :
   ex. en déclarant [vp-inf → vger] (la fermeture) ;
3. les *chunks* syntagmes verbaux infinitifs, vp-inf, constitués d'un verbe au participe présent (sans être précédé par une propositti on) sont réétiquetés comme vptpr, verbe au participe présent :
   ex. (des combattants) ayant perdu [vp-inf → vptpr]
4. les *chunks* ip-det, suivis d'un syntagme nominal sont regroupés avec ce dernier pour constituer ensemble un *chunk* ip :
   ex. combien d' [ip-det] + habitants [np] → combien d'habitants [ip],
   ex. (à) quel [ip-det] + point [np] (la sidérurgie est ...) → quel point [ip],
   ex. quels [ip-det] + horizons [np] (... s'offraient ...) → quels horizons [ip] ;
5. les *chunks* ip-det, « quel », suivis d'un camb, « que », sont regroupés avec ce dernier pour constituer ensemble un *chunk* cs :
   ex. quels [ip-det] + que [camb] (leurs efforts) → quels que [cs] ;
6. d'autres *chunks* ip-det « quel », et les *chunks* « combien » non suivis d'un « de » constituent tout seuls un *chunk* ip :
   ex. combien [ip] (la France représente ...),
   ex. quelles [ip] (en seraient les conséquences) ;
7. les quantifieurs « beaucoup », « assez », « trop », « tant », « tellement » et « moins », étiquetés advp, constituent un *chunk* nominal np, avec le *chunk* prépositionnel pp commençant par la préposition « de » qui les suit :
   ex. beaucoup [advp] + de friches industrielles [pp] → beaucoup de friches industrielles [np] ;
   ex. trop [advp] + de concessions [pp] → trop de concessions [np].

Dans la dernière règle, très importante pour l'analyse syntaxique postérieure, résident cependant quelques problèmes.

Premièrement, la règle ne permet pas de constituer un *chunk* nominal quand le *chunker* n'a pas regroupé le « de » et le syntagme nominal suivant le quantifieur en un *chunk* prépositionnel. Cette opération, bien que peu compliquée, n'a pas été implémentée car nous avons limité la fenêtre d'examen pour le post-*tagging* à deux *chunks* consécutifs.

Deuxièmement, le traitement du mot « peu » nécessite plus de règles : en effet, à l'heure actuelle, le *tagger* (ou le *tokenizer*) ne traite pas de façon particulière la séquence « un peu » (et d'autres séquences dérivées telles que « un petit peu » ou « un tout petit peu ») et il arrive que le *chunker* produise une analyse erronée telle que : un tout petit [np] + peu [advp] + de place [pp]. Afin de traiter correctement ces cas, il faudrait plusieurs autres règles plus précises et les appliquer dans une étape très tôt dans la chaîne de traitement.

Enfin, la règle traitant les quantifieurs ne traite pas le mot « plus », qui constitue également un *chunk* nominal de la même manière. En effet, « plus » étant très polysémique, la simple application de cette règle risque de provoquer une analyse erronée, car il n'est pas possible de distinguer le « plus » de négation de celui constituant effectivement le *chunk* nominal. Par exemple, avec la simple application de cette règle, l'exemple suivant serait analysé de manière erronée comme suit :

> (un grand nombre de personnes ne se soucient) plus [advp] + de leur santé [pp] → plus de leur santé [np].

Par ailleurs, l'expression « un grand nombre de » est regroupée par le *tokenizer* (qui précède le *tagger*, développé spécialement pour ce dernier par une équipe de Paris 7) mais étiquetée comme np. Cet étiquetage pose des problème pour l'analyse syntaxique postérieure : si on définit « un grand nombre de » comme une unité, on devrait lui attribuer une étiquette du type déterminant qui pourrait être suivi d'un syntagme nominal, ou si on préfère garder l'étiquette np, il faudrait considérer « un grand nombre » comme une unité, « de » constituant un *chunk* prépositionnel avec le syntagme nominal qui le suit.

9.2.6 Module de pré-traitement 3 : chu2pl

Le module chu2pl crée une liste PROLOG à partir de l'ensemble d'étiquettes de *chunk* extraites du résultat fourni par le module postChunking. La liste PROLOG ainsi construite, peut être traitée directement comme unité par le module principal qui est un programme PROLOG.

Au moment de l'extraction des étiquettes de *chunk*, le module ne se contente pas de recopier le résultat entré : il recatégorise certaines étiquettes selon nos besoins pour la reconnaissance des propositions.

Les exemples de résultats fournis par le module postChunking pour les deux phrases d'exemple introduites dans la section précédente et de la liste PROLOG produite par le module chu2pl sont présentés dans l'annexe C.3.

9.2.7 Module de post-traitement : pl2prop

Le module pl2prop transforme les résultats du module principal – listes PROLOG constituées de suites d'étiquettes de propositions et de *chunk* – en liste au format xml, constituée non seulement des étiquettes mais aussi des chaînes de caractères initiales, que l'on peut retrouver dans le texte initial, à l'aide du fichier produit par le module chu2pl décrit précédemment (les exemples sont présentés dans l'annexe C.4).

Ce résultat, écrit en xml, peut être affiché à l'aide d'un navigateur sous un format plus agréable à lire, par la définition d'une feuille de style. La figure 9.7 en est un exemple avec une feuille de style que nous avons définie à cet effet.

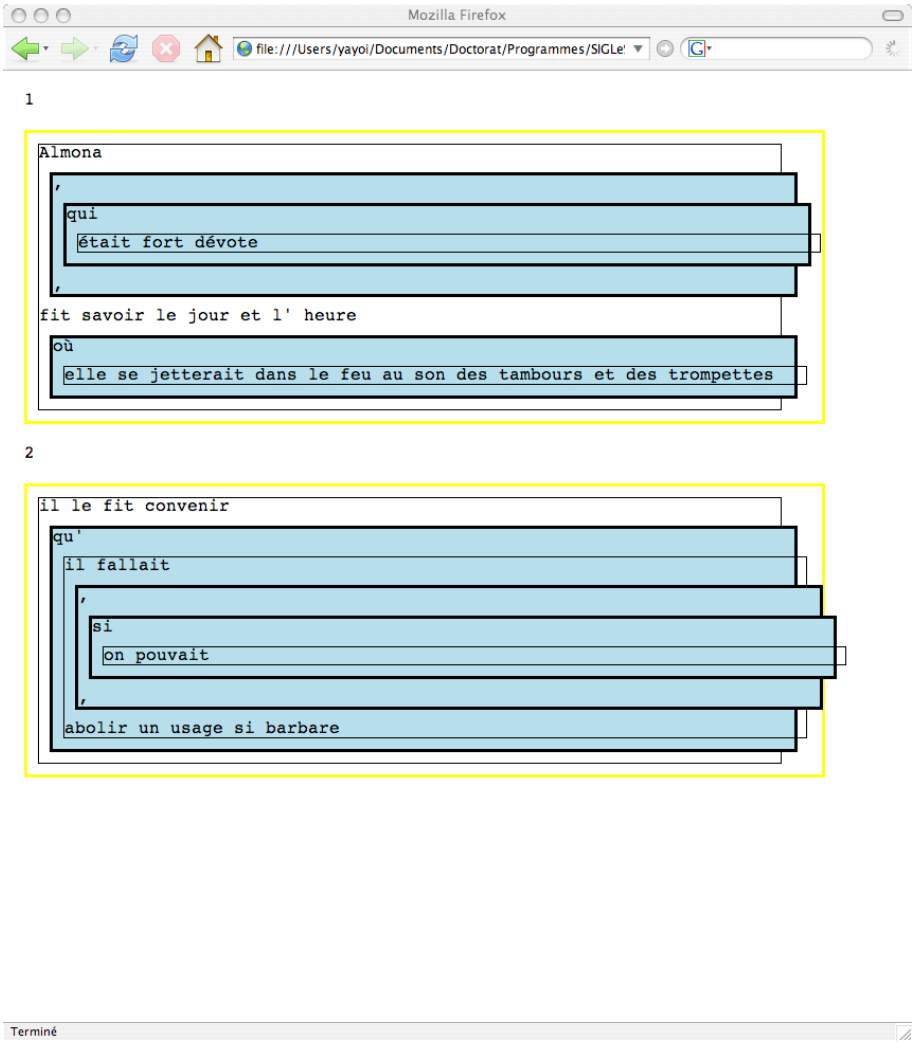


FIG. 9.7 – Résultat affiché sur un navigateur

9.3 Évaluation du système

Une évaluation a été réalisée avec quatre corpus[8] : G8 (53 phrases), Unicode (274 phrases), Zadig (extrait de 1206 phrases) et LMD (1713 phrases).

9.3.1 Résultat quantitatif

Le tableau 9.8 présente le résultat quantitatif de l'évaluation[9].

	G8	Unicode	Zadig	LMD
Nombre de phrases	53	274	1206	1713
Rappel	0,962	0,814	0,886	0,849
Précision 1	0,980	0,962	0,928	0,892
Précision 2	1,000	0,978	0,953	0,980
Précision (Préc. 1 × 2)	0,980	0,941	0,884	0,874

TAB. 9.8 – Résultat de la détection des propositions

9.3.2 Taux de rappel

Les taux de rappel sont relativement bas. Certaines erreurs sont dues à l'absence de règles adéquates, qui devront être rajoutées au fur et à mesure de l'entraînement sur de nouveaux corpus.

Mais, la grande majorité des échecs provient du résultat erroné des prétraitements (segmentation et *tagging*) à savoir plus de 90% pour G8 et Unicode. Le *tagger* a particulièrement mal supporté les séquences de symboles, ce qui explique un rappel médiocre de Unicode.

La stratégie de correction et d'adaptation des résultats de *tagging* et de *chunking* a été efficace : suite à l'introduction du module modifTag, le taux d'échec a baissé de 12 %. Ce résultat confirme l'importance de l'interaction entre l'analyse morpho-lexicale et l'opération ultérieure. D'une part, il existe des problèmes qui ne peuvent être résolus qu'avec une analyse syntaxique plus large qu'un contexte immédiat et d'autre part, surtout, les étiquettes nécessaires ne peuvent pas être entièrement définies *a priori* sans connaître leur utilisation postérieure.

[8] Des informations plus détaillées sur les corpus sont présentées dans la liste des corpus utilisés, page 547 et suivantes.

[9] Le rappel est défini comme la proportion du nombre de phrases dont l'analyse a abouti sur le nombre total de phrases. La précision 1 est définie comme la proportion du nombre de phrases dont les frontières de propositions sont correctement détectées, sur le nombre total d'analyses de phrases ayant abouti. La précision 2 correspond à la proportion du nombre de phrases dont les relations des propositions sont correctement analysées, sur le nombre total de phrases dont les frontières sont correctement détectées.

9.3.3 Taux de précision

Les taux de précision sont relativement élevés : les propositions sont généralement bien détectées et leurs relations sont bien analysées. L'utilisation d'une CFG a permis l'analyse correcte des structures imbriquées, difficiles à résoudre pour les méthodes avec une expression régulière.

Exemples de résultat d'analyse correct

La phrase à propositions multiples :

> si ces chiffres peuvent susciter l'étonnement, la triste vérité est que les habitants de Reay Road et des autres poches de misère qui prolifèrent n'ont pas mieux où aller

a été analysée sans problème (cf. figure 9.9) : la proposition 1, type racine, a deux fils : proposition 2, étiquetée subP, et proposition 3, étiquetée subQ. La proposition 3 a elle-même un fils, indexé 4 du type subR.

XML

```
<prop id='1' etq='racine' pere='0' fils ='2;3;'>
   [subP], la triste vérité est [subQ]
</prop>
<prop id='2' etq='subP' pere='1' fils ="">
   si ces chiffres peuvent susciter l'étonnement
</prop>
<prop id='3' etq='subQ' pere='1' fils ='4;'>
   que les habitants de Reay Road et des autres poches de misère [subR] n'ont pas
mieux où aller
</prop>
<prop id='4' etq='subR' pere='3' fils ="">
   qui prolifèrent
</prop>
```

FIG. 9.9 – Résultat d'analyse correct I

De même, la phrase :

> En arrivant aux frontières qui séparent l'Arabie pétrée de la Syrie, comme il passait près d'un château assez fort, des arabes armés en sortirent.

a été analysée correctement (cf. figure 9.10 page ci-contre) comme constituée de la proposition 1, type racine, qui a deux fils. Le premier est la proposition 2, étiquetée ED (Elément Détaché exra-prédicatif), qui a elle-même un fils, indexé 3 du type subR. Et le second fils de la racine est la proposition 4 étiquetée subP.

Enfin l'analyse de la phrase :

XML

```
<prop id='1' etq='racine' pere='0' fils ='2;4;'>
   [ED] [subP], des arabes armés en sortirent
</prop>
<prop id='2' etq='ED' pere='1' fils ='3;'>
   en arrivant aux frontières [subR],
</prop>
<prop id='3' etq='subR' pere='2' fils =''>
   qui séparent l'Arabie pétrée de la Syrie
</prop>
<prop id='4' etq='subP' pere='1' fils =''>
   comme il passait près_d' un château assez fort
</prop>
```

FIG. 9.10 – Résultat d'analyse correct II

> Tout ce qui passe sur mes terres est à moi, dit -il, aussi bien que ce que je trouve sur les terres des autres.

est comme montré dans la figure 9.11 : la proposition 1, type racine, a trois fils, proposition 2 étiquetée subR, proposition 3 étiquetée Incidente, proposition 4 du type subR.

XML

```
<prop id='1' etq='racine' pere='0' fils ='2;'>
   j' étais au désespoir de voir [subQ]
</prop>
<prop id='2' etq='subQ' pere='1' fils ='3;'>
   que  [ED] la destinée ne m' eût pas réservé ma portion
</prop>
<prop id='3' etq='ED' pere='2' fils ='4;'>
   dans toute la terre [subR],
</prop>
<prop id='4' etq='subR' pere='3' fils =''>
   , qui appartient également aux hommes
</prop>
```

FIG. 9.11 – Résultat d'analyse correct III

Mais, comme le montrent les chiffres du tableau 9.8 page 349, des erreurs se produisent tout de même dans des situations plus ou moins complexes.

Deux plans d'analyse

L'examen de la précision est réalisée en deux temps : du point de vue de l'analyse linéaire (Préc. 1 dans le tableau), et du point de vue de l'analyse structurale (Préc. 2). Quand on parle de l'identification des propositions, il s'agit souvent de détecter simplement les frontières des propositions. C'est une analyse linéaire qui considère qu'une phrase est constituée d'une juxtaposition de propositions. En revanche, notre analyse est du type structural qui tient compte de l'enchâssement, donc de la relation entre les propositions. Ainsi, nous avons évalué le résultat d'abord sur le plan de l'analyse linéaire puis sur le plan de l'analyse structurale.

9.3.4 Taux de précision 1 : analyse linéaire

Constat général sur le résultat d'analyse linéaire

L'analyse linéaire concerne donc juste la détection des frontières de propositions.

Les précisions élevées de G8 et Unicode proviennent du fait que leurs phrases ont une structure relativement simple. Les erreurs sur les détections de frontières se limitent essentiellement aux phrases contenant plusieurs virgules, notamment dans les structures de coordination.

Erreurs dues à la présence importante des virgules

Pour la phrase suivante, l'analyse est perturbée par la présence importante de virgules (« | » indique les frontières de propositions détectées, le symbole « * » indique le caractère erroné des frontières détectées) :

> On l'utilise, par exemple, pour comprendre les gènes | qui entrent en jeu dans la formation du cœur, des cellules sanguines, des muscles, des reins |* , de l'intestin, des yeux et enfin du cerveau.

Les règles traitant les structures détachées étant prioritaires selon l'ordre d'application dans notre grammaire, lorsqu'il y a plusieurs virgules, le système tente d'abord de reconnaître les propositions ou les compléments incidents en créant des paires de virgules. Ainsi, la première virgule fait une paire avec la dernière virgule, reconnaissant un complément incident : « par exemple, pour comprendre les gènes qui entrent en jeu dans la formation du cœur, des cellules sanguines, des muscles, des reins, de l'intestin ». Puis, la deuxième virgule fait une autre paire avec l'avant dernière virgule, reconnaissant un autre complément incident : « pour comprendre les gènes qui entrent en jeu dans la formation du cœur, des cellules sanguines, des muscles, des reins ». Enfin, à l'intérieur de cette structure, le système a détecté une relative : « qui entrent en jeu dans la formation du cœur, des cellules sanguines, des muscles, des reins »

Dans le cas de la phrase suivante, la présence d'autres virgules en position postérieure a empêché d'interpréter la première virgule comme le marqueur de fin d'une proposition détachée :

> Alors ils survivent ici, sur la route, jour après jour, malgré la pollution, la chaleur insupportable, la malnutrition, la saleté, le grondement des camions | qui passent à toute allure, les accidents, les maladies, les rats énormes et les corbeaux, les caniveaux puants |* , le dégoût des passants mieux lotis et les inondations de la mousson.

Afin d'analyser correctement ces cas, plusieurs possibilités existent. Il est sans doute possible de réduire les erreurs par la définition de contraintes telles que l'interdiction des structures récursives des propositions ou des syntagmes détachés enchâssés. Certains problèmes peuvent être résolus par l'introduction d'informations supplémentaires telles que la structure argumentale des verbes. D'autres nécessiteraient peut-être l'analyse préalable des structures de coordination, voire la conjonction de ces deux solutions. Il serait également intéressant d'envisager dans une étape de pré-traitement, l'analyse des virgules afin de les distinguer en deux types, opérateur unaire ou binaire, et de reconnaître pour le deuxième type les paires qui vont ensemble. Dans tous les cas, l'amélioration risquant de multiplier les calculs, il faut examiner différentes solutions afin de déterminer celle qui, à la fois, fournit des résultats intéressants tout en étant opérationnelle dans une implémentation réelle.

Erreurs liées à l'interprétation d'un connecteur

Par ailleurs, dans certains cas, bien qu'assez limités, un connecteur introduisant une proposition est interprété comme précédant un syntagme (ou vice versa), perturbant alors l'ensemble de l'analyse de la phrase.

> Et dire | qu'au moment de son apogée, dans les années 1950, Cockerill employait encore plus de 25 000 personnes, **que** la ville de Seraing |* était toujours noire de fumée, de bruit, de monde, de travail.

Dans cette phrase, « que » a été interprété comme introducteur d'un syntagme et non d'une proposition. En effet, la règle définissant la phrase constituée d'un sujet et d'un prédicat étant prioritaire sur les autres types de phrases dans l'ordre d'application, le prédicat de la subordonnée « était toujours noire de fumée… » est interprété comme celui de la proposition racine, ce qui a impliqué l'analyse de « que » comme introducteur d'un syntagme.

Interprétation difficile du rattachement des circonstants en fin de phrase

Comme nous l'avons signalé dans la section 4.8.2, l'analyse des circonstants détachés en fin de phrase est très difficile, car ils peuvent appartenir à la proposition qui les précède, mais la proposition qui les précède directement peut aussi être une incidente et ils peuvent appartenir à la racine. Dans la phrase suivante :

> (Ce qui) signifie l'extinction des hauts-fourneaux | qui produisaient de la fonte depuis près de deux siècles, **avec à la clé des milliers d'emplois sacrifiés**.

le circonstant détaché en fin de phrase « avec à la clé des milliers d'emplois sacrifiés » est interprété, de manière erronée, comme appartenant à la relative qui le précède, car cette dernière n'a pas été traitée comme une incidente du fait de l'absence de virgule au début.

Les circonstants liés en fin de phrase peuvent également être ambigus comme dans la phrase :

> Dites ce | que vous voulez **sur le voile**

le syntagme prépositionnel « sur le voile » peut être inclus aussi bien dans la racine que dans la subordonnée comme le système l'a fait. Seul le contexte permet une interprétation correcte. Dans le cas de notre corpus où on parle de la loi sur le voile, ce circonstant devait être analysé comme appartenant à la racine.

La résolution de ces problèmes de circonstant final est beaucoup plus délicate que les deux premiers types d'erreurs décrits précédemment, car elle nécessite des informations beaucoup plus difficiles à manipuler, à savoir des connaissances sémantiques voire extra-linguistiques.

Distinction entre l'intégrative et la relative « qui »

Le dernier exemple d'erreur concerne la difficulté de distinction entre l'intégrative et la relative « qui ». Le système n'a pas réussi à analyser correctement les intégratives lorsqu'elles suivent une préposition.

Dans la phrase suivante, alors que ce sont de parfaits exemples d'intégratives « qui détient des armes à feu » (= celui qui détient des armes à feu), « qui possède les richesses » (= celui qui possède les richesses), qui fonctionnent toutes seules comme un syntagme nominal et constituent avec la préposition « de » qui les précède un syntagme prépositionnel, les subordonnées sont interprétées comme introduites par un relatif précédé par une préposition :

> (Car omettre ces actes de résistance, ces victoires même limitées du « petit peuple » américain, reviendrait à faire croire que)
> le pouvoir est seulement entre les mains | de **qui détient des armes à feu** | , de **qui possède les richesses**.

avec les règles destinées à traiter les subordonnées telles que :

> il admet aussi un Être supérieur | , **de qui** la forme et la matière dépendent.

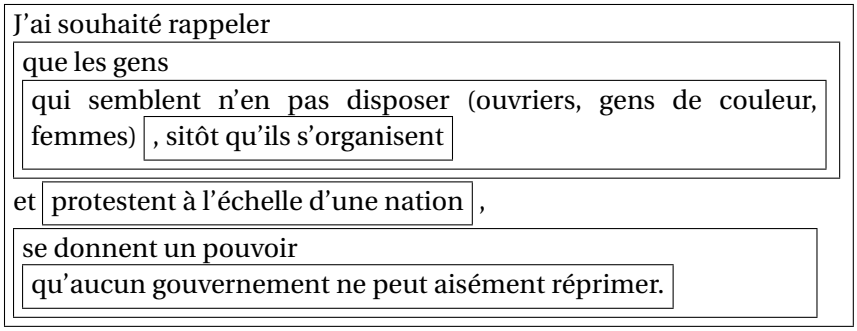
Dans notre réalisation, les règles traitant les subordonnées déterminantes (y compris celles précédées par une préposition) sont prioritaires par rapport aux subordonnées substantives à position SN considérées comme rares. La réalisation d'une analyse correcte n'est pas impossible mais elle nécessiterait l'introduction d'informations lexicales beaucoup plus précises du type animé ou non animé, informations très coûteuses en terme de calcul, et dont l'utilité est très restreinte dans notre opération.

9.3.5 Taux de précision 2 : analyse structurale

L'analyse structurale est celle des relations entre les propositions. Dans la phrase suivante, la détection des frontières est assez simple et le système réussit sans problème comme :

> J'ai souhaité rappeler | que les gens | qui semblent n'en pas disposer (ouvriers, gens de couleur, femmes) | , sitôt qu'ils s'organisent | et protestent à l'échelle d'une nation | , se donnent un pouvoir | qu'aucun gouvernement ne peut aisément réprimer.

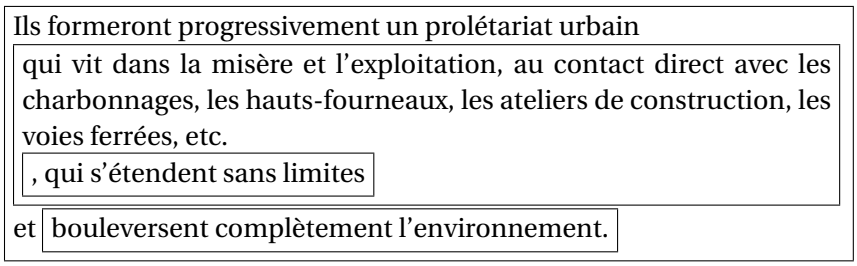
alors que l'analyse des relations est nettement plus compliquée et le système fournit un résultat erroné comme suit :



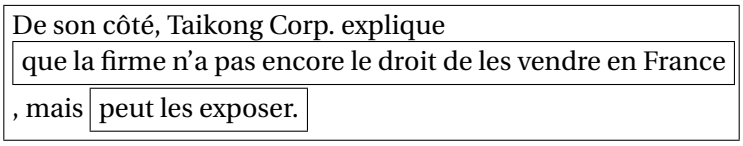
Le système est incapable de traiter, en tant que telles, les suboordonnées coordonnées sans pronom. Si bien que la proposition « protestent à l'échelle d'une nation » ne peut être interprétée que comme proposition coordonnée à la racine, ce qui a perturbé complètement le reste de l'analyse et empêché la mise en relation du sujet « **les gens** qui semblent n'en pas disposer (ouvriers, gens de couleur, femmes), sitôt qu'ils s'organisent et protestent à l'échelle d'une nation » avec le prédicat « se donnent un pouvoir [...] ».

Coordination des suboordonnées mal analysée faute de relatif

Les subordonnées sans pronom sont une des principales sources d'erreur de l'analyse structurale. Outre l'exemple précédent, la phrase suivante est également un exemple.



Pour ces deux cas, l'introduction de traits grammaticaux tels que la personne et le nombre permettrait une analyse correcte, mais dans le cas de la phrase suivante :



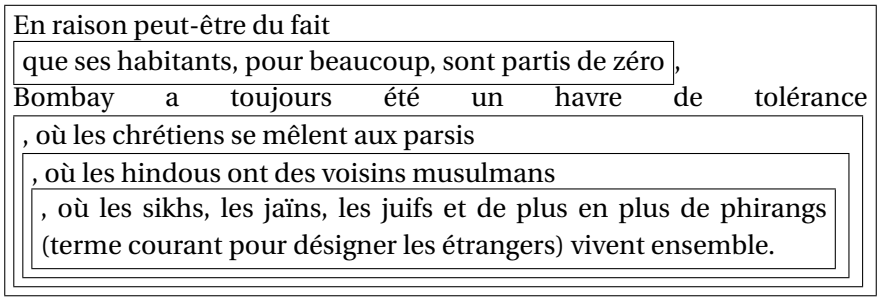
il n'est pas possible pour le système, même avec la prise en compte de ces traits grammaticaux, de choisir l'analyse correcte parmi les différents candidats.

Coordination des relatives mal analysée à cause de l'ambiguïté de la virgule

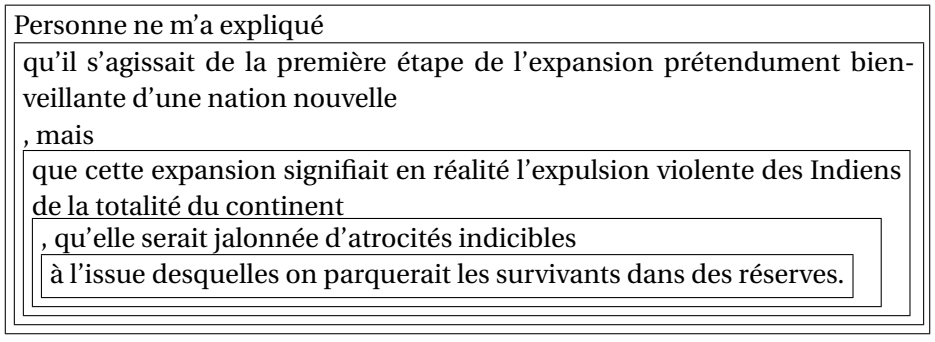
La coordination des subordonnées peut également être mal analysée lorsque la coordination est réalisée par l'ajout d'une virgule. En effet, une relative non coordonnée peut tout à fait être précédée par une virgule comme dans la phrase :

> On enseigne à tous les écoliers américains le massacre de Boston**, qui se déroula à la veille de la guerre d'Indépendance contre la couronne anglaise**.

Mais, comme la coordination des subordonnées peut être réalisée également par une virgule, cette structure de subordonnées précédée par une virgule est ambiguë pour le système et l'analyse par la relative simple non coordonnée est prioritaire, entraînant ainsi des résultats erronés :



Dans le résultat d'analyse suivant :



la deuxième subordonnée (que cette expansion signifiait en réalité l'expulsion violente...) est correctement interprétée comme complétive coordonnée à la première (qu'il s'agissait de la première étape de l'expansion...), mais la troisième

(qu'elle serait jalonnée d'atrocités indicibles...) est mal analysée, faute de conjonction de coordination, comme étant une simple relative précédée par une virgule.

Afin de résoudre ce problème, il est également nécessaire de réaliser une analyse plus précise telle que celle permettant de distinguer les relatives des complétives ou de déterminer l'antécédent des relatives pour interpréter correctement leur coordination. Cependant, ce type de calcul est, encore une fois, très coûteux et risquerait de rendre le système peu opérationnel.

Relations ambiguës

Lorsque la phrase contient trois propositions (ou plus), le rattachement de la troisième peut être ambigu et il est difficile dans ce cas de réaliser une évaluation de résultat. Ainsi, nous avons considéré certaines phrases comme ambiguës et ne les avons pas comptées parmi les erreurs.

La troisième proposition peut être une coordonnée comme dans la phrase :

Paris avait estimé, à l'époque
, qu'une référence aux valeurs religieuses n'était pas acceptable
car elle soulevait des problèmes politiques et constitutionnels en France.

Le système l'interprète dans ce cas comme coordonnée à la racine (Paris avait estimé X car Y), mais elle peut tout à fait être rattachée à la subordonnée (une référence aux valeurs religieuses n'était pas acceptable car Y).

La troisième proposition peut être une circonstantielle comme dans la phrase :

Il peut donc arriver
que, par le jeu de ces dédoublements, les organisations de l'Opus Dei soient financées plusieurs fois
sans que personne ne s'en aperçoive.

Le système l'interprète dans ce cas comme appartenant à la subordonnée (les organisations soient financées plusieurs fois sans que personne ne s'en aperçoive), mais elle peut être interprétée comme rattachée à la racine (Il peut arriver X sans que personne ne s'en aperçoive).

Faux étiquetages de subordonnées

Dans certains cas, tout en obtenant un résultat correct du point de vue aussi bien de la détection des frontières de proposition que de la résolution de leurs relations, on peut avoir un résultat erroné quant au type d'étiquette attribuée aux propositions détectées.

Ce mécanisme discriminant les subordonnées selon leur fréquence s'est montré efficace dans certaines situations où la phrase aurait pu être ambiguë et a permis une analyse correcte. Par exemple, lorsque une subordonnée introduite par

un connecteur Camb apparaissant en fin de phrase est directement précédée par un syntagme nominal telle que :

ils détestent le peuple américain quand il ne leur ressemble pas.

elle peut être soit une relative, soit une circonstancielle. Ainsi, la subordonnée « quand il ne leur ressemble pas » peut être traitée comme une relative et non pas comme une circonstancielle. Mais, l'introduction de l'opposition des subordonnées rares/fréquentes a permis, de manière économique, de favoriser l'analyse avec les subordonnées fréquentes, fournissant un résultat correct.

Néanmoins, il existe d'autres cas où ce mécanisme ne suffit pas à trouver le résultat adéquat. Par exemple, quand une subordonnée introduite par un connecteur Camb suit un verbe, elle est considérée, à cause de l'ordre de priorité des règles, non pas comme une circonstancielle, mais comme une subordonnée de complément (percontative) comme c'est le cas dans la phrase :

c'est facile à dire quand on n'est pas concerné dans sa chair.

Pour régler ce problème, il faudrait, comme pour d'autres problèmes, recourir à des informations supplémentaires et réaliser une analyse beaucoup plus fine au prix de l'augmentation des calculs nécessaires.

9.3.6 Fréquence des subordonnées

Pour vérifier nos hypothèses concernant l'opposition fréquent/rare des subordonnées, nous avons compté manuellement les occurrences de chaque type de subordonnées dans le résultat de deux corpus : LMD (qui contenait au total 501 connecteurs de la famille « qu- ») et du corpus Zadig (516 connecteurs).

Le tableau 9.12 page suivante montre le résultat d'une étude comparative présentant nos hypothèses (colonne HYP) et le résultat de comptage (colonnes LMD et ZDG). Cette étude a confirmé à peu près notre définition du qualificatif rare/fréquent des subordonnées.

En dépit de ce à quoi nous nous attendions, nous n'avons pas constaté de très grandes différences entre ces deux corpus de nature différente. La différence est constatée à un niveau plus précis entre les connecteurs employés dans chaque catégorie.

9.3.7 Remarques sur le temps de calcul

Afin d'éviter la répétition des même calculs dus au retour en arrière, l'analyse est réalisée à l'aide de l'interpréteur en analyseur tabulaire. Le temps de calcul est incomparablement amélioré grâce à l'introduction de cet interpréteur. Mais l'utilisation de mémoire est déjà très importante, et si nous envisagions l'introduction de plus d'informations, serait impératif le recours à un autre algorithme plus efficace.

Nous avons aussi rajouté une fonction de contrôle de temps de calcul. La figure 9.13 page 360 présente l'évolution du temps de calcul et le rappel selon la

Occurrence = %

		Int/Fin			post-V			post-N			Autres SN		
		HYP	LMD	ZDG	HYP	LMD	ZDG	HYP	LMD	ZDG	HYP	LMD	ZDG
Sub.	Intégrative pro.				△	0	0				△	0,4	0,8
	Percontative				✔	2	4				△	0	0
	Complétive				✔	20	27				△	0,2	0
Adj.								✔	2	0,3			
	Relative							✔	60	57			
Adv.	Intégrative adv.	✔	15	11				△	0	0			

✔ = fréquent ; △ = moins fréquent / rare

TAB. 9.12 – Fréquence des subordonnées

limitation du temps de calcul définie. Les résultats présentés et analysés jusqu'ici sont obtenus avec comme limite de temps de calcul 180 secondes par phrase. Avec cette limite, le temps de calcul moyen d'une phrase était de 3 secondes pour le corpus LMD. Mais, avec le temps maximum à 0,1 seconde, le résultat est déjà intéressant avec un rappel à plus de 80% et un temps de calcul moyen de 0,04 seconde. On a constaté un bon équilibre avec le temps maximum à 10 secondes : l'augmentation du rappel est encore significative alors que le temps de calcul reste raisonnable à savoir en moyenne 0,4 seconde par phrase.

9.4 Conclusion et pistes d'amélioration

Notre système de détection des propositions a fourni des résultats assez satisfaisants avec des taux de rappel et de précision élevés. Les erreurs sur la détection des frontières de propositions, à part celles provenant des traitements antérieurs, se limitent essentiellement aux phrases contenant plusieurs virgules dans les structures de coordination. Les erreurs sur la détermination des relations entre les propositions détectées se distinguent en trois types, mais tous les trois montrent bien la limite de l'analyse avec des informations très restreintes. Cependant, l'enrichissement des informations se traduit directement par des calculs très coûteux, ce qui risquerait de rendre le système peu opérationnel. L'introduction de l'opposition fréquent/rare des subordonnées a permis une amélioration de la précision de manière économique et assez efficace.

Les résultats de notre système sont prometteurs. Ils semblent confirmer que sont utiles voire indispensables la remise en cause des habitudes classiques ainsi

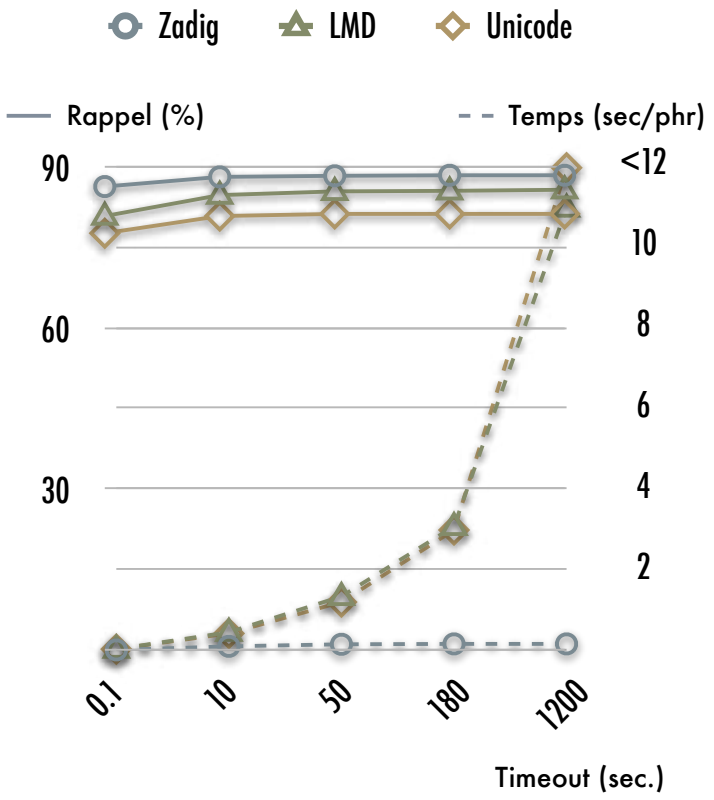


FIG. 9.13 – Limitation du temps de calcul et rappel

que le recours aux travaux linguistiques pointus sur des sujets connexes. L'utilisation des connaissances linguistiques favorise par ailleurs le développement non seulement d'un grand système complet, mais aussi d'un petit outil qui peut fournir une analyse assez détaillée sur un sujet précis avec des informations très limitées.

En dépit de ces résultats satisfaisants, nous pouvons tout de même imaginer trois grandes pistes d'amélioration : amélioration des modules de pré-traitement ; perfectionnement de l'analyse principale par l'introduction de plus d'informations ; affinement des étiquettes par la réalisation d'une analyse sémantique des connecteurs.

9.4.1 Amélioration des modules de pré-traitement

Comme nous l'avons déjà abordé, la majorité des cas où l'analyse a totalement échoué est due aux résultats erronés de *tagging* et de *chunking* et l'in-

troduction des modules de pré-traitement s'est montré très efficace. Cependant, dans le cadre de la présente thèse, les correction et modification réalisées sont en fait assez limitées. Afin de rendre plus opérationnels les résultats du *tagger* et d'augmenter le rappel, il est indispensable d'améliorer encore les modules de pré-traitement.

Il existe encore plusieurs pistes d'amélioration : beaucoup de formes ont un indice de catégorie grammaticale plus ou moins fiable. Huot dit dans son ouvrage (Huot, 2001, p. 5) :

> « pour plus de 80%, le stock de mots constituant le lexique du français moderne provient du latin. [...] une grande partie de ces mots français venus du latin sont aussi des mots construits, en ce sens qu'ils sont constitués d'éléments différents mais organisés selon des principes précis, qui les rendent par là-même repérables. »

Il est vraiment dommage de ne pas profiter pleinement de cette calculabilité de la langue – d'autant plus que le français est sans doute une des langues présentant le plus de calculabilité – et de faire tout avec la consultation des dictionnaires et des calculs probabilistes.

L'approche linguistique que nous avons adoptée est très efficace et applicable sans doute aux résultats de l'étiquetage réalisé par d'autres *taggers* probabilistes avec quelques adaptations des étiquettes utilisées. Cette approche hybride qui combine des méthodes probabiliste et linguistique représente probablement la solution la plus efficace et la plus raisonnable, du moins en l'état actuel des capacités du traitement automatique des langues.

Par ailleurs, les récents travaux d'une équipe de l'IGM (Blanc et al., 2007) proposent un système de segmentation en « *super-chunks* », unités différentes des *chunks* ordinaires, intégrant jusqu'aux attachements adjectivaux et prépositionnels. Le système fournit des résultats très intéressants qui atteignent une précision et un rappel de 92,9% et 98,7%. L'utilisation d'un système présentant de telles performances améliorerait sans doute le résultat de la détection des propositions et, qui plus est, puisqu'il réalise même des attachements adjectivaux et prépositionnels, l'opération réalisée par notre module principal serait allégée, réduisant ainsi le temps de calcul.

9.4.2 Exploitation de plus d'informations

Les résultats de la détection des propositions pourraient être améliorés par l'utilisation de plus d'informations. Mais l'introduction d'informations supplémentaires signifie une augmentation des calculs nécessaires. L'important est de savoir tracer la limite afin de ne pas perdre l'intérêt d'un petit outil opérationnel. Par ailleurs, est également envisageable une amélioration par l'utilisation d'autres outils extérieurs fournissant certaines informations précises très fiables, tels que ILIMP (Danlos, 2005), étiqueteur du pronom impersonnel « il », qui permettrait à notre système d'améliorer le résultat de la détection des complétives suivant la proposition racine à « il » impersonnel.

9.4.3 Affinement des étiquettes

Le troisième type d'amélioration est l'affinement des étiquettes pour les propositions détectées. En effet, comme on peut le constater en se référant au tableau 4.7 page 175, après la détection des propositions selon la position, on peut déterminer finalement la nature, plus usuelle, du caractère syntactico-sémantique des propositions. Par exemple, dans le cas où la subordonnée détectée contient le connecteur « quand », il est possible de déduire qu'elle est intégrative à valeur temporelle, si elle est en position post-nominale ou en position initiale/finale, ou qu'elle est percontative (interrogative à valeur temporelle), si le connecteur est en position post-verbale ou à une autre position SN, sans aucune ambiguïté.

Examinons le cas des deux connecteurs les plus ambigus « comme » et « que ».

Subordonnées en « comme »

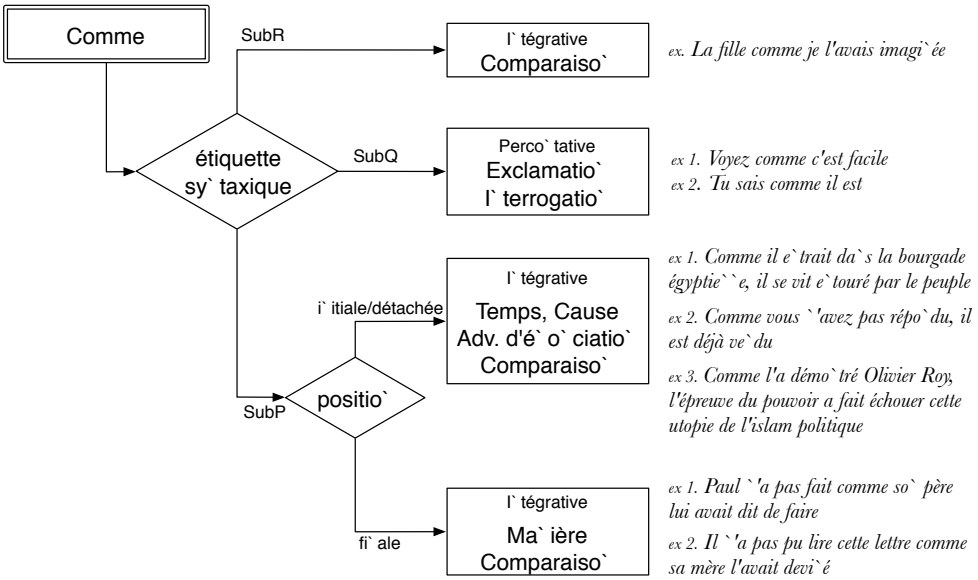


FIG. 9.14 – Étiquetage syntactico-sémantique des subordonnées en « comme »

La figure 9.14 est un schéma représentant l'étiquetage syntactico-sémantique des subordonnées en « comme ». Une subordonnée en « comme » est étiquetée comme intégrative à valeur de comparaison si son étiquette syntaxique est subR (subordonnée déterminante), ou comme percontative d'exclamation si son étiquette syntaxique est subQ (subordonnée complément). Dans le cas où son étiquette syntaxique est subP (subordonnée circonstancielle), elle a une valeur temporelle, causale, comparative ou elle est adverbe d'énonciation si elle est en position initiale ou détachée, et elle exprime la manière ou la comparaison si elle est en position finale.

Subordonnées en « que »

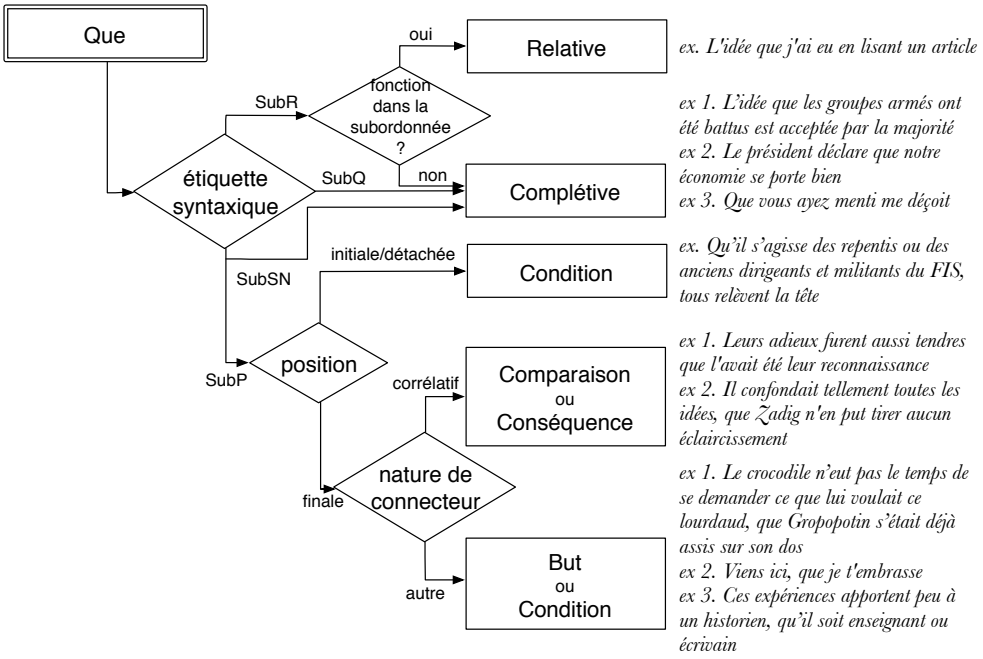


FIG. 9.15 – Étiquetage syntactico-sémantique des subordonnées en « que »

La figure 9.15 est un schéma représentant l'étiquetage syntactico-sémantique des subordonnées en « que ». Une subordonnée en « que » reçoit l'étiquette syntactico-sémantique « relative » ou « complétive » si son étiquette syntaxique est subR (subordonnée déterminante), ou complétive si son étiquette syntaxique est subQ (subordonnée complément) ou subSN. Dans le cas où son étiquette syntaxique est subP (subordonnée circonstancielle), elle a une valeur de condition si elle est en position initiale ou détachée, et elle a une valeur de comparaison, de conséquence, de but ou de condition, si elle est en position finale.

Opération nécessitant cependant des précautions

L'attribution des étiquettes syntactico-sémantiques favoriserait sans doute l'alignement, et permettrait peut-être d'élargir le champ d'application. Néanmoins, la définition de la valeur sémantique des subordonnées circonstancielles est très délicate et l'analyse varie selon les linguistes. Il nous serait donc nécessaire de réaliser une étude beaucoup plus approfondie et une évaluation du rapport entre faisabilité et apport réel.

CHAPITRE 10

RECONNAISSANCE DES PROPOSITIONS JAPONAISES : ÉTAT DE L'ART

Nous nous intéressons dans ce chapitre aux méthodes existantes relatives à la reconnaissance des propositions japonaises. Nous allons tout d'abord présenter les techniques de segmentation partielle en propositions (§ 10.1) avant de passer à la description des systèmes d'identification des propositions et d'analyse syntaxique (§ 10.2).

10.1 Segmentation partielle dans le cadre de l'amélioration d'une opération

Beaucoup d'études sur la segmentation des phrases en propositions, ou plutôt en certains types de propositions, ont été réalisées dans le cadre de l'amélioration d'autres opérations telles que l'analyse syntaxique ou la traduction automatique.

10.1.1 Méthodes basées sur la définition des motifs

Dans l'article de Saito et al. (1978) est décrite une segmentation des phrases des sections dîtes « Revendications » des brevets, toujours très longues, afin d'améliorer la qualité des résultats et de réduire le temps de calcul de l'analyse syntaxique. La méthode proposée segmente ces phrases en détectant les verbes coordonnés et les syntagmes nominaux coordonnés – contenant souvent des subordonnées relatives – à l'aide de certains mots clés (ou expressions) particuliers au style des brevets, et les virgules.

Kimura et al. (1993) et Kim & Ehara (1993) réalisent une segmentation des phrases dans le pré-traitement de la traduction automatique du japonais vers l'an-

glais. Wakao et al. (1998) font part de l'efficacité des segmentations dans le cadre de la synthèse automatique, et Ehara et al. (2000) pour la création des sous-titres de nouvelles télévisées.

Ces travaux définissent tous d'abord les types de propositions dont la segmentation est jugée utile pour l'amélioration de la traduction postérieure et réalisent la détection de ces propositions en utilisant comme repères les indicateurs de ces subordonnées. De plus, Kimura *et al.* tiennent compte également de la longueur des phrases : si la phrase contient un/des indicateurs de subordonnée à détecter sans pour autant avoir la longueur minimale définie, la segmentation n'est pas réalisée ; à l'inverse, si elle a une longueur supérieure au seuil défini sans comprendre d'indicateurs clés, la segmentation est tout de même réalisée. Par ailleurs, Kim et Ehara procèdent également au regroupement en syntagmes et utilisent, en plus des informations morphologiques, celles sur ces syntagmes au moment de la reconnaissance des motifs définissant les frontières de segmentation.

10.1.2 Méthode basée sur l'analyse des structures conjonctives

Enfin, Takeishi et Hayashi proposent dans Takeishi & Hayashi (1992) une méthode de segmentation basée sur l'analyse des structures conjonctives, conçue dans le but d'améliorer les applications d'amélioration de la rédaction.

En considérant que la méthode linéaire qui ne tient compte que des informations du niveau morphologique, voire des caractères, n'est pas capable de segmenter correctement les phrases, les auteurs proposent une méthode qui calcule les relations entre les propositions pour déterminer la segmentation la plus adéquate.

De plus, étant donné les coûts de calcul élevés de l'analyse syntaxique, ils proposent également une méthode d'analyse des structures conjonctives qui n'exploite que les informations que portent les connecteurs de proposition, sans recourir à une analyse syntaxique complète. Cette méthode d'analyse des structures conjonctives est basée sur la classification des propositions définie par le linguiste Minami (1974), et réalise une analyse des structures en déterminant les relations d'inclusion entre les propositions à l'aide de cette classification.

10.1.3 Opérations supplémentaires

Il est à noter qu'étant donné l'omission fréquente des constituants (ou des éléments de constituants) dans la phrase japonaise, dans les deux travaux (Kimura et al., 1993 ; Kim & Ehara, 1993) réalisés dans le cadre de la traduction automatique du japonais vers l'anglais, la segmentation se poursuit par l'opération de reconstitution du sujet et du prédicat. De plus, les applications telles que la synthèse automatique, la création de sous-titres ou l'amélioration de la rédaction nécessitent que les parties segmentées aient toutes une forme de phrase autonome et complète, si bien que la segmentation réalisée dans le cadre de ces travaux est également complétée par une opération de transformation des prédicats en une

forme adéquate, à savoir la forme conclusive, et par l'ajout de mots de liaison selon le sens exprimé par les connecteurs supprimés.

10.2 Segmentation en propositions

Il n'existe pas beaucoup de systèmes de détection des propositions proprement dit, et même un seul à notre connaissance : CBAP. Cependant, il existe des analyseurs syntaxiques partiels permettant de détecter les propositions sans être munis de véritable fonction de reconnaissance de propositions. Nous présenterons également deux systèmes robustes d'analyse syntaxique, qui pourraient sans doute fournir des résultats satisfaisants : KNP et CaboCha.

10.2.1 Détecteur de propositions CBAP

Étant donné l'absence de système détectant tous les types de propositions, Maruyama et al. (2004) proposent le programme CBAP (*Clause Boundaries Annotation Program*) qui réalise la détection de la totalité des frontières de propositions présentes dans une phrase japonaise et l'identification de leur type.

Le système reçoit comme entrée le résultat d'une analyse morphologique, et détecte 147 types de frontières de proposition par la reconnaissance de motifs représentant certaines séquences locales, de une à trois unités morphologiques.

CBAP a obtenu un rappel de 97,15 à 99,57% selon les corpus, lors d'une évaluation avec cinq corpus (transcription d'émissions, recueil d'articles d'un journal télévisé, articles de journaux, transcription de modèles de conversation, expressions de base pour le voyage).

Les erreurs se distinguent en trois types : erreurs dues au résultat erroné de l'analyse morphologique, celles dues à l'absence de définition des motifs correspondants et celles liées aux cas difficiles pour tout traitement automatique.

Les erreurs sont principalement dues à des résultats erronés de l'analyse morphologique, mais il y a tout de même 57 cas d'erreurs liés à l'absence de définition de structure. Les auteurs parlent d'une amélioration assez aisée de ce défaut par l'ajout de définitions et modifications des règles. Cependant, la construction d'une liste totalement exhaustive étant extrêmement difficile voire impossible, ce type d'erreur persistera toujours même avec l'enrichissement de la liste par entraînement avec des corpus de grande taille et de différente nature.

Un cas difficile est, par exemple, celui de la coordination de plusieurs prédicats avec l'ellipse des copules ou des parties variables. Les qualificatifs aux formes qualificative et adverbiale posent également des problèmes d'ambiguïté très difficiles à résoudre. Ils peuvent être interprétés soit comme le prédicat de la proposition, soit comme le complément ou le complément secondaire. Étant donné que s'ils jouent effectivement le rôle de prédicat, la limite de proposition doit suivre cet élément, l'identification de leur fonction a une grande influence sur le résultat de la détection.

Les motifs définissant une suite d'unités morphologiques d'une zone très restreinte permettent également, d'après les auteurs, un traitement efficace des corpus oraux qui ne contiennent pas de ponctuation. Mais cette non prise en compte de la ponctuation peut représenter également un défaut pour les corpus écrits : certains signes de ponctuation sont justement là pour marquer la frontière entre deux propositions.

Dernière remarque intéressante : ce système, et la plupart des autres présentés précédemment, définit une frontière après la particule *wa*, c'est-à-dire après le syntagme thématisé. Ce qui signifie qu'ils considèrent le syntagme thématisé comme une unité de même niveau qu'une proposition, que ce soit pour des raisons linguistiques ou purement pratiques. CBAP, conçu sur la base des travaux de syntacticiens comme Mikami et Minami, exclut également de la proposition certains éléments tels que les interjections ou les marqueurs discursifs.

La problématique de ce système unique dans son genre précédant nos travaux a été examinée avec notre propre évaluation. Le résultat de ces études sera présenté dans la section 11.1.

10.2.2 Analyseur syntaxique KNP

Le système d'analyse syntaxique du japonais KNP *(Kurohashi-Nagao Parser)* (Kurohashi, 2000 ; Kurohashi & Nagao, 1991, 1992) a été développé à l'Université de Kyoto.

Le système KNP reçoit comme entrée un texte segmenté en mots et analysé morphologiquement. Il réalise le regroupement des mots en syntagmes minimums, *bunsetsu*, et la détermination des relations de dépendance entre ces unités.

Sa caractéristique principale se trouve dans sa capacité de détection des structures de coordination (syntagmes ou propositions) et ce sans utiliser, à la différence de la plupart des autres systèmes, aucun algorithme « classique » d'analyse syntaxique.

Kurohashi et Nagao posent comme hypothèse que les constructions de coordination possédant une certaine ressemblance, il est possible de traiter ces structures en détectant leur ressemblance. Cependant, selon eux « détecter leur ressemblance avec les méthodes traditionnelles ne peut pas être une solution ». Ils ont alors développé l'analyseur KNP, capable d'analyser ces structures de coordination, inspiré du principe de correspondance des méthodes de programmation dynamique, utilisées largement en traitement de la parole.

KNP interprète bien le schéma global des relations entre les syntagmes. En particulier, l'analyse de certaines relations de coordination entre syntagmes et de relations entre la proposition principale et la proposition subordonnée relative est un travail considérable de KNP.

Une étude plus détaillée de cet analyseur est réalisée dans Nakamura-Delloye (2003a).

10.2.3 Analyseur des relations de dépendance CaboCha

L'analyseur CaboCha (Kudo & Matsumoto, 2002) calcule, tout comme le système KNP, les relations de dépendance entre les syntagmes de la phrase japonaise. Mais contrairement à ce premier basé sur les règles écrites à la main, le système CaboCha recourt à une méthode statistique.

Le système utilise pour l'analyse des relations la méthode d'analyse en cascade, algorithme utilisé pour le *chunking* par Abney et décrit précédemment (voir la section 8.2.2).

Tous les éléments de la phrase (séquence de *chunks*) sont d'abord étiquetés par *O*, tag « indéfini ». Lors de la première phase, tous les *chunks* considérés comme dépendant du *chunk* les suivant directement sont étiquetés par le tag *D*. Au cours de la seconde phase, tous les *chunks* suivant directement le *chunk* étiqueté par *O* sont supprimés. On réalise ces opérations de manière itérative jusqu'à ce qu'il ne reste qu'un seul *chunk*.

Cet algorithme exploitant efficacement le caractère « centripète accusé » (les régis précèdent toujours les régissants) de la langue japonaise, est très simple et facile à implémenter. De plus, contrairement aux méthodes classiques – qui calculent d'abord la probabilité de dépendance entre toutes les combinaisons possibles de deux syntagmes pour obtenir l'ensemble des combinaisons les plus probables par programmation dynamique –, c'est une méthode plus efficace nécessitant moins de calcul.

L'algorithme n'impose aucune méthode pour la détermination de la probabilité de dépendance entre deux syntagmes. Mais, pour CaboCha, un algorithme d'apprentissage machine, *Support Vector Machine* (SVM) est utilisé.

Les auteurs ont comparé le résultat de leur système avec celui obtenu par une méthode classique (Kudo & Matsumoto, 2000). Alors que la qualité d'analyse des structures est légèrement meilleure (taux de réussite de 89,29% contre 89,09%), ils ont obtenu une amélioration considérable du temps de calcul (0,5 sec/phrase contre 2,1) et du temps d'apprentissage (10 heures environ contre 2 semaines).

10.2.4 Possibilité d'utilisation d'un analyseur des relations dépendancielles pour la détection des propositions

Ces deux analyseurs syntaxiques, KNP et CaboCha, réalisent non pas une analyse syntaxique au sens classique du terme (à savoir une analyse en constituants et une résolution de leur fonction syntaxique), mais la détermination des relations de dépendance entre les syntagmes minimums ou les *chunks*. L'avantage de cette analyse est la robustesse : les deux systèmes fournissent toujours une réponse quelle que soit la phrase étudiée. Cette solution semble en fait particulièrement adaptée à l'analyse d'une langue telle que le japonais dépendant fortement du contexte, pour laquelle tous les éléments de phrase, y compris le nominatif, mais sauf le prédicat (ou le mot variable) sont facultatifs et susceptibles d'être omis.

Les systèmes fournissent leurs résultats aussi bien sous un format facile à trai-

ter par une machine que sous la forme d'un arbre de dépendance. Définir les types d'unités pouvant constituer le noyau de la proposition, et expliciter les informations sur les relations de dépendance entre les syntagmes et sur chaque unité composant les syntagmes, permettrait sans doute une détection relativement facile des propositions qui composent la phrase.

CHAPITRE 11

NOTRE SYSTÈME DE DÉTECTION AUTOMATIQUE DES PROPOSITIONS JAPONAISES : SIGLé JP

Nous présentons dans le présent chapitre la réalisation d'un système de détection automatique des propositions japonaises, basé sur nos études linguistiques. Nous abordons tout d'abord les problèmes du système existant CBAP (§ 11.1), puis nous proposons une solution par l'utilisation d'un système d'analyse des relations dépendancielles (§ 11.2). Nous passons ensuite à l'exposé sur le fonctionnement général du système (§ 11.3) avant de présenter la description détaillée de chaque opération : pré-traitement des séquences entourées de parenthèses (§ 11.4), attribution des traits morpho-syntaxiques aux chunks (§ 11.5), regroupement des chunks (§ 11.6), reconstitution finale des propositions et détermination de leur type (§ 11.7) et interface pour l'affichage du résultat (§ 11.8). Le chapitre se terminera par l'évaluation de notre système (§ 11.9) suivie d'une brève conclusion et des perspectives (§ 11.10).

11.1 Problèmes du système existant

Pour la détection des propositions japonaises, nous avons décidé de profiter pleinement des systèmes existants. Nous avons donc réalisé notre propre évaluation, avec le corpus Yomiuri (cf. Liste des corpus utilisés (page 550)), du système CBAP (Maruyama et al., 2004) afin de vérifier son efficacité et de déterminer d'éventuels problèmes que nous pourrions rencontrer lors de son utilisation dans le cadre de l'alignement des propositions.

11.1.1 Résultat de notre évaluation du système CBAP

Nous avons constaté trois types d'erreurs :

1. détection erronée due aux erreurs de l'analyse morphologique ;
2. frontières finales non détectées ;
3. impossibilité de détection des frontières initiales ;

Outre ces erreurs, il existe beaucoup de cas délicats liés aux séquences plus ou moins figées et lexicalisées.

Détection erronée due aux erreurs de l'analyse morphologique

Comme les auteurs le signalent après l'évaluation de leur système (cf. § 10.2.1), il existe des erreurs provenant des résultats erronés de l'analyse morphologique.

En particulier, nous constatons beaucoup de détections erronées des frontières après la particule *de* étiquetée incorrectement comme la copule à la forme adverbiale (8 phrases sur 175), et les cas inverses, c'est-à-dire les cas d'absence de détection de frontières après la copule à la forme adverbiale *de*, incorrectement analysée comme la particule *de* (3 phrases sur 175).

Frontières finales non détectées

Il existe deux autres cas où la fin de phrase n'est pas détectée : l'absence de copule et les subordonnées déterminantes avec prédicat adjectif. Ces erreurs sont également signalées par les auteurs dans leur article.

Le premier type concerne les propositions terminées par un substantif ou le radical d'un qualificatif en *na* (5 phrases sur 175). Dans ces propositions, la copule ou la partie variable étant omise, le système ne sait pas détecter leur fin.

Le second type concerne les propositions déterminantes avec prédicat adjectif (2 phrases sur 175). Étant donnée leur forme identique, la distinction entre le prédicat adjectif constituant la proposition déterminante et le qualificatif seul à l'emploi déterminant est délicate. Généralement, comme le font les auteurs de CBAP, on se contente de considérer le qualificatif ayant un ou plusieurs compléments comme le prédicat d'une proposition (ex. 1) et le qualificatif sans complément comme le qualificatif seul à l'emploi déterminant (ex. 2).

1. 悪い　学生 (*warui - gakusei*)
   (mauvais - élève) « élève délinquant »
2. 成績が　悪い　学生 (*seiseki ga - warui - gakusei*)
   (résultat [ga] - mauvais - élève) « élève dont les notes sont mauvaises »

Mais, comme l'expliquent les auteurs, la distinction selon la présence ou non des compléments est impossible pour la méthode adoptée, basée sur l'expression régulière exploitant uniquement des contextes immédiats.

Problème de l'impossibilité de détection des frontières initiales

Ce sont les erreurs les plus importantes et les plus gênantes pour notre traitement ultérieur.

En effet, tout comme d'autres méthodes basées sur l'expression régulière, le système ne détecte qu'une seule extrémité de la proposition, à savoir la frontière finale (la frontière initiale dans le cas de la détection des propositions en français ou en anglais), cette extrémité étant considérée comme le début de la proposition suivante (la fin de la proposition précédente dans le cas du traitement en français ou en anglais).

Comme nous l'avons déjà vu dans l'état de l'art des méthodes de détection pour le français et l'anglais, cette hypothèse entraîne une analyse erronée des propositions imbriquées. Par exemple, dans la phrase :

現在、多くの　国公立、私立　大学が
(*genzai - ôkuno - kokkôritsu - shiritsu - daigaku ga* -)
(actuel - grand nombre de - national et publique - privé - université [ga] -)
社会人　でも　受講できる　公開講座を　設けている。
(*shakaijin - demo - jukô dekiru* - *kôkaikôza wo - môketeiru*)
(personne travaillant - même - pouvoir assister - cours ouverts [wo] - installer [état])

« Aujourd'hui, beaucoup d'universités, publiques ou privées, proposent des cours ouverts auxquels les personnes travaillant peuvent assister. »

où la proposition subordonnée « personnes travaillant peuvent assister » déterminant le SN « cours ouverts » est enchâssée entre le syntagme en *ga* (nominatif) et le syntagme en *wo* (accusatif), le système ne détecte que la fin de la subordonnée et la phrase est segmentée en deux : la subordonnée déterminante comportant non seulement son complément mais aussi les compléments du prédicat, et la racine avec uniquement son complément accusatif qui apparaît après la limite finale de la subordonnée.

Cet enchâssement ne se limitant pas à la subordonnée déterminante, le même type d'erreur est constaté dans d'autres subordonnées : complétives et subordonnées adverbiales.

Les guillemets n'étant pas pris en compte, lorsque le discours rapporté est constitué de plusieurs propositions, la phrase est également mal segmentée.

Le syntagme thématisé en *wa* est également extrait mais lorsque d'autres éléments de phrases le précèdent, la segmentation n'ayant lieu qu'à la fin du syntagme thématisé, les éléments précédents sont inclus dans le syntagme thématisé.

Dans les résultats de CBAP, nous constatons un grand nombre d'erreurs de ce type. Cette absence de détection des frontières initiales est due sans doute à l'objectif fixé par ses auteurs. En effet, le système a été développé en vue du traitement des langues orales et présente comme avantage la possibilité d'un traitement en temps-réel. Pour cette application visée, l'identification des frontières initiales était probablement une opération trop lourde qui risquait d'annuler complètement le caractère opérationnel du système.

Toutefois, dans notre utilisation, la détection des deux extrémités de propositions est indispensable et l'amélioration du système serait inévitable si nous décidions d'employer le système CBAP pour la détection des propositions du japonais dans notre chaîne de traitement de l'alignement.

Problème des séquences plus ou moins figées

Enfin, ce dernier problème est le point le plus délicat et le plus difficile dépendant étroitement de la définition de la proposition.

Comme nous l'avons déjà abordé dans les études linguistiques (cf. § 7.4), certaines formes des mots variables – notamment les formes neutres des verbes – constituent parfois des locutions ou des expressions plus ou moins figées. Dans ces séquences, les mots variables perdent souvent leur fonction prédicative (ou plutôt leur capacité phrasogénératrice selon la terminologie usuelle de la linguistique japonaise), mais la détermination de l'absence ou non de cette fonction est très difficile, ce qui pose des problèmes cruciaux lors de la définition de la proposition en japonais.

Le système CBAP considère, simplement, comme des locutions les séquences avec un mot variable définies comme non indicatrices de la frontière par la grammaire du système. Mais la liste étant incomplète, nous constatons beaucoup de propositions extraites qui ne semblent pas conformes au statut de « proposition ».

Par ailleurs, il existe en japonais beaucoup de substantifs, peu autonomes, qui deviennent des éléments tout à fait autonomes lorsqu'ils sont déterminés par une subordonnée déterminante, et qui constituent une subordonnée « intégrative »[1] substantive ou adverbiale. Nous avons regroupé ces substantifs avec d'autres unités similaires sous le nom de *kyûchakugo*, mots agglutinants, dans les études linguistiques (cf. § 7.8).

Mais, contrairement au français dans lequel seuls les mots « qu- » ont la possibilité de jouer le rôle de cheville dans la structure de la subordination, en japonais, ces substantifs ne sont pas limités à un type particulier et il est difficile de constituer une liste fermée de ces substantifs dits « formels ».

Ainsi, dans les résultats du système, nous constatons des substantifs formels, laissés en dehors des subordonnées déterminantes qui les précèdent, constituant non pas une subordonnée intégrative, mais un substantif précédé par une subordonnée déterminante.

11.1.2 Difficultés pour l'adaptation à notre opération d'alignement

Comme nous venons de le voir, il resterait beaucoup d'améliorations à apporter au système CBAP pour pouvoir l'utiliser dans nos travaux de détection des

[1] Ces subordonnées sont intégratives dans le sens où la base, non autonome, des propositions déterminantes est « intégrée » dans ces dernières pour constituer une subordonnée substantive (ou adverbiale), contrairement aux subordonnées classiques dont la base reste dans la proposition racine, seules les déterminantes étant extraites comme des subordonnées.

propositions pour l'alignement :

1. amélioration permettant d'analyser correctement les structures enchâssées ;
2. élaboration d'une liste et/ou de règles permettant de déterminer le plus possible les locutions figées comprenant les mots variables ;
3. constitution d'une liste plus complète des substantifs formels et des règles permettant de les inclure dans la subordonnée déterminante qui les précède pour constituer ensemble une subordonnée « intégrative ».

De plus, pour aligner les propositions, nous sont également nécessaires des informations sur les relations entre les propositions constituant la phrase. Il faudrait donc également créer une fonction de résolution des relations.

Ainsi, l'amélioration de ce système nous demanderait plus de travail que nous ne pouvons raisonnablement envisager dans le cadre de la présente thèse : pour identifier correctement les frontières initiales des propositions enchâssées, il nous faudrait au moins une analyse syntaxique partielle.

11.2 Solution aux problèmes par l'utilisation d'un analyseur syntaxique

Nous allons maintenant proposer une solution par l'utilisation d'un analyseur syntaxique. Nous énumérons tout d'abord les problèmes, avant de montrer comment détecter les propositions à partir du résultat de CaboCha.

11.2.1 Problèmes à résoudre

Les problèmes sont communs avec ceux déjà posés lors de l'évaluation de CBAP :

1. distinction des syntagmes à mot variable que l'on peut considérer comme des propositions, des autres ;
2. traitement des substantifs peu autonomes qui constituent, en étant déterminés par une subordonnée déterminante, une subordonnée « intégrative ».

Pour résoudre ces problèmes, il faut :

1. définir les règles permettant de distinguer les syntagmes à mot variable propositionnels et non-propositionnels ;
2. construire la liste des substantifs formels constituant des subordonnées « intégratives ».

Ces résolutions ne sont pas des opérations aisées, mais contrairement à CBAP, CaboCha permet théoriquement :

- de détecter les subordonnées même imbriquées ;
- de construire, puisqu'il fournit les informations sur les relations de dépendance entre les chunks, un graphe de relations.

11.2.2 Méthode de détermination des propositions à partir du résultat du système CaboCha

La figure 11.1 présente le résultat de l'analyse d'une phrase par le système CaboCha. Le système fournit comme résultat la liste des *chunks* constituant la phrase avec leurs relations de dépendance (les lignes commençant par un symbole « * » dans la figure). Le résultat comporte également les informations sur les unités morpho-lexicales que contient chaque *chunk* (les lignes sans « * » suivant celle contenant les informations sur le *chunk*; chaque ligne comporte le résultat de l'analyse morphologique d'une unité morpho-lexicale constituant le *chunk*).

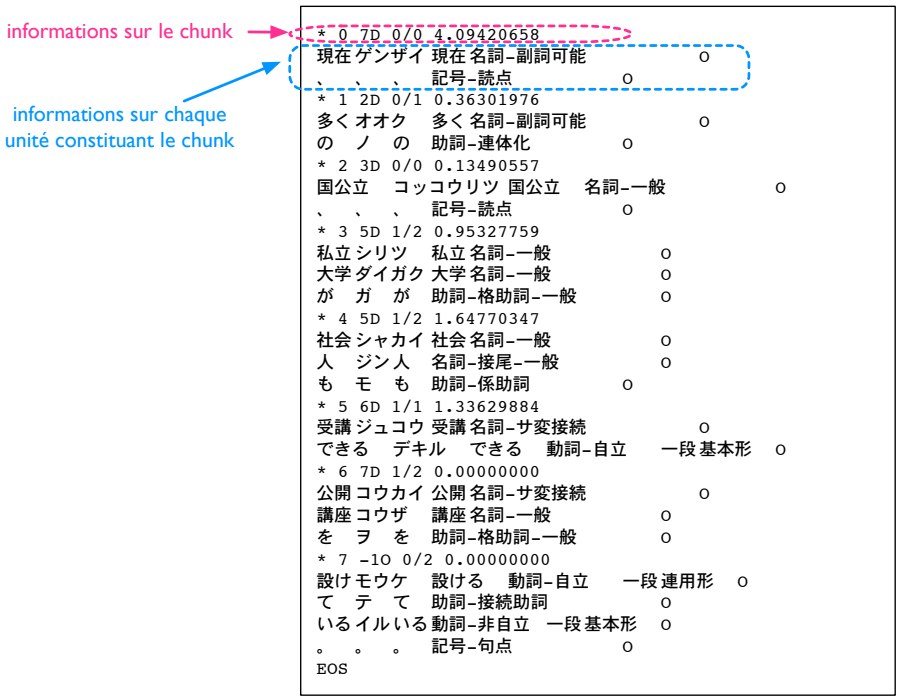


FIG. 11.1 – Résultat d'analyse par CaboCha I

Pour faciliter la lecture du résultat, on peut le représenter sous forme d'un graphe comme dans la figure 11.2 page suivante.

On peut constater que la dépendance du *chunk* étiqueté 3 est mal analysée : son arc de dépendance se dirige vers le *chunk* 5, prédicat de la subordonnée déterminante, alors qu'il doit s'orienter vers le *chunk* 7, prédicat principal. La figure 11.3 page ci-contre montre le graphe correspondant au résultat correct.

Supposons maintenant que les résultats du système CaboCha soient corrects et réfléchissons comment nous pourrions extraire les propositions à partir de ses résultats.

Le premier repère pour la proposition japonaise est le mot variable : c'est lui

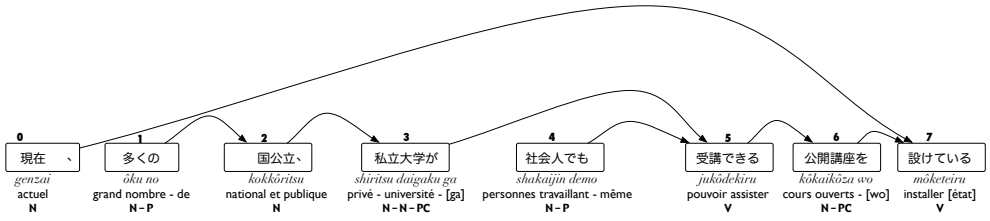


FIG. 11.2 – Graphe représentant le résultat d'analyse par CaboCha

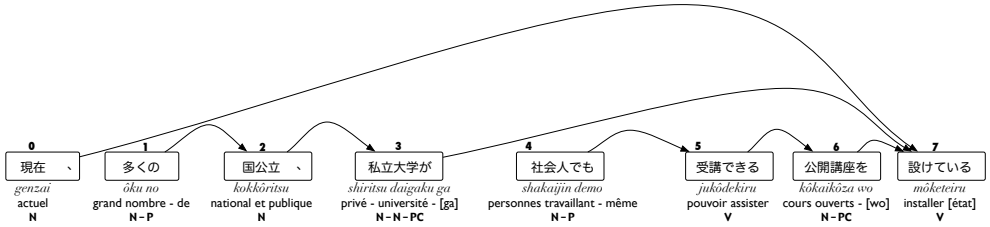


FIG. 11.3 – Graphe correspondant au résultat correct

qui est l'élément principal du prédicat et le prédicat est le seul élément obligatoire dans la proposition japonaise. On extrait donc le *chunk* contenant un mot variable et tous les *chunks* qui dépendent, directement ou indirectement, de lui. Ainsi, on peut détecter les syntagmes à mot variable avec leurs compléments, même dans les structures enchâssées (cf. figure 11.4).

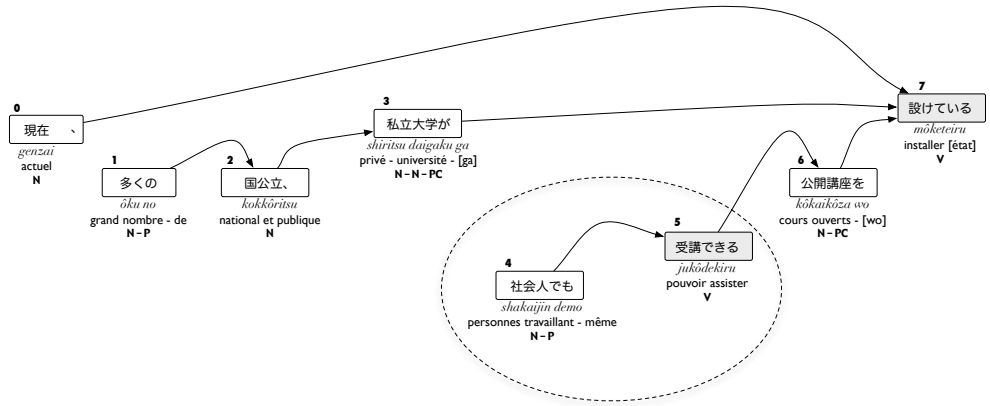


FIG. 11.4 – Détection des propositions à partir du résultat de CaboCha

11.2.3 Solutions aux deux autres problèmes

Pour réaliser l'opération de détection des propositions à partir des résultats de CaboCha, deux problèmes liés à la définition même de la proposition doivent être résolus.

En nous appuyant sur nos études linguistiques, nous proposons les solutions suivantes.

Distinction entre les syntagmes à mot variable propositionnels et non-propositionnels

Nous avons déjà défini dans les études linguistiques (cf. § 7.4) trois règles permettant la détermination des syntagmes à mot variable non-propositionnels. Les première (pour les mots variables supports ou auxiliaires) et deuxième (pour les locutions figées) règles servent à la détermination préalable de ces syntagmes afin de constituer une liste cohérente.

En revanche, la troisième est dédiée à la reconnaissance automatique des syntagmes à mot variable non-propositionnels. Nous nous basons sur cette règle pour la détection des propositions. Rappelons la règle (définie dans § 7.4.3) :

Règle 3 (pour l'identification dynamique) Lorsqu'un syntagme terminé par un mot variable à une forme neutre ne comprend aucun complément, il est considéré comme un syntagme non-propositionnel dépendant du prédicat apparaissant à une position postérieure. Lorsqu'un syntagme terminé par un mot variable conclusif ne comprend aucun complément et qu'il est précédé directement par un mot variable à une forme neutre, il est considéré comme constituant un mot variable composé avec celui qui le précède directement.

Liste des substantifs formels constituant des subordonnées « intégratives »

Nous avons déjà défini les connecteurs appelés agglutinants dans les études linguistiques (cf. § 7.8.2). Ce sont des mots suivant une forme autonome, qui ne sont ni substantifs autonomes, ni auxiliaires, ni particules conjonctives. Ces connecteurs, éventuellement suivis de particules, constituent, selon notre typologie, une subordonnée dite avec connecteur agglutinant.

Sur le plan pratique, nous considérons comme mots agglutinants, les unités définies dans ipadic[2] comme substantifs non autonomes et particules adverbiales, ainsi que les mots cités comme *kyûchakugo* par Sakuma (1940b).

[2] Dictionnaire électronique pour le TAL utilisé par l'analyseur morphologique ChaSen que nous utilisons dans notre réalisation informatique.

11.3 Procédure générale

Notre système de détection des propositions est constitué de trois modules (implémentés en Perl) comme représenté figure 11.5.

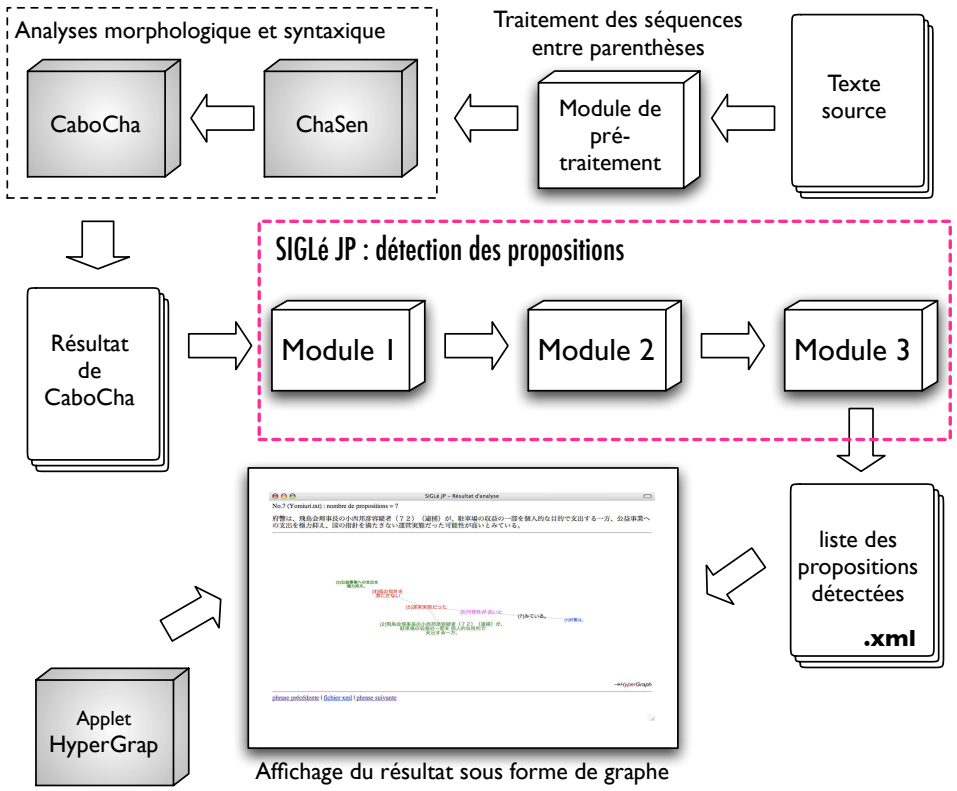


FIG. 11.5 – Procédure générale du système de détection des propositions SIGLé JP

11.3.1 Prétraitement

Analyses préparatoires par deux analyseurs extérieurs

Le pré-traitement consiste en deux tâches, analyse morphologique et analyse des relations dépendancielles entre les constituants, réalisées par deux systèmes extérieurs.

Notre système reconnaît les propositions à partir du texte source segmenté en *chunks* par le système CaboCha, qui analyse également leurs relations de dépendance syntaxique.

Mais, avant le *chunking* à l'aide de CaboCha, le texte source est tout d'abord segmenté et étiqueté morphologiquement par l'analyseur morphologique du ja-

ponais ChaSen[3].

Pré-traitement des séquences entourées de parenthèses

Ces deux analyseurs extérieurs tracent les frontières de la fin de phrase, en considérant comme séparateurs de phrase le point final japonais et le retour à la ligne. Cependant, le point final peut apparaître à l'intérieur d'une phrase : c'est le cas des séquences entourées de parenthèses ou de guillemets, constituées de plusieurs phrases. Afin de pouvoir segmenter correctement même les phrases contenant d'autres phrases, nous avons introduit un module de pré-traitement destiné à extraire toute séquence entourée de parenthèses ou de guillemets (description détaillée dans § 11.4).

11.3.2 Premier module

Ce module reçoit comme entrée le fichier résultant d'une analyse du système CaboCha (cf. figure 11.1 page 376). Il en extrait les informations nécessaires et attribue à chaque *chunk* des traits indiquant sa nature morpho-syntaxique (description détaillée dans § 11.5).

En outre, il réalise également la modification des étiquettes attribuées aux mots par l'analyseur morphologique.

Les étiquettes des mots agglutinants, cités par Sakuma, mais étiquetés comme substantifs autonomes, sont modifiées en substantifs non-autonomes. Par ailleurs, les mots variables, tels que *suru* ou *aru*, fonctionnant comme des auxiliaires ou des verbes de support lorsqu'ils se mettent derrière une proposition terminée par un connecteur agglutinant sont également modifiés et marqués comme mots variables de support. La liste des mots modifiés par le module 1 est présentée dans l'annexe D.1.

11.3.3 Deuxième module

Le deuxième module reçoit les résultats du premier module et réalise le regroupement des *chunks*, en fonction de leurs traits morpho-syntaxiques attribués par le premier module, de manière à obtenir les segments composés des constituants continus de la proposition (description détaillée dans § 11.6).

11.3.4 Troisième module

Le dernier module finalise la reconstitution de la proposition par le regroupement notamment des constituants discontinus. Cette dernière opération comporte également l'insertion des « thèmes faibles » dans la proposition. À cet effet,

[3] L'analyseur morphologique du japonais ChaSen est développé par l'équipe du *Computational Linguistics Laboratory* du NAIST (*Nara Institute of Science and Technology*) et est un logiciel libre disponible sur `http://chasen.naist.jp/hiki/ChaSen/`

le module réalise d'abord une nouvelle analyse des *chunks* en *wa* pour déterminer leur nature selon leur contexte syntaxique.

Ce dernier module fournit finalement la liste des propositions ainsi reconstituées avec les informations sur leur type et leurs relations, au format xml.

Le type de chaque proposition détectée (ou regroupée) est déterminé par ce troisième module, selon différents traits attribués à chaque proposition par le module précédent (description détaillée dans § 11.7).

11.3.5 Interface pour l'affichage du résultat sous forme de graphe

En outre, le système offre également la possibilité d'afficher le résultat sous un format plus convivial. Le fichier résultat est transformé par un script en un format xml adéquat, permettant l'affichage du résultat sous forme d'un graphe à l'aide de l'applet JAVA HyperGraph, dans une fenêtre d'un navigateur Internet (description détaillée dans § 11.8).

11.4 Pré-traitement : extraction des séquences entre parenthèses ou entre guillemets

Bien qu'ils constituent rarement un sujet de préoccupation des chercheurs, les parenthèses, les guillemets ou d'autres symboles de ponctuation, sont des éléments syntaxiques très importants, surtout dans les applications du TAL. Il n'existe quasiment aucun type de texte qui puisse être correctement analysé sans un traitement préalable (ou simultané) de ces signes typographiques.

Dans le cas de la segmentation en phrases, bien que le point final japonais soit extrêmement fiable par rapport au point final français ou anglais – qui, très polysémique, rend l'opération de segmentation en phrases très complexe –, segmenter après un symbole séparateur, point final japonais ou retour à la ligne, n'est pas suffisant pour obtenir l'ensemble correct des phrases constituant le texte source.

Le point final peut apparaître à l'intérieur d'une phrase. Il existe en effet beaucoup de phrases où apparaissent des séquences entourées de parenthèses ou de guillemets, constituées de plusieurs phrases.

Nous avons donc réalisé un module de pré-traitement qui extrait tous les éléments entre parenthèses ou guillemets, de manière à remplacer ces ensembles par une sorte de boîte noire, pour que les analyseurs extérieurs réalisent des analyses en considérant comme un symbole les éléments entre parenthèses et comme un SN les éléments entre guillemets.

11.4.1 Problème de la segmentation en phrases

Considérons la phrase suivante :

首相は　官邸で　記者団に対し「教育の再生は　私の内閣で最も重要な課題だ。その課題に向かって　しっかりと　政策を作っていく　ために　改

正は　重要だ」と　語った。
shushô wa - kantei de - kishadan nitaishi - "kyôiku no saisei wa - watashi no naikaku de mottomo jûyôna kadaida. - sono kadaini mukatte - shikkarito - seisakuwo tsukutteiku - tameni - kaisei wa - jûyôda" - to - katatta
Premier ministre [*wa*] - dans la résidence officielle du premier ministre - vis à vis des journalistes - "renaissance de l'éducation [*wa*] - la question la plus importante pour mon gouvernement. - face à cette question - de manière fiable - aller créer des projets politiques - afin de - modification [*wa*] - important - [citation] - raconter [passé]
« Le premier ministre a affirmé dans sa résidence officielle devant les journalistes : "La renaissance de l'éducation est la question la plus importante pour mon gouvernement. Afin de créer des projets politiques dans ce but, une amélioration est importante." »

Cette phrase comprend un discours direct correspondant au commentaire du premier ministre. Le discours inséré est constitué lui-même de deux phrases dont la première est terminée par un point final.

Lorsqu'on la soumet à des analyseurs qui reconnaissent les frontières des phrases à l'aide des points finaux et des retours à la ligne, cette phrase est segmentée en deux phrases, séparées par le point final de la première phrase du discours entre guillemets, entraînant ainsi une mauvaise segmentation produisant deux phrases mal formées.

11.4.2 Extraction des séquences entourées de parenthèses

Afin d'éviter cette erreur de segmentation en phrases, notre module de prétraitement extrait tous les éléments entre parenthèses ou guillemets de manière à remplacer ces ensembles par une sorte de boîte noire. La séquence extraite est traitée, elle-même de manière récursive, pour être segmentée en phrases. Elle est remplacée dans la phrase initiale par des symboles servant à indiquer la phrase correspondante extraite, et à l'aide desquels on peut retrouver la position d'apparition initiale des phrases extraites.

Ainsi, la phrase d'exemple est segmentée comme montré dans la figure 11.6 page ci-contre.

La phrase indexée 79 contient une séquence entourée de guillemets. Cette séquence est extraite et segmentée elle-même en deux phrases, indexées @1 et @2 ayant comme père la phrase 79. Dans la phrase initiale, on trouve également les symboles /@1/ et /@2/, indiquant la position où apparaissent ces deux phrases extraites.

11.4.3 Analyse postérieure par des systèmes extérieurs

Une fois que ces séquences sont extraites et remplacées par des symboles, l'analyseur morphologique considère comme des symboles les éléments entre parenthèses et comme des SN les éléments entre guillemets. Cette analyse est réalisable grâce à la possibilité de définition par l'utilisateur de règles propres pour certaines séquences, option fournie par l'analyseur ChaSen. Nous avons à cet effet

```
<s id='79'>
  首相は官邸で記者団に対し「/@1//@2/」と語った。
</s>
<s id='@1' pere='79'>
  教育の再生は私の内閣で最も重要な課題だ。
</s>
<s id='@2' pere='79'>
  その課題に向かってしっかりと政策を作っていくために改正は重要だ
</s>
```

FIG. 11.6 – Résultat de la segmentation par le module de pré-traitement

défini tout simplement, dans le fichier .rc, les règles pour les séquences entourées de guillemets ou de parenthèses.

11.4.4 Réinsertion des séquences extraites

Les relations entre les phrases initiales et leurs séquences extraites ne sont prises en compte qu'après la détection des propositions. Mais une fois que la reconnaissance des propositions est terminée, les séquences extraites sont réintégrées dans la phrase : les syntagmes non-propositionnels sont insérés à leur position initiale dans la phrase ; les propositions constituant les séquences extraites sont incluses dans la liste des propositions de la phrase initiale.

11.5 Détermination des traits morpho-syntaxiques des *chunks*

Le premier module extrait du résultat de CaboCha, reçu en entrée, les informations nécessaires tout en attribuant à chaque *chunk* des traits indiquant sa nature morpho-syntaxique. Les traits du *chunk*, qui servent au regroupement des *chunks*, c'est-à-dire à la reconstitution des propositions, sont déterminés selon la nature des mots constituant le *chunk*.

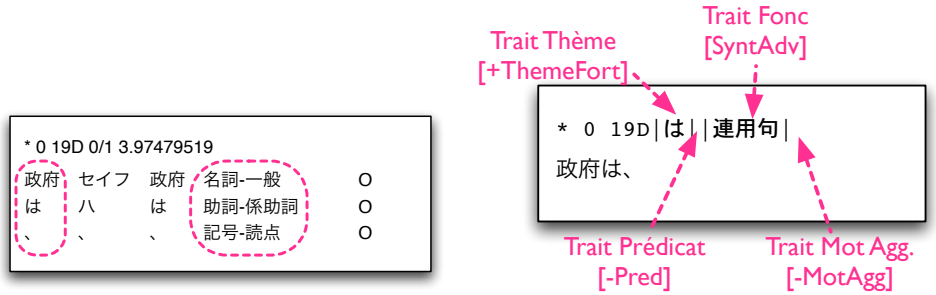
11.5.1 Principe de la méthode de détermination des traits

L'algorithme de détermination des traits consiste à lire les lignes correspondant aux étiquettes des constituants du *chunk* une par une dans l'ordre de leur apparition (c'est-à-dire dans le même ordre que celui de leur occurrence dans la réalisation linéaire). Chaque fois qu'une nouvelle ligne est lue, les traits morpho-

syntaxiques que le constituant concerné donne au *chunk* qu'il constitue sont déterminés, et les variables dédiées au stockage des traits du *chunk* sont mises à jour.

Cette méthode, très simple, profite en fait d'une particularité du japonais : en japonais, les éléments décisifs de la fonction syntaxique jouée par leur syntagme se situent toujours en position postérieure, les unités pouvant apparaître derrière eux étant extrêmement limitées, à savoir les particules de mise en relief.

La figure 11.7 montre le résultat de la détermination des traits d'un *chunk* (figure 11.7(b)) à partir d'un résultat de *chunking* (figure 11.7(a)). Le *chunk* à considérer est constitué de trois éléments : le substantif-commun(名詞-一般,*meishi-ippan*) « 政府 (*seifu*, gouvernement) » est suivi d'une particule-particule *kakari* (助詞-係助詞, *joshi-kakarijoshi*) « は (*wa*, [thème]) », le chunk se terminant par une virgule « 、 » ([virgule]) étiquetée ponctuation-virgule (記号-読点, *kigô-tôten*). Par examen de chaque constituant, le *chunk* obtient finalement le trait du thème fort et le trait du syntagme adverbial.



(a) (b)

FIG. 11.7 – Détermination des traits d'un *chunk*

La procédure détaillée et des exemples d'application de la méthode sont présentés dans l'annexe D.2.

11.6 Premier regroupement des *chunks*

Le deuxième module reçoit les résultats du premier module et réalise, en fonctions de leurs traits morpho-syntaxiques attribués par le premier module, le premier regroupement des *chunks* qui consiste à regrouper les constituants continus de la même proposition.

11.6.1 Principe du regroupement des *chunks*

Le regroupement consiste en la fusion du *chunk* ou segment – unités résultant d'une fusion – avec un autre *chunk* ou segment qui le suit directement et qui dépend syntaxiquement de lui, et ce de manière itérative jusqu'à l'épuisement des possibilités de fusion.

La fusion des segments est réalisée non seulement en cas d'absence de relation de dépendance entre les deux segments considérés, mais aussi dans le cas où l'unité précédente est un élément considéré comme régissant principal de la proposition, en d'autres termes un élément marquant la fin de la proposition.

Fusions en cascade

La figure 11.8 (voir page suivante) montre le déroulement du regroupement selon cette méthode. Les cercles numérotés représentent des *chunks* ; les cercles colorés correspondent aux éléments marquant la fin de la proposition ; les arcs indiquent les relations de dépendance entre les *chunks*.

Nous comparons toujours l'unité considérée seulement avec l'unité suivante. Ainsi, au premier tour (voir le cadre 1 de la figure 11.8), les couples de chunks adjacents, 2-3, 4-5 et 6-7 sont regroupés. De plus, le *chunk* 8 est regroupé avec le segment précédent déjà fusionné, constituant finalement le nouveau segment 6-7-8.

Après la fusion, l'arc de dépendance partant du dernier *chunk* du segment résultant est attaché à ce nouveau segment. De même, tous les arcs atteignant un des chunks constituant le segment, sont également associés au segment qu'ils constituent (voir le cadre 1b de la figure 11.8).

Au second tour (cf. cadre 2 dans la figure 11.8), le segment 4-5 est fusionné avec le segment précédent 2-3 qui lui est relié directement par l'arc de dépendance. En revanche, le segment 6-7-8 ne peut pas être fusionné avec le nouveau segment 2-3-4-5 car ce dernier est terminé par l'élément marquant la fin de la proposition.

Ainsi, le graphe initial contenant huit nœuds est réduit à un autre avec seulement trois nœuds. Le *chunk* 1 n'a pas été fusionné alors que ce n'est pas un élément marquant la fin de la proposition. C'est typiquement le cas des phrases comportant une subordonnée enchâssée. Mais, dans le cas où le *chunk* isolé est effectivement un élément appartenant à la proposition (c'est-à-dire que ce n'est pas un élément externe), l'opération d'insertion de cet élément dans la proposition est réalisée par le module suivant, module 3.

La procédure détaillée et des exemples d'application de la méthode sont présentés dans l'annexe D.3.

11.7 Reconstitution finale des propositions et détermination de leur type

Le dernier module finalise la reconstitution de la proposition par le regroupement notamment des constituants discontinus et par l'insertion des « thèmes faibles » dans la proposition et fournit finalement la liste des propositions ainsi reconstituées avec les informations sur leur type et leurs relations au format xml.

Le module réalise donc trois tâches : la réanalyse des *chunks* en *wa*, le regroupement des constituants et la détermination du type de chaque proposition dé-

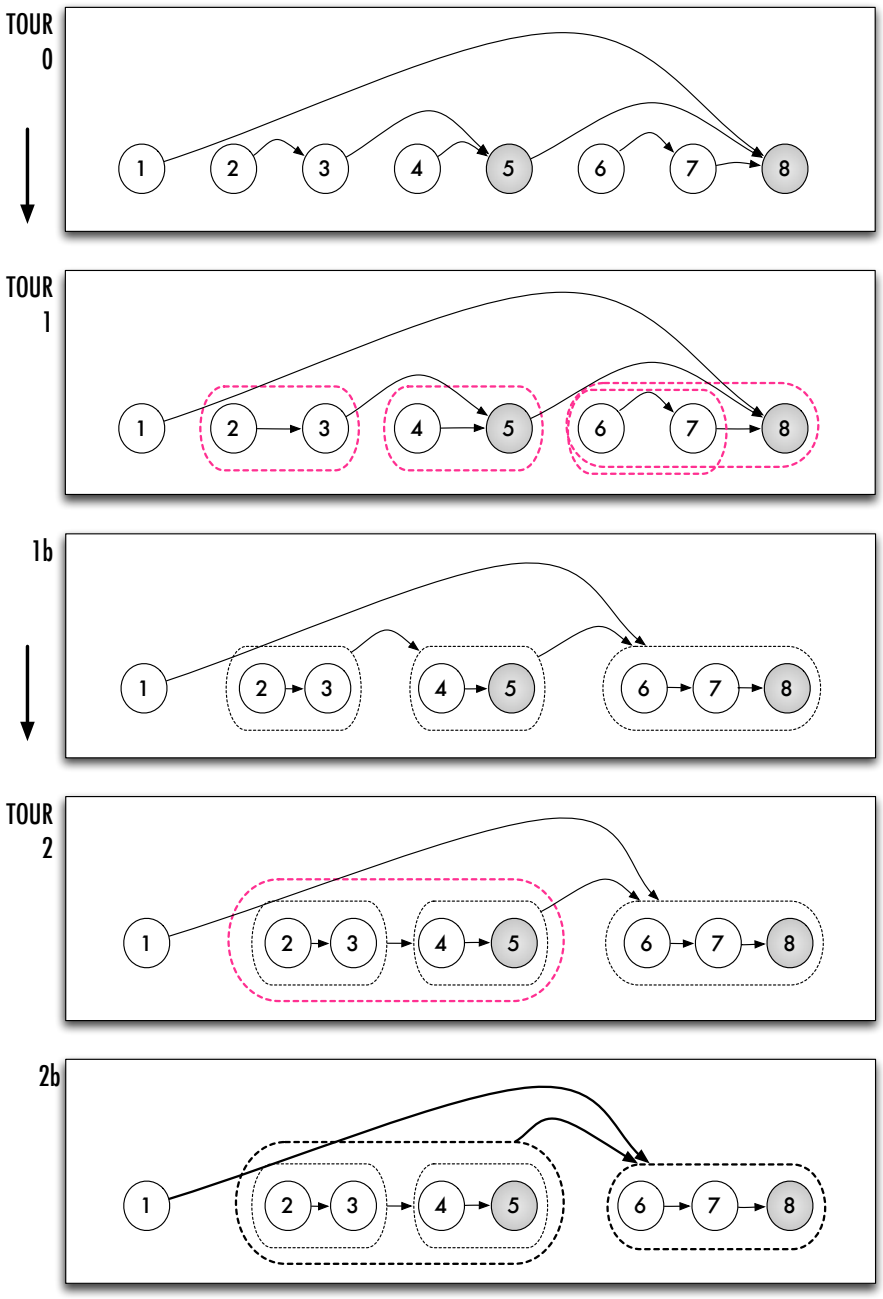


FIG. 11.8 – Principe du regroupement des *chunks*

tectée (ou regroupée).

11.7.1 Réanalyse des *chunks* en *wa*

Le trait « Thème » est déterminé par le premier module, et aux syntagmes en *wa* est déjà attribué le trait [+thème faible] ou [+thème fort] selon la présence ou non d'une particule de cas précédant la particule *wa*.

Mais, comme nous l'avons défini dans nos études linguistiques (cf. § 6.4.6), la nature du syntagme en *wa* change, non seulement selon les constituants internes des *chunks* ou segments, mais aussi selon leur contexte syntaxique.

Les règles de détermination du trait Thème des *chunks* liées aux conditions syntaxiques sont les suivantes :

1. les *chunks* en *wa*-fort apparaissant à l'intérieur de la portée d'un autre *wa*-fort sont des *chunks* en *wa*-faible à portée restreinte ;
2. les *chunks* en *wa*-fort n'apparaissent pas à l'intérieur de l'unité déterminée par une proposition déterminante : en d'autres termes, s'il existe une proposition déterminante située à une place antérieure à un *chunk* en *wa*-fort, et dépendant d'un élément situé postérieurement à ce dernier, ce *chunk* en *wa*-fort est un *chunk* en *wa*-faible ;
3. les *chunks* en *wa*-fort n'apparaissent pas à une place postérieure à un complément essentiel en *ga* ou en *wo* : en d'autres termes, s'il existe un syntagme en particule de cas *ga* ou *wo*, situé à une place antérieure à un *chunk* en *wa*-fort, et dépendant d'un élément situé postérieurement à ce dernier, ce *chunk* en *wa*-fort est un *chunk* en *wa*-faible.

Les syntagmes en *wa* sont ainsi définitivement distingués en deux types : thème-fort et thème-faible. Les thèmes forts sont considérés comme des éléments entrant en relation avec le reste de phrase ou la proposition qui les suit, et donc externes à la proposition. Ils sont alors extraits de la phrase, séparément de toutes les propositions constituant la phrase considérée, tandis que les thèmes faibles sont intégrés à l'intérieur d'une proposition.

Par ailleurs, tous les éléments précédant les thèmes forts ainsi déterminés sont considérés comme des éléments externes à la proposition et sont extraits de la phrase, séparément de toutes les propositions constituant la phrase considérée.

11.7.2 Regroupement des constituants

Le regroupement des constituants par ce dernier module concerne notamment les éléments discontinus. Le cas typique de dispersion des constituants est la séparation de deux segments par l'insertion d'une proposition déterminante. Par ailleurs, la réintégration des thèmes faibles est également réalisée dans cette étape.

Pour représenter le caractère non-linéaire de la structure d'une phrase, la présence d'une subordonnée extraite est marquée dans la proposition régissante par une indication entre crochets telle que :

abc [sub. déterminante] *def.*

Selon la méthode de représentation des relations de dépendance que nous avons définie dans la section 7.14, les relations entre la proposition et les éléments externes ou les thèmes forts sont marquées par un arc s'étendant des éléments externes ou des thèmes forts vers le prédicat de la proposition entrant en relation avec ces premiers. Mais, ces éléments n'étant pas des constituants de la proposition, leur présence n'est pas marquée dans la séquence textuelle de la proposition. De même, dans le cas où la sous-phrase racine est précédée par une sous-phrase subordonnée ayant son propre thème fort, leur relation est marquée par un arc, mais la présence de la sous-phrase précédente n'est pas marquée dans la séquence textuelle de la sous-phrase racine.

11.7.3 Détermination du type de proposition

Les propositions ainsi reconstituées sont toutes caractérisées pour les quatre traits (trait Thème, trait Prédicat, trait Fonction, trait Proposition à connecteur Agglutinant) par une valeur. À partir de la valeur de ces quatre traits, la détermination du type de proposition est réalisée.

La figure 11.9 page suivante montre le résultat fourni par le troisième module (cadre du bas) à partir du résultat du premier regroupement réalisé par le deuxième module (cadre du haut) pour l'analyse de la phrase :

治験は、新薬について、製薬会社が
厚労省に　承認申請する　際に
必要な　安全性、有効性の　データを　集める　ために　実施される。
(*chiken wa - shinyaku nitsuite - seiyaku gaisha ga*
- kôrôshô ni - shônin shinsei suru - sai ni
- hitsuyôna - anzensei - yûkôsei no - dêta wo - atsumeru - tameni - jisshi sareru)
(essai clinique [wa] - pour les nouveaux médicaments - les sociétés pharmaceutiques [ga]
- le Ministère de la santé et du travail [ni] - demander l'autorisation - à l'occasion de
- nécessaire - sécurité - efficacité [no] - données [wo] - collecter - afin de - réaliser)

« Des essais cliniques de médicaments sont réalisés pour les nouveaux produits, afin d'obtenir des données sur leur sécurité et leur efficacité, nécessaires lorsqu'une société pharmaceutique fait une demande d'autorisation auprès du Ministère de la santé et du travail. »

Les règles et la description détaillée d'application des règles pour l'exemple cité sont présentées dans l'annexe D.4.

11.8 Interface pour l'affichage du résultat

Le résultat fourni par le système SIGLéJP peut être affiché sous un format plus agréable à l'aide du logiciel libre HyperGraph[4].

[4] `http://hypergraph.sourceforge.net/`

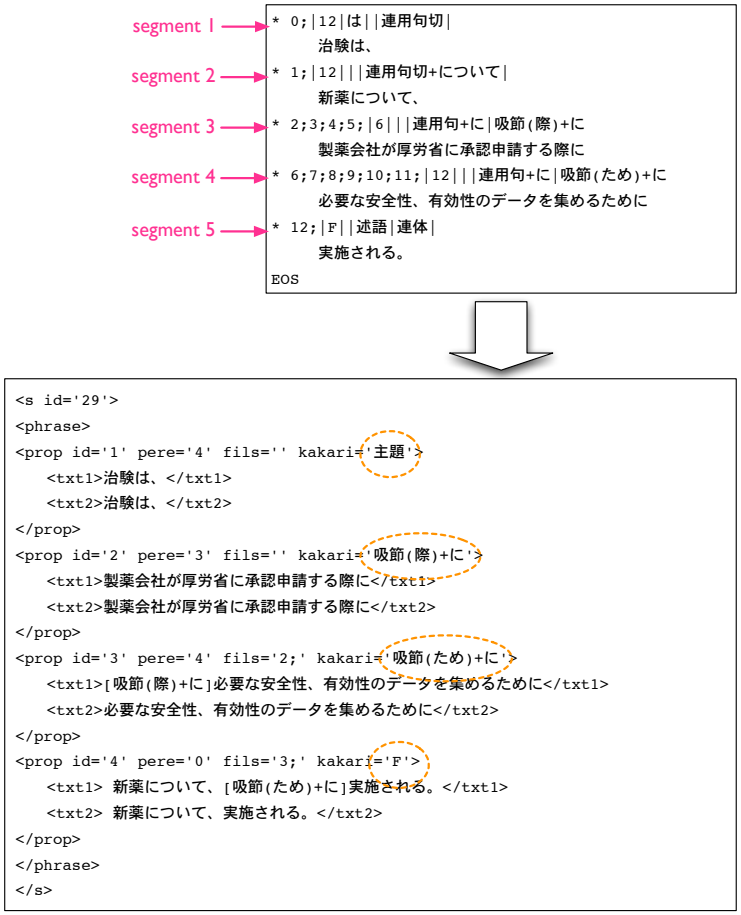


FIG. 11.9 – Exemple du résultat de la détermination du type de proposition

Le fichier résultat, une fois transformé au format xml adéquat par un script, permet d'afficher le résultat sous forme d'un graphe à l'aide de l'applet JAVA HyperGraph, dans la fenêtre d'un navigateur Internet comme montré figure 11.10 (voir page suivante).

Une telle représentation graphique facilite considérablement les traitements postérieurs permettant de mieux exploiter et de profiter des résultats : la vérification du résultat d'analyse est beaucoup plus facile, donc favorise la détermination des problèmes de l'analyse et d'éventuelles améliorations, non seulement pour le système lui-même mais aussi pour les traitements antérieurs à savoir les analyses morphologique et syntaxique ; cette représentation conviviale peut également être une grande aide à la description syntaxique, permettant ainsi de proposer aux linguistes un outil de recherche très efficace.

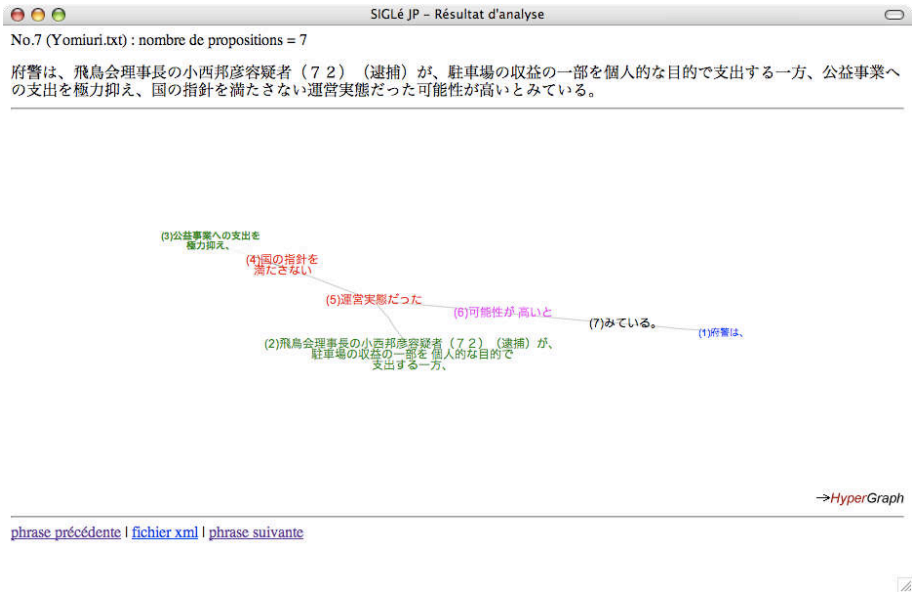


FIG. 11.10 – Affichage du résultat sous forme d'un graphe

11.9 Évaluation

Nous avons réalisé l'évaluation de notre système avec quatre corpus[5] :

1. **traduction d'un article de journal français** : corpus LMDJP2 (124 phrases, LMD ci-après) ;
2. **traduction d'un brevet technique franças** : corpus Brevet1 (158 phrases, Brevet ci-après) ;
3. **articles de journal japonais** : corpus Asahi (112 phrases, Asahi ci-après) ;
4. **texte littéraire** : corpus FdT, extrait du roman « *Fin de Temps* » de MURAKAMI Haruki (149 phrases, Murakami ci-après) ;

Ces corpus, non utilisés lors du développement du système, ont été choisis en tenant compte de la différence des styles, due non seulement au genre du texte (journal, roman, etc.), mais aussi au fait que certains sont des traductions.

11.9.1 Caractéristiques des corpus et méthodologie de l'évaluation

Caractéristiques des corpus

Les lignes A, B et C du tableau 11.13 page 393 présentent les principales caractéristiques de chaque corpus en chiffres[6], la figure 11.11 page ci-contre montrant

[5] Pour le contenu détaillé de chaque corpus, voir la Liste des corpus utilisés (page 547).

[6] L'interprétation du nombre de propositions nécessite cependant une certaine prudence. En effet, ces chiffres contiennent les unités considérées comme externes à toute proposition telles que

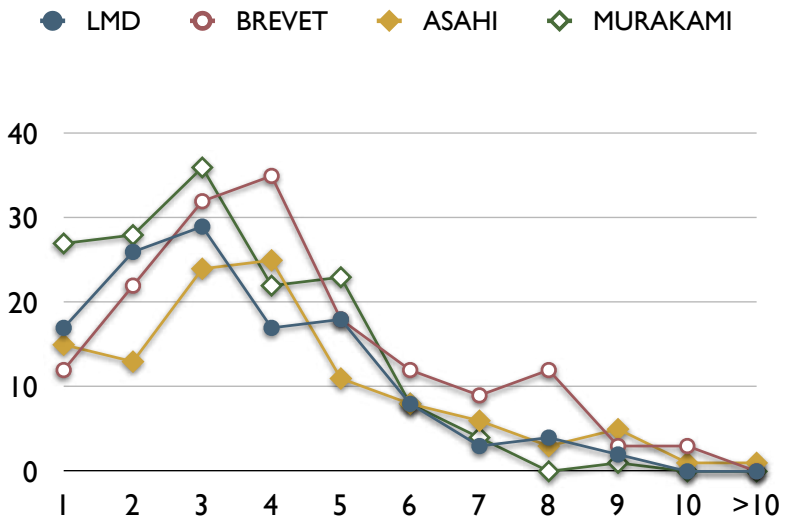


FIG. 11.11 – Distribution des phrases en fonction du nombre de propositions qu'elles contiennent

quant à elle la distribution des phrases en fonction du nombre de propositions qu'elles contiennent.

Le nombre moyen de propositions dans une phrase est compris entre 3 et 4 pour l'ensemble du corpus. Mais la répartition du nombre des phrases en fonction du nombre de propositions qu'elles contiennent varie selon les corpus. Le nombre de phrases contenant plus de cinq propositions diminue nettement pour les corpus « LMD » et « Murakami » tandis que le corpus « Brevet » contient un nombre significatif de phrases avec huit propositions.

Ce constat est encore plus net lorsque nous nous référons à la figure 11.12 (voir page suivante) qui montre pour chaque corpus les proportions de phrases classées selon le nombre de propositions contenues.

Pour le corpus « Murakami », les phrases contenant moins de six propositions représentent plus de 90% et pour « LMD », plus de 85%. En revanche, dans le corpus « Brevet », elles représentent à peine 75%, et dans « Asahi » aussi, moins de 80%. Les deux premiers corpus peuvent donc être qualifiés de « constitués de phrases relativement brèves » et les deux derniers, au contraire, caractérisés par leurs phrases longues.

les thèmes ou les syntagmes adverbiaux cadratifs. Mais notre définition de leur statut est encore plutôt expérimentale et le nombre de propositions peut changer selon cette définition. De plus, il y a encore beaucoup de constituants pour lesquels nous avons du mal à nous forger un avis sur leur statut de proposition. Nous discuterons de ces problèmes plus précisément dans l'analyse des résultats.

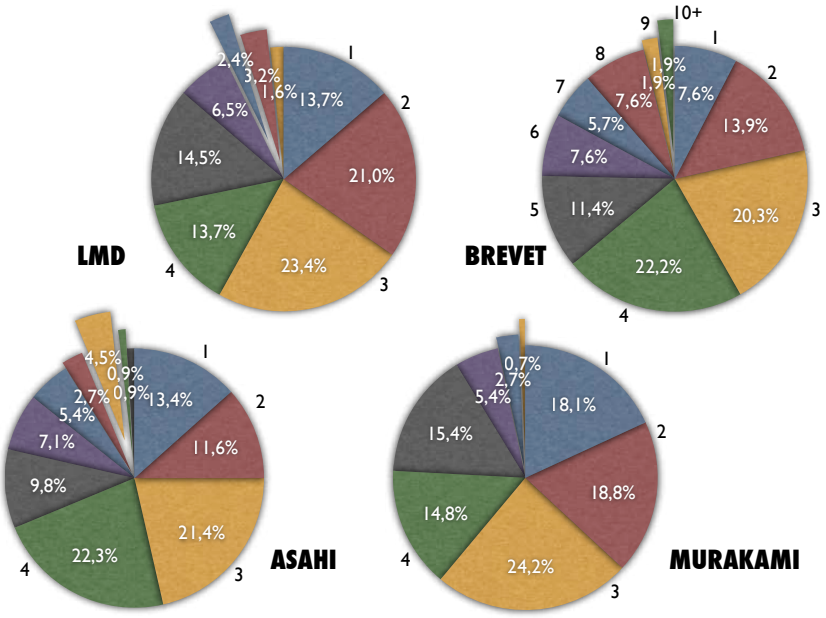


FIG. 11.12 – Proportions de phrases selon le nombre de propositions contenues, par corpus

Méthodologie : évaluation en deux temps

L'ensemble des résultats est présenté dans le tableau 11.13 page ci-contre.

Les résultats fournis par le système sont évalués sur deux axes différents : analyse linéaire et analyse structurale.

L'analyse linéaire concerne les frontières entre les propositions détectées par le système. L'analyse structurale est la détermination des relations de dépendance entre les propositions.

Par exemple, dans le résultat d'analyse suivant (|| indique la frontière détectée) :

私は(A) || 目を閉じて、(B) || 眼鏡のレンズを洗うように(C) || 右の脳と左の脳をからっぽにした。(D)

watashi wa || *me wo tojite* || *megane no renzu wo arau yôni* || *migi no nô to hidari no nô wo karappo ni shita*

moi [*wa*] || fermer les yeux || comme on nettoie des verres de lunettes || vider le cerveau droit et le cerveau gauche [passé]

« Les yeux fermés, j'ai vidé mon cerveau droit et mon cerveau gauche, comme on nettoie des verres de lunettes. »

		LMD	Brevet	Asahi	Murakami
A	Nombre de phrases	124	158	112	149
B	Nombre de propositions	433	670	450	479
C	Nombre moyen de propositions dans une phrase	3,49	4,24	3,99	3,21
D	Nombre de propositions détectées	444	672	453	490
E	Nombre de propositions détectées **correctement**	389	589	391	426
F	Rappel (= **E/B**)	0,898	0,879	0,869	0,891
G	Précision (= **E/D**)	0,876	0,876	0,863	0,869
H	**Analyse linéaire** = nombre de phrases correctement analysées (**H/A %**)	99 (80%)	119 (75%)	85 (76%)	120 (81%)
I	**Analyse structurale** = nombre de phrases correctement analysées (**I/H %**)	94 (95%)	107 (90%)	79 (93%)	111 (93%)

TAB. 11.13 – Caractéristiques des corpus et résultat de l'évaluation

la phrase étant effectivement constituée d'un thème (A) et de trois propositions (B), (C) et (D), toutes les frontières sont correctement détectées et l'analyse linéaire de cette phrase est donc réussie.

En revanche, la proposition (B) est analysée comme dépendant de la proposition (C) et non de (D). Ainsi, l'analyse structurale de cette phrase est considérée erronée.

Les lignes de D à H du tableau présentent l'évaluation des résultats de l'analyse linéaire, et la dernière ligne I montre celle des résultats de l'analyse structurale.

11.9.2 Évaluation de l'analyse linéaire

Il existe deux types de frontières à détecter : frontière initiale et frontière finale.

Les frontières finales sont généralement bien marquées par la présence du prédicat de la proposition. D'ailleurs, le système CBAP se base sur ce fait pour la détection des frontières finales. Pour notre système, les erreurs sur la détection des frontières finales sont dues principalement à une mauvaise analyse morphologique, tout comme pour le système CBAP.

La frontière initiale d'une proposition coïncide généralement avec la frontière finale de la proposition précédente (ou d'une autre unité équivalente). Mais dans le cas de la subordonnée enchâssée, sa frontière initiale apparaît au milieu de la proposition qui l'entoure. Nous avons besoin, dans ce cas, d'identifier spécifiquement les frontières initiales, ce qui est une opération non réalisée par le système antérieur CBAP, mais présent dans le nôtre.

Les erreurs sur la détection des frontières initiales proviennent essentiellement de la mauvaise analyse, par le système extérieur CaboCha, des relations de

dépendance des compléments ainsi que celle des structures de coordination.

Nous allons maintenant étudier plus précisément différents exemples de détection (ou de non-détection) des frontières finales et initiales. L'analyse des résultats de la détection finale est réalisée notamment par comparaison avec les résultats du système antérieur dédié spécifiquement à cette tâche.

Détections des frontières finales

Les auteurs du système CBAP analysent les erreurs constatées dans les résultats et les distinguent en quatre principaux types :

1. les problèmes des règles de détection ;
2. les erreurs liées aux verbes composés ;
3. les frontières fondamentalement difficiles à détecter ;
4. les erreurs dues à l'analyse morphologique.

Nous examinons comment notre système a réagi devant chaque type d'erreur.

1. Problèmes des règles de détection

Notre définition des propositions, englobant l'ensemble des structures à mot variable à la forme autonome suivi d'un mot non autonome ou d'une particule adverbiale sous le nom de proposition à connecteur agglutinant, semble fonctionner de manière plus robuste que le système antérieur, qui énumère les mots susceptibles de constituer les propositions.

En effet, avec notre définition des propositions, l'absence de définition d'un mot agglutinant provoque une erreur de découpage – c'est-à-dire que la frontière est détectée avant le mot agglutinant au lieu de l'identifier après –, mais l'omission complète de détection d'une frontière faute de règle adéquate est théoriquement impossible.

Ce fait confirme la meilleure efficacité de notre définition des propositions, et peut-être, du moins sur le plan pratique, sa justesse.

2. Erreurs liées aux verbes composés

Les verbes composés fonctionnant ensemble comme un seul verbe sont parfois segmentés incorrectement par le système CBAP. En revanche, pour notre système, le problème se pose plutôt en sens inverse.

Suivant notre définition de la proposition, notre système fusionne, en fait, tous les verbes à la forme neutre avec les verbes les suivant directement sans aucun complément, d'où le risque de non-segmentation mais pas de sur-segmentation.

Avec cette méthode, nous avons eu plusieurs résultats de fusion de verbes aspectuels avec le verbe noyau précédent, comme dans la phrase suivante correctement analysée, contrairement au système CBAP qui a détecté incorrectement une frontière entre ces deux verbes :

立ち返って || みる べき なのだ
tachikaette || *miru - beki - nanoda*
revenir || essayer - devoir - en effet
« En effet, (nous) devons essayer de revenir (au ...) »

Par ailleurs, nous considérons également que les verbes à une forme neutre sans aucun complément n'ont pas de fonction de prédicat apte à constituer une proposition. Avec cette définition, les verbes utilisés comme des adverbes ne sont pas non plus détectés faussement comme des propositions.

Dans la phrase suivante, alors que le système CBAP a détecté deux propositions en déterminant, de manière erronée, une frontière après le mot variable « *dô yatte* » (comment) comme indiqué dans l'exemple, notre système ne l'a pas considéré comme un prédicat et n'y a détecté qu'une seule proposition :

日々を どうやって || 切り抜ける か
hibi wo - dô yatte || *kirinukeru - ka*
jours [*wo*] - comment || se débrouiller - [interrogation]
« Comment se débrouille-t-on au quotidien »

Dans les résultats d'analyse, nous n'avons pas constaté de détection omise à cause de ces règles, mais elles restent heuristiques et le risque qu'elles entraînent des erreurs de segmentation existe.

Par ailleurs, l'introduction de la notion de mot variable support a apporté dans les résultats de notre méthode une amélioration de l'analyse des verbes sans autonomie. Par exemple, les phrases suivantes :

出動型の 警察活動が 重視される ようになった
shutsudô gata no - keisatsu katsudô ga - jûshi sareru - yô ni natta
type d'intervention [*no*] - activités policières [*ga*] - être considéré comme important - devenir [passé]
« la priorité a été donnée à une police d'intervention. »

最大の危険の一つが おそらく そこに 巣食っている からだ。
saidai no kiken no hitotsu ga - osoraku - soko ni - sukutteiru - karada
un des dangers les plus importants [*ga*] - sans doute - là-bas [*ni*] - nicher [état] - parce que
« Car l'un des risques majeurs réside sans doute là. »

sont analysées par notre système comme une proposition, alors que le système CBAP en détecte deux comme :

出動型の 警察活動が 重視される ように || なった
shutsudô gata no - keisatsu katsudô ga - jûshi sareru yô ni || *natta*

最大の危険の一つが おそらく そこに 巣食っている から || だ。
saidai no kiken no hitotsu ga - osoraku - soko ni - sukutteiru - kara || *da*

Dans ces phrases, les mots variables isolés à la fin sont en général considérés comme des auxiliaires ou comme fonctionnant comme des auxiliaires et on ne leur accorde pas d'autonomie pour constituer seuls le prédicat.

3. Frontières fondamentalement difficiles à détecter

La méthode de CBAP, basée sur des patrons créés à partir d'une grammaire locale, compte parmi les frontières difficiles à détecter celles produites par les prédicats nominaux et qualificatifs.

Prédicats nominaux À l'aide des résultats de l'analyse des relations de dépendance réalisée par CaboCha, notre système a réussi à détecter les propositions à prédicat nominal sans mot variable, telles que (la séquence soulignée correspond au prédicat nominal sans mot variable) :

今年６月に　省昇格への法案を　閣議決定、||　先の通常国会に　提出した
kotoshi roku gatsu ni - shô shôkaku e no hôan wo - kakugi kettei || sakino tsûjô kokkai ni - teishutsu shita
en juin cette année - projet de loi sur la promotion au statut de ministère [*wo*] - décision au Conseil de cabinet || dernière session ordinaire du Parlement [*ni*] - soumettre [passé]
« En juin, le projet de loi sur la promotion au statut de ministère a été décidé au Conseil de cabinet et il a été soumis à la dernière session ordinaire du Parlement. »

Cette structure, brève, est très utilisée dans le style journalistique où il existe une limitation de l'espace très stricte. Néanmoins, cette méthode totalement dépendante de l'analyse des relations dépendancielles, risque également des détections erronées comme (* indique la nature erronée des frontières détectées) :

約２万７千円分について　収支報告書を　修正、||*　削除した
yaku 2 man 7 sen yen bun ni tsuite - shûshi hôkoku sho wo - shûsei || sakujo shita*
pour ce qui correspond à 27 000 yens environ - rapport sur les recettes et les dépenses [*wo*] - modification ||* supprimer [passé]
« (On a) réalisé une modification et/ou une suppression dans le rapport des recettes et des dépenses pour des notes correspondant à 27 000 yens environ. »

Dans cette phrase, le substantif « *shûsei* » (modification) est coordonné avec « *sakujo* » (suppression), mais le complément en *wo* étant analysé comme dépendant de ce premier, le système a considéré, de manière erronée, la suite « *shûshi hôkoku sho wo - shûsei* » comme une proposition.

Prédicats qualificatifs Notre système détecte également les propositions à prédicat qualificatif sous la condition que celui-ci régisse au moins un complément.

変速機の直流母線上の過電圧が　管理可能な　||　従来の変速機の　機能
hensokuki no chokuryû bosen jô no kaden'atsu ga - kanri kanô na || jûrai no hensokuki no - kinô
les surtensions sur le bus continu du variateur de vitesse [*ga*] - être possible à gérer || variateur de vitesse classique [*no*] - fonctionnement
« fonctionnement selon l'art antérieur d'un variateur de vitesse classique apte à gérer les surtensions sur le bus continu du variateur de vitesse »

Néanmoins, les propositions détectées ainsi sont presque toutes des syntagmes non-propositionnels tels que :

非常に　無防備な　||*　気持ちになった
hijôni - mubôbina || kimochi ni natta*
extrêmement - démilitarisé ||* se sentir [passé]
« (je me suis) senti sans aucune défense. »

（彼らは||）おそろしく　神経質で　用心深く　||*　映った
(karera wa ||) - osoroshiku - shinkeishitsu de - yôjinbukaku || utsutta*
(eux [*wa*] ||) - horriblement - nerveux - prudent ||* me paraître [passé]
« Ils me paraissaient horriblement nerveux et prudents. »

Ces erreurs sont dues à la condition trop simpliste : « le prédicat doit régir au moins un complément ». Afin d'éviter ces détections erronées, il faudrait des conditions plus précises qui limitent les types de compléments régis, par exemple les compléments en particule de cas.

4. Erreurs dues à l'analyse morphologique

Les erreurs de l'analyse morphologique ont une influence directe sur les résultats, aussi bien pour le système antérieur que pour le nôtre.

Les erreurs constatées les plus importantes de ce type proviennent d'une mauvaise analyse de l'unité morpho-lexicale pluricatégorielle *de* (soit particule de cas, soit mot variable ou terminaison du mot variable). L'étiquetage de la particule de cas comme mot variable provoque la détection erronée des frontières finales et dans le cas inverse, l'absence de détection de frontière.

En outre, les substantifs risquent également d'être mal analysés et étiquetés verbes à la forme neutre.

Nous avons constaté que certaines de ces erreurs étaient corrigeables par l'introduction d'un filtrage basé sur les connaissances linguistiques.

Toutefois, nous n'avons pas élaboré cette fonction, car nous avons considéré l'amélioration des résultats de l'étiquetage morphologique comme une tâche à part entière, qui n'est pas censée être intégrée dans une autre opération de niveau différent telle que la détection des propositions.

Détection des frontières initiales

Les erreurs sur la détection des frontières initiales concernent uniquement des structures enchâssées pour lesquelles il existe deux types de structures : subordonnées déterminant un syntagme nominal coordonné avec un autre syntagme nominal qui le précède ; subordonnées enchâssées entre deux compléments du prédicat principal.

Dans les deux cas, le système CBAP ne détecte pas de frontière initiale, entraînant ainsi des segmentations erronées.

Avec l'analyse des relations de dépendance, il est possible, théoriquement, de détecter correctement ces structures imbriquées. Mais, du fait de la difficulté de

détermination de ces relations, cette opération supplémentaire semble plutôt rajouter – au lieu de supprimer – des sources d'erreurs potentielles.

La détermination difficile des éléments coordonnées implique finalement des résultats semblables à ceux du système CBAP, comme (les séquences soulignées et indexées correspondent aux syntagmes nominaux coordonnés ; || entre parenthèses indique une frontière non détectée) :

首相を経ていた || 法案提出や(A)、(||) 海上警備行動発令の 承認を得る || 閣議要求など(B)

shushô wo eteita || *hôan teishutsu ya(A)* - *kaijô keibi kôdô hatsurei no* - *shônin wo eru* || *kakugi yôkyû nado*

(qu')on passait par le Premier Ministre || - soumission du projet de loi [coordination](A) - (qu')on obtient l'autorisation de promulgation des activités de surveillance maritime || demande d'une réunion du Conseil de cabinet(B)

« la soumission d'un projet de loi jusqu'alors via par le Premier Ministre, et la demande d'une réunion du Conseil de cabinet pour obtenir l'autorisation de promulgation des activités de surveillance maritime »

Le syntagme nominal (A), coordonné avec le SN (B), est mal analysé et est inclus dans la subordonnée déterminant le syntagme (B). Nous ne constatons pas, pour ce type de structure, d'amélioration suite à l'utilisation des résultats de l'analyse syntaxique.

En revanche, beaucoup de subordonnées imbriquées entre deux compléments sont correctement détectées grâce à l'analyse des relations dépendancielles.

この二つの要因が(A) || かねて 賃金社会の到来によって 改善されていた(B) || 労働者層の生活不安を 再び助長することになった(C)

kono futatsu no yôin ga(A) || *kanete* - *chingin shakai no tôrai niyotte* - *kaizen sareteita(B)* || *rôdôsha sô no seikatsu fuan wo* - *futatabi jochôsuru kotoni natta(C)*

ces deux facteurs [*ga*] || - précédemment - suite à l'avènement d'une société salariale - être amélioré [état - passé]] || inquiétude pour la vie des milieux populaires - accroître à nouveau [passé]

« Ces deux facteurs ont accru la précarisation des conditions des milieux populaires, que l'avènement d'une société salariale avait contribué à réduire »

Le constituant (A) est bien analysé comme un complément appartenant à (C), les frontières initiale et finale de la subordonnée (B) étant par conséquent bien déterminées.

Mais de par la nature « probabiliste » de l'analyseur, il lui arrive autant de bien analyser, que de se tromper.

それだけの準備を(A) (||) ガン・ファイトにのぞむ前の(B) || 『ワーロック』のヘンリー・フォンダみたいに 手際よくすませた(C)

soredake no junbi wo(A) (||) *gan faito ni nozomu mae no(B)* || *"wârokku" no henrî fonda mitai ni* - *tegiwayoku sumaseta(C)*

cet ensemble de préparations [*wo*] (||) - avant de se rendre au combat au pistolet [*no*] || comme Henry Fonda dans "Warlock" - terminer avec adresse [passé]

« (J'ai) terminé cet ensemble de préparations avec adresse comme Henry Fonda dans "Warlock" avant de se rendre au combat au pistolet »

La constituant (A) est mal analysé car comme dépendant du prédicat de (B) et non de (C), inclus ainsi dans la subordonnée à connecteur agglutinant *mae* (avant).

La fausse analyse des relations provoque non seulement des sous-détections mais aussi des sur-détections de frontières.

ナイアガラの滝の上から(A) || 落とされたり、(B) || あるいは 北海で氷づけになったりするのだ(C)

naiagara no taki no ue kara(A) || *otosaretari(B)* || *aruiwa - hokkai de koori zukeni nattari suru noda(C)*

du haut des chutes du Niagara || se faire tomber [énumération] || ou - être trempé dans les glaces de la mer du Nord [énumération]

« (Il) se fait entre autres tomber du haut des chutes du Niagara ou tremper dans les glaces de la mer du Nord »

Le constituant (A) dépendant de (B) qui le suit directement est analysé incorrectement comme dépendant de (C). Ainsi, le constituant (B) est considéré comme subordonnée enchâssée et une frontière est faussement détectée après le constituant (A).

À cause notamment de ces cas de sur-détections, la détection des propositions imbriquées n'a pas été améliorée en dépit de l'introduction de l'analyse des relations de dépendance.

En effet, pour le corpus Asahi par exemple, les phrases contenant des structures imbriquées nécessitant une analyse syntaxique pour leur segmentation correcte – c'est-à-dire, des phrases que le système CBAP ne sait pas analyser correctement – représentent 13 phrases sur 112, soit 12%. Néanmoins, dans ces 13 phrases, celles correctement analysées grâce à l'analyse syntaxique CaboCha se limitaient à 5. En revanche, le nombre de phrases mal analysées du fait d'erreurs dans l'analyse des relations de dépendance s'élève à 19. Ce constat nous oblige finalement à remettre en cause l'utilité de cette opération supplémentaire.

11.9.3 Évaluation de l'analyse structurale

La ligne I du tableau 11.13 page 393 indique le nombre de phrases, parmi celles correctement segmentées, dont l'analyse des relations entre les propositions détectées est également correcte.

L'analyse des relations de dépendance entre les compléments et les prédicats a de l'influence sur les résultats de l'analyse linéaire, et l'analyse structurale reflète les résultats de l'analyse des relations dépendancielles entre les prédicats.

La figure 11.14 (voir page suivante) est un exemple de résultat correct que nous avons obtenu pour la phrase suivante, très longue, constituée de six propositions :

「在日朝鮮人科学技術者の親睦（しんぼく）団体」とされる一方で、０３年に摘発された 都内メーカーのミサイル関連機器不正輸出事件では 科

```
<s id='62'>
<prop id='1' pere='6' fils='' kakari='吸節(一方)+で'>
    <txt1>「在日朝鮮人科学技術者の親睦（しんぼく）団体」とされる一方で、</txt1>
</prop>
<prop id='2' pere='5' fils='' kakari='連体'>
    <txt1>０３年に摘発された</txt1>
</prop>
<prop id='3' pere='4' fils='' kakari='中立'>
    <txt1>科協系企業が関与して</txt1>
</prop>
<prop id='4' pere='5' fils='3;' kakari='吸節(こと)+が'>
    <txt1>[中立]北朝鮮にも送っていたことが</txt1>
</prop>
<prop id='5' pere='6' fils='2;4;' kakari='吸節(など)'>
    <txt1>[連体]都内メーカーのミサイル関連機器不正輸出事件では[吸節(こと)+が]判明するなど、</txt1>
</prop>
<prop id='6' pere='0' fils='1;5;' kakari='F'>
    <txt1>[吸節(一方)+で][吸節(など)]技術・物資流出への関与が指摘されてきた。</txt1>
</prop>
</s>
```

FIG. 11.14 – Résultat de l'analyse d'une phrase, sous forme xml

協系企業が関与して北朝鮮にも送っていたことが　判明するなど、技術・物資流出への関与が指摘されてきた。

"zainichi chôsenjin kagaku gijutsusha no shinboku (shinboku) dantai" to sareru ippô de - 03 nen ni tekihatsusareta - tonai mêkâ no misairu kanren kiki fusei yushutsu jiken dewa - kakyôkei kigyô ga kan'yo shite - kitachôsen nimo okutteita koto ga - hanmeisuru nado - gijutsu· busshi ryûshutsu eno kan'yo ga shitekisarete kita

Alors qu'elle est considérée comme un organisme amical des scientifiques nord-coréens demeurant au Japon - (qui) a été dénoncé en 2003 - dans l'affaire d'exportation clandestine d'équipements pour missiles par un fabriquant de Tokyo - le(s) filiale(s) de l'Association Kakyô y participe(nt) - le fait qu'(on les) envoyait aussi en Corée du Nord [*ga*] - comme par exemple, a été découvert - faire remarquer son implication dans les fuites techniques et matérielles [passé]

« Alors que [cette association] est considérée comme un organisme amical des scientifiques nord-coréens demeurant au Japon, on faisait remarquer son implication dans des fuites techniques et matérielles, comme par exemple dans l'affaire d'exportation clandestine d'équipements pour missiles par un fabriquant de Tokyo, dénoncée en 2003, où il a été découvert qu'on les envoyait aussi en Corée du Nord, avec le concours de(s) filiale(s) de l'Association Kakyô »

La figure 11.15 page suivante est une représentation en graphe du résultat de l'exemple. On voit que toutes les relations sont correctement déterminées.

Comme le résultat le montre, les relations entre les prédicats sont relativement bien déterminées, mais il est tout de même rare que les phrases longues, comme celle de l'exemple, soient entièrement correctement analysées.

11.9.4 Évaluation des autres tâches réalisées par le système

Nous avons réalisé en plus de la reconnaissance des propositions deux tâches d'extraction des deux autres unités jugées situées au même niveau que les propo-

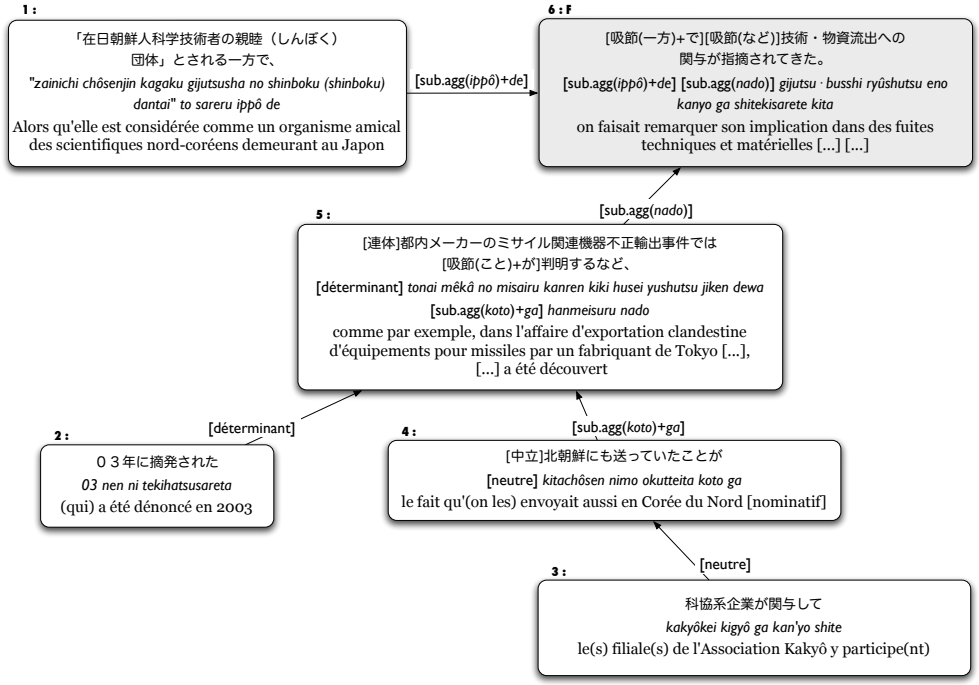


FIG. 11.15 – Résultat de l'analyse d'une phrase, sous forme de graphe

sitions : les thèmes et les éléments externes.

Les résultats d'extraction n'ont pas vraiment été évalués car le statut même de ces unités n'est pas encore bien défini et nous avons considéré ne pas encore être aptes à déterminer la justesse de l'analyse. Cette analyse n'en est donc encore qu'au stade expérimental, mais nous continuerons à examiner les effets de nos définitions sur ces sujets avec les résultats de l'alignement avec les propositions françaises.

Par ailleurs, comme déjà mentionné, nous avons également introduit un pré-traitement des séquences entourées de deux symboles, parenthèses et guillemets. Nous aborderons donc aussi les remarques sur les résultats de ce pré-traitement.

Analyse des syntagmes en *wa*

Nous avons initialement distingué les SN en *wa* sans particule de cas, thèmes forts, des SN en *wa* avec particule de cas, thèmes faibles. Ensuite, nous avons ré-analysé les thèmes forts et considéré ceux qui apparaissaient à l'intérieur de la portée d'un autre thème fort comme des thèmes faibles. Puis les thèmes forts sont considérés comme extérieurs à la proposition, mais les thèmes faibles sont en revanche inclus dans la proposition.

Les résultats de cette analyse sont encourageants : l'ensemble des schémas de

phrase que nous avons eu avec cette méthode nous semblent confirmer la justesse de notre distinction des thèmes en deux types.

Par exemple, dans la structure racine de la phrase suivante :

前代表は、堀江前社長から　ニッポン放送の経営権を　取りたい　趣旨を
聞いた　ことは　認めた　ものの、
「時価総額１８００億円のニッポン放送の経営権を　本気で　取るつもり
だとは　到底　思えなかった」と　し、
「ライブドア一流の思いつきの面白おかしい大言壮語を　聞いたという
受け止め方だった」と　述べた。

zendaihyô wa - horie zen shachô kara - nippon hôsô no keieiken wo - toritai shushi wo - kiita - koto wa - mitometa - monono
"jika sôgaku 1800 oku en no nippon hôsô no keieiken wo - honkide - toru tsumori da to wa - tôtei - omoenakatta" - to - shi
raibudoâ ryû no omoitsuki no omoshiro okashii taigen sôgo wo - kiita toiu uketomekata datta" - to - nobeta

ancien représentant [*wa*] - ancien président Horie [*kara*] - pouvoir gestionnaire [*wo*] -vouloir prendre - résumé [*wo*] - écouter - le fait de [*wa*] - reconnaître [passé] - [concession] - "pouvoir gestionnaire de la Chaîne Nippon qui coûte actuellement 180 milliards de yens [*wo*] - sérieusement - compter prendre [citation] [*wa*] - absolument - ne pas pouvoir croire [passé]" - [citation] - faire

"fanfaronnade folle et drôle, idée spontanée à la Livedoor [*wo*] - entendre [passé] [déterminant] - être une interprétation [passé]" - [citation] - dire [passé]

« L'ancien représentant a affirmé, tout en reconnaissant qu'il avait entendu l'ancien président Horie exprimer son désir de prendre la direction de la Chaîne Nippon, qu'il n'avait absolument pas cru qu'il avait vraiment l'intention de prendre la direction de la Chaîne Nippon qui coûtait à l'époque 180 milliards de yens, et qu'il s'était dit avoir entendu une fanfaronnade folle et drôle, dans la veine des idées spontanées à la Livedoor »

existent deux syntagmes en *wa*, « *zendaihyô wa* (ancien représentant [*wa*]) » et « *... wo kiita koto wa* (le fait d'avoir entendu ... [*wa*]) ». Alors que le premier « *zendaihyô wa* (ancien représentant [*wa*]) » porte sur l'ensemble de la phrase et entre en relation avec les trois prédicats « *mitometa (monono)* (reconnaître) », « *(to) shi* (faire) », « *(to) nobeta* (dire) », le second « *... wo kiita koto wa* (le fait d'avoir entendu ... [*wa*]) » n'entre en relation qu'avec le prédicat adjacent « *mitometa (monono)* (reconnaître) ».

Dans le résultat de l'analyse, ce second syntagme en *wa* est considéré comme thème faible à portée limitée, inclus dans la proposition constituée du prédicat « *mitometa (monono)* (reconnaître) », et seul le premier est analysé comme thème fort, thème de la phrase entrant en relation avec l'ensemble du reste de la phrase.

Toutefois, comme nous l'avons déjà dit, la définition du statut de syntagme en *wa* est un sujet contenant encore beaucoup de points à étudier et il existe aussi dans le résultat d'analyse des exemples mettant en cause notre définition du thème syntaxique. Les syntagmes en *wa* contenant également une particule de cas, par exemple, considérés comme thèmes faibles quel que soit le contexte d'apparition, nous paraissent parfois fonctionner comme un thème.

Par ailleurs, cette méthode possède comme défaut la dépendance totale au résultat de l'analyse syntaxique. Les SN en *wa* possèdent des propriétés non encore exploitées qui pourraient servir au contraire à l'amélioration de l'analyse syntaxique.

Extraction des éléments externes

Nous avons considéré comme des éléments externes tous les syntagmes n'appartenant pas au thème et préposés par rapport à ce dernier.

Tous les éléments extraits par cette règle semblaient correspondre à cette qualification de « externes », mais des problèmes se posent, surtout dans le sens de la sous-détection. En effet, dans les phrases sans SN thématisé, apparaissent également des éléments externes.

Comme nous l'avons répété lors des études linguistiques, ce sujet n'est pas encore suffisamment étudié, et il reste encore beaucoup de questions à régler.

Traitement des éléments entre guillemets et parenthèses

Nous avons réalisé un pré-traitement qui extrait tous les éléments entre parenthèses ou guillemets de manière à remplacer ces ensembles par une sorte de boîte noire pour que les analyseurs extérieurs réalisent des analyses en considérant comme un symbole les éléments entre parenthèses et comme un SN les éléments entre guillemets.

Les résultats avec ce module de pré-traitement étaient plus propres et bien entendu plus corrects.

Cependant, nous avons également rencontré quelques problèmes. En effet, les guillemets n'entourent pas forcément un ensemble cohérent et il arrive de n'enchâsser qu'une partie de la proposition. Dans la phrase suivante :

前代表が　同放送社株を　「買えるだけ買え」と　部下に　指示した
zen daihyô ga - dôhôsôshakabu wo - "kaerudake kae" to - buka ni - shijishita

ancien président [*ga*] - actions de cette chaîne de radio [*wo*] - "achetez autant que possible" [citation] - subordonné [*ni*] - ordonner [passé]

« L'ancien président a ordonné à son/ses subordonné(s) d'"acheter le plus possible" des actions de cette chaîne de radio »

Le complément accusatif en *wo* extérieur aux guillemets dépend en réalité du prédicat situé entre guillemets. Les éléments entre guillemets étant traités comme des SN, la détection des propositions a bien entendu échoué. Cette méthode de la boîte noire est en fait efficace dans le cas où les guillemets ou les parenthèses entourent plusieurs phrases. Il serait donc nécessaire d'ajouter une condition supplémentaire qui permettrait de ne pas appliquer cette méthode de la boîte noire lorsque les éléments entourés ne contiennent pas plus d'un prédicat.

Par ailleurs, dans le corpus littéraire, les symboles ne se limitent pas aux parenthèses et aux guillemets : il contient plusieurs passages dont une partie est isolée

par le symbole « – » (tiret). Mais, le traitement de ce symbole est aussi délicat, car, contrairement aux parenthèses et guillemets, il sert aussi bien à l'ouverture qu'à la fermeture de la partie isolée. Cette ambiguïté est d'autant plus dangereuse que ce symbole peut être utilisé seul comme dans la phrase :

とにかく正確な数字を確認すること　－　救済は　それによって　もたらされる　はずだった

tonikaku seikakuna sûji wo kakuninsuru koto – kyûsai wa - sore niyotte - motarasareru - hazudatta

vérifier les chiffres exacts avant tout – sauvetage [*wa*] - par cela - me/nous être rapporté - devoir[passé]

« Vérifier avant tout les chiffres exacts – ce qui devait m'apporter le sauvetage »

De plus, les énumérations perturbent également l'analyse syntaxique. Dans notre corpus Asahi, deux phrases contenant des énumérations apparaissent, entraînant une analyse des relations de dépendance pour ces phrases (fournie par le système CaboCha) complètement fausse. Si l'analyseur syntaxique utilisé ne prend pas en compte ce style, il est impératif d'effectuer un traitement préalable afin d'obtenir les résultats souhaités.

11.9.5 Remarques sur les différences des résultats entre les corpus

Nous avons utilisé quatre corpus de nature différente pour déterminer l'influence de cette différence sur la performance du système.

Les chiffres des résultats semblent refléter, non pas leur différence de nature mais plutôt la longueur des phrases. Leur différence de nature se retrouve plus sur les types des erreurs détectés.

Pour le corpus littéraire, nous avons constaté plusieurs cas d'absence de définition des mots agglutinants, analysés comme des substantifs classiques. Les symboles typographiques utilisés sont également variés.

Le corpus journalistique est caractérisé par des phrases brèves dont la syntaxe est particulière et qui nécessitent un traitement particulier pour la détermination des prédicats. L'énumération caractérise également ce type de texte et pose également des problèmes lors de l'analyse syntaxique.

11.10 Conclusion et perspectives

Nous avons présenté un système de détection des propositions utilisant un analyseur des relations de dépendance entre les constituants de la phrase.

Le résultat de notre propre évaluation du système existant CBAP, nous a amené à la conclusion que l'adaptation de ce système à notre opération serait difficile. Nous avons donc conçu une méthode de détection des propositions à l'aide d'un analyseur syntaxique : le système d'analyse des relations de dépendance CaboCha. Cette utilisation d'un analyseur des relations de dépendance permet de reconnaître les propositions même imbriquées que le système antérieur CBAP ne peut pas identifier.

Nous avons réalisé deux types d'évaluation : avec le résultat de l'analyse linéaire et avec celui de l'analyse structurale.

L'analyse linéaire est améliorée par rapport au système antérieur. Cette amélioration est due, d'une part à notre définition des propositions, notamment celles à connecteur agglutinant, et de l'autre à l'utilisation d'informations syntaxiques fournies par un moyen extérieur.

En revanche, nous n'avons pas pu obtenir l'amélioration souhaitée pour l'analyse structurale. Cet échec provient principalement du bruit produit par l'analyseur syntaxique extérieur lors de la résolution des relations dépendancielles.

Ce résultat montre que l'utilisation d'un système supplémentaire augmente les informations exploitables mais il multiplie également – et malheureusement – le bruit qui risque d'annuler l'intérêt même du recours à cette analyse.

Nous avons, face à cette réalité, deux possibilités : améliorer le résultat des traitements en amont (analyses morphologique et syntaxique) pour l'utiliser ensuite dans notre opération de détection des propositions ; ou alors, suivre plutôt la voie du système antérieur qui ne recourt pas à l'analyse syntaxique.

Comme nous l'avons dit, nous avons réalisé de manière expérimentale quelques fonctions de filtrage des erreurs fréquentes de l'analyse morphologique. Cette expérience nous a montré qu'une certaine amélioration peut être envisageable par l'introduction de connaissances linguistiques. De même, dans les résultats de l'analyseur probabiliste des relations dépendancielles, nous avons constaté plusieurs erreurs, linguistiquement impossibles, qui pourraient être corrigées par un petit module de post-traitement basé sur les règles linguistiques.

Mais il est également possible d'abandonner totalement le recours à un analyseur syntaxique et d'améliorer plutôt le système antérieur. Dans ce cas, il faudrait plusieurs petits modules spécifiques à une tâche donnée, notamment un module de détermination des compléments d'un prédicat ou un de reconnaissance des éléments coordonnés. Ces modules devraient exploiter un maximum de connaissances linguistiques afin de permettre l'analyse fiable d'une tâche très limitée en utilisant jusqu'à certaines informations sémantiques.

Ce schéma de la chaîne de traitement des petits modules est en fait déjà proposé dans Danlos (2005). Danlos y défend, face aux analyseurs syntaxiques complets suffisamment performants qui n'existent pas aujourd'hui, la pertinence des petits outils modestes destinés à une fonctionnalité bien particulière.

CHAPITRE 12

ALIGNEMENT DES PROPOSITIONS : ÉTAT DE L'ART

Nous abordons dans ce chapitre les travaux existants sur l'alignement des propositions ou sur un sujet connexe. Nous allons tout d'abord passer en revue l'état actuel (§ 12.1) avant de présenter deux méthodes adaptant une technique d'alignement des phrases (§ 12.2). Nous examinerons également les travaux sur l'alignement manuel des propositions (§ 12.3), avant d'aborder les méthodes d'alignement des syntagmes à l'aide d'arbres syntaxiques (§ 12.4).

12.1 Bref aperçu panoramique

Comme déjà précisé dans la section 1.5.3, nous n'avons trouvé que très peu de travaux sur l'alignement des propositions.

Nous pouvons tout de même citer ceux de Piperidis, Papageorgiou et Boutsis (Boutsis & Piperidis, 1998 ; Piperidis et al., 2000), sur les textes parallèles anglais-grec et ceux de Wang & Ren (2005) sur la paire japonais-chinois. Ces deux travaux recourent tous les deux à une technique existante d'alignement des phrases.

Il n'existe à ce jour aucune étude sur l'alignement des propositions traitant le japonais, avec le français, ou même avec l'anglais. Il existe cependant un article portant sur l'alignement manuel des propositions anglais-japonais. Cet article (Kashioka et al., 2003) présente une méthode d'alignement manuel, mais expose aussi des remarques intéressantes pour la réalisation d'un système automatique. En effet, les auteurs constatent beaucoup de croisements des alignements. Nous avons donc besoin, pour automatiser la tâche d'alignement des propositions, de concevoir un algorithme qui ne présuppose pas le parallélisme et qui utilise par exemple une structure non linéaire mais à deux dimensions, telle que les graphes.

Cette idée d'alignement à l'aide de graphes n'est pas nouvelle. Comme nous l'avons déjà mentionné dans la section 1.5.2, plusieurs études de structures inférieures à la proposition utilisant les arbres syntaxiques ont été réalisées. Nous étudierons donc également ces méthodes d'une manière plus approfondie.

12.2 Méthodes adaptant une technique d'alignement des phrases

12.2.1 Méthode proposée par Piperidis *et al.*

L'ensemble de la procédure se déroule comme suit :

1. *tagging* ;
2. reconnaissance des propositions ;
3. alignement des phrases ;
4. alignement des propositions.

La méthode de reconnaissance des propositions a déjà été présentée dans la section 8.2.3.

Comme Brown et al. (1991) l'ont proposé pour l'alignement des phrases (cf. la section 2.2), les auteurs considèrent chaque paire de phrases alignées comme une séquence de perles constituées de plusieurs propositions supposées traductions les unes des autres. L'alignement est alors considéré comme la maximisation de la distribution jointe des probabilités de chaque perle.

La probabilité de la perle des propositions *sc* et *tc* est calculée à partir des trois probabilités suivantes :

- probabilité du type de traduction : Pr_{1-0}, Pr_{0-1}, Pr_{1-1}, Pr_{1-2}, Pr_{2-1} et Pr_{2-2} avec la même valeur que celle utilisée pour l'alignement des phrases ;
- probabilité basée sur les longueurs des propositions : $\mathsf{Pr}(l(sc), l(tc))$ calculée avec le même modèle statistique de longueur en caractères utilisé pour l'alignement des phrases ;
- probabilité basée sur les mots alignés constituant les propositions : seuls les mots lexicaux sont pris en compte. Un mot lexical de *tc* peut correspondre à seulement zéro ou un mot lexical de *sc*.
  Soient scw_i et tcw_j, mots lexicaux contenus dans les propositions *sc* et *tc*, la probabilité $\mathsf{Pr}(\{scw_1, ..., scw_v\}, \{tcw_1, ..., tcw_w\})$ est calculée à partir des probabilités de co-occurrence des paires de mots $\mathsf{Pr}(scw_i, tcw_j)$.

Ainsi, la probabilité d'une perle de propositions peut être décrite par la formule suivante :

$$\mathsf{Pr}(\textit{perle}) = \mathsf{Pr}_{n-m} \cdot \mathsf{Pr}\big(l(sc), l(tc)\big) \cdot \mathsf{Pr}\big(\{scw_1, ..., scw_v\}, \{tcw_1, ..., tcw_w\}\big)$$

Une des remarques importantes que l'on peut faire concernant cette méthode est l'adoption de la même probabilité du type de traduction que celle calculée pour l'alignement des phrases. Intuitivement, la proposition correspondant, dans

la plupart des cas, à une proposition au sens logique, elle constitue une unité plus commune à différentes langues que la phrase – qui peut, elle, être constituée de la conjonction de plusieurs propositions dont le choix diffère sans doute selon les langues. Si bien que l'utilisation des mêmes probabilités du type de traduction semble difficile à justifier.

12.2.2 Méthode proposée par Wang et Ren

Wang et Ren améliorent la méthode d'appariement des phrases basée sur les longueurs des textes par l'introduction d'un calcul de similarité basé sur l'information portée par les idéogrammes Han. Les auteurs combinent l'information statistique et celle sur les idéogrammes Han pour trouver avec une méthode de programmation dynamique l'alignement présentant le coût le moins élevé. Néanmoins, leur méthode ne traite pas en réalité de propositions selon notre définition. La proposition est définie dans leurs travaux comme l'unité entourée de certains types de séparateurs graphiques tels que des virgules. Ce qui nous a amené à la considérer comme une technique d'alignement de phrases plutôt que de propositions.

12.3 Alignement manuel des propositions anglais-japonais

Kashioka et al. (2003) présentent la constitution d'un corpus parallèle avec alignement au niveau des propositions, réalisée dans un but d'utilisation pour la traduction automatique des monologues (e.g. nouvelles télévisées, conférences, présentations techniques).

En effet, suite à la constatation que les monologues ont tendance à être plus longs que les conversations, les auteurs sont convaincus de la nécessité d'une unité de traduction autre que la phrase. La proposition japonaise contenant un syntagme verbal, ils la considèrent comme une unité syntaxiquement suffisante et sémantiquement significative, donc meilleur candidat pour l'unité de traduction automatique.

La constitution du corpus s'est déroulée comme suit. Le corpus de départ a d'abord été formé du recueil de 250 séances (soit 15 313 phrases) de transcription du programme télévisé « *asu wo yomu* » qui est constitué de 10 minutes de présentation d'un événement actuel par un commentateur. Puis les opérations suivantes ont été exécutées[1] :

1. Analyse morphologique des transcriptions à l'aide de l'analyseur morphologique du japonais ChaSen.
2. Détection des frontières de propositions par CBAP.

[1] Seules les opérations d'analyse morphologique et de détection des propositions du texte japonais étaient automatisées, le reste étant réalisé manuellement.

3. Traduction humaine des phrases avec prise en compte des frontières de propositions.

4. Division des phrases traduites en anglais en segments correspondant aux propositions japonaises – réalisée par une personne qui n'est pas un traducteur.

5. Annotation du numéro de ligne du segment anglais correspondant aux propositions japonaises.

Les 15 313 phrases sources en japonais ont été traduites en 15 275 phrases anglaises. Dans les phrases japonaises, ont été détectées par le système 70 989 frontières de propositions, et les phrases anglaises divisées en 73 755 segments.

Un point intéressant est que pour 6 280 propositions japonaises, soit 8,8% de la totalité, le segment anglais correspondant n'a pas été trouvé. Dans près de la moitié des cas (à savoir 2 973 propositions), ces propositions se trouvent en fin de phrase, ce qui signifie que dans 20% de la totalité de phrases, la dernière proposition n'a pas de segment correspondant en anglais. Ce sont probablement des mots dits exprimer la modalité, ces mots étant concentrés aux extrémités dans les phrases japonaises. Il est important, lorsqu'on envisage l'alignement des propositions, d'étudier les différentes possibilités de traduction (ou éventuellement suppression) de ces mots afin de pas être perturbé par ce problème qui peut avoir une influence sur la définition même de la proposition japonaise à adopter.

Une autre remarque faite par les auteurs porte sur la différence d'ordre des propositions japonaises et des segments anglais correspondants : on constate beaucoup de croisements des alignements. Ce qui confirme le non-parallélisme de l'alignement des propositions par contraste avec l'alignement des phrases quasiment parallèle. Nous avons donc besoin, pour automatiser la tâche d'alignement des propositions, de concevoir un autre algorithme qui ne présuppose pas le parallélisme et qui utilise une structure non linéaire mais à deux dimensions.

12.4 Alignement des unités sous-phrastiques à l'aide de graphes

On peut envisager deux types d'alignements avec des arbres syntaxiques : alignement total pour lequel l'ensemble des paires des unités alignées couvre tous les éléments des deux arbres de départ ; alignement hiérarchique qui ne cherche pas à mettre en correspondance des ensembles d'unités d'un niveau donné, mais qui tente d'établir des liens entre les paires de certains des syntagmes, et ce à tout niveau.

Nous nous intéressons maintenant aux méthodes d'alignement hiérarchique avant d'aborder celles d'alignement total.

12.4.1 Approches pour l'alignement hiérarchique

Pour les méthodes visant un alignement hiérarchique, il existe notamment les travaux de Kaji et al. (1992), et ceux de Imamura (2000) basés sur la technique de ces derniers. La méthode de Kaji consiste en :

(a) analyse de la phrase japonaise ;

(b) analyse de la phrase anglaise ;

(c) mise en correspondance des mots entre les phrases japonaise et anglaise ;

(d) appariement des syntagmes correspondants.

Une fois les mots mis en correspondance à l'aide d'un dictionnaire bilingue, l'alignement des syntagmes est réalisé en cherchant pour tout syntagme X de la phrase japonaise un syntagme anglais Y contenant tous les mots appariés avec ceux du syntagme X et n'incluant aucun mot apparié avec un mot n'appartenant pas au syntagme X. Imamura améliore cette méthode d'alignement des syntagmes de tout niveau notamment par la mise à profit des informations sur la nature syntaxique des syntagmes. L'introduction de ces informations supplémentaires permet d'empêcher l'alignement des unités trop petites ayant des étiquettes syntaxiques différentes.

12.4.2 Méthodes visant l'alignement total

Les travaux de Matsumoto et *al.* (Ishimoto et al., 1993 ; Matsumoto et al., 1993) proposent une méthode permettant de trouver des correspondances structurelles entre des phrases parallèles à l'aide du résultat d'une analyse syntaxique en arbre de dépendance, capable de représenter des ambiguïtés syntaxiques. La mise en correspondance détaillée des structures est automatiquement obtenue par comparaison des sous-arbres. La similarité entre une paire de sous-arbres est calculée sur la base des mots correspondants contenus dans les structures considérées. La mise en correspondance des mots est réalisée à l'aide d'un thesaurus. Dans cette méthode, le problème d'appariement est posé comme la recherche d'un appariement *one-to-one* entre les décompositions de deux arbres initiaux, qui maximise la somme des valeurs de similarité entre les sous-arbres appariés.

La méthode proposée par Watanabe et al. (2000) est proche des travaux de Kaji et al. (1992) dans la mesure où dans ces méthodes, les mots correspondants servent à ancrer les textes pour former les segments à extraire. La différence est que la méthode de Watanabe utilise les arbres de dépendance alors que celle de Kaji recourt à l'analyse en constituants. La recherche des structures correspondantes de cette méthode est constituée de trois étapes : construction d'un arbre de dépendance des textes initiaux dans les deux langues ; appariement des mots à l'aide d'un dictionnaire bilingue ; recherche d'un ensemble de structures correspondantes. L'appariement des syntagmes consiste lui-même en quatre opérations :

– Dans la première étape, on cherche toutes les paires des nœuds ancres $W(s_1, t_1)$ et $W(s_2, t_2)$ tels qu'il n'existe aucun nœud ancre entre s_1 et s_2, et

on crée les paires candidates $P(LT(s_1, s_2), T(t_1, t_2))$ où $LT(s_1, s_2)$ correspond au chemin formé entre les nœuds s_1 et s_2, $T(t_1, t_2)$ représentant l'arbre minimal incluant tous les nœuds entre t_1 et t_2.

– Dans la deuxième étape, on vérifie que tous les mots formant les nœuds ancres avec les mots appartenant à P sont également tous inclus dans P. Dans le cas contraire, P est fusionné avec une autre structure candidate de manière à constituer la paire de structures incluant tous les mots formant les nœuds ancres appartenant à celle-ci.
– Dans la troisième étape, toute paire P_x qui partage certains nœuds (sauf les nœuds ancres) avec P_y est fusionnée avec ce dernier de manière à former une paire de structures plus grandes.
– Dans la quatrième étape, on cherche dans les deux arbres tous les chemins $LT(n_1, n_2)$ tels que n_1 appartient à P, tous les autres nœuds n'étant inclus dans aucune structure alignée. Pour chacun de ces chemins, si n_1 n'est pas un nœud ancre, tous les autres nœuds sont aussi inclus dans P. Si n_1 est un nœud ancre, une nouvelle correspondance syntagmatique est créée à partir du chemin considéré.

CHAPITRE 13

NOTRE SYSTÈME D'ALIGNEMENT DE PROPOSITIONS : Mizolé

みぞれ　霙【midzore】n.

1. grésil, neige fondue. **2.** dessert en glaçon râpé au sirop. **3.** radis blanc râpé. **4.** inform. ***MIZOLé*** système réalisant l'alignement des propositions sur la base de l'approche spectrale de l'alignement des graphes ou de la méthode inspirée de la classification ascendante hiérarchique.

Nous présentons, dans le présent chapitre, deux méthodes d'alignement des propositions : l'une basée sur les méthodes d'appariement des graphes et une autre inspirée de la classification ascendante hiérarchique (CAH). Nous allons d'abord décrire, afin de clarifier les conditions du développement, nos données d'entrée (§ 13.1) avant de présenter la problématique et notre choix de solution (§ 13.2). Puis, l'exposé se poursuivra par la description des deux méthodes : par appariement des graphes (§ 13.3) et par CAH (§ 13.4). L'exposé se terminera par l'analyse des résultats obtenus (§ 13.5) et une discussion sur les pistes d'amélioration (§ 13.6).

13.1 Étapes précédant l'alignement des propositions

La figure 13.1 représente l'ensemble des étapes précédant l'alignement des propositions.

Les corpus initiaux d'entrée sont des textes parallèles français-japonais. Ils sont d'abord segmentés en phrases (respectivement A et B dans la figure) et alignés au niveau phrastique par notre système d'alignement des phrases (indexé 1 dans la figure, cf. § 3). Nous réalisons ensuite pour chaque phrase du corpus la détection des propositions ainsi que leur mise en relation à l'aide de nos détecteurs de propositions du français (indexé 2, cf. § 13.1.1 et § 9) et du japonais (indexé 3, cf. § 13.1.1 et § 11), permettant de créer directement les arbres de propositions représentant leurs relations de dépendance (arbres des propositions, ci-après). Nous procédons ensuite à la création des perles (indexée 4, cf. § 13.1.2), consistant en la constitution, à l'aide du résultat d'alignement des phrases (C), des paires des ensembles de phrases alignées – appelées perles – avec les phrases maintenant segmentées en propositions (D et E). Nous réalisons enfin l'alignement au niveau de proposition des perles de phrases alignées ainsi créées (F), une par une.

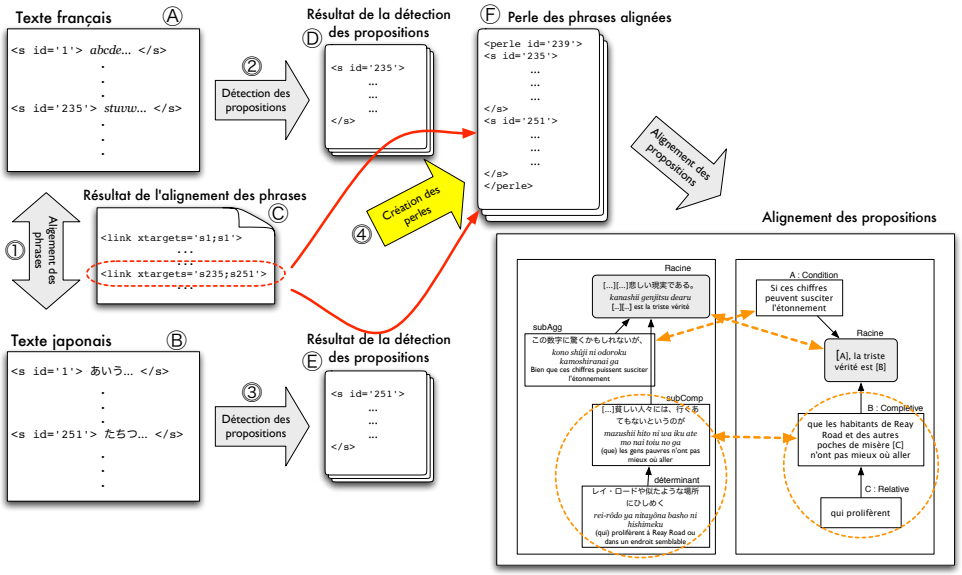


FIG. 13.1 – Étapes précédant l'alignement des propositions

13.1.1 Rappel : brève description de la détection des propositions

Notre détecteur de propositions du français, basé sur notre définition des propositions (cf. chapitre 4), identifie quatre types de propositions (racine, coordonnée, incidente et subordonnée) et les subordonnées sont étiquetées avec

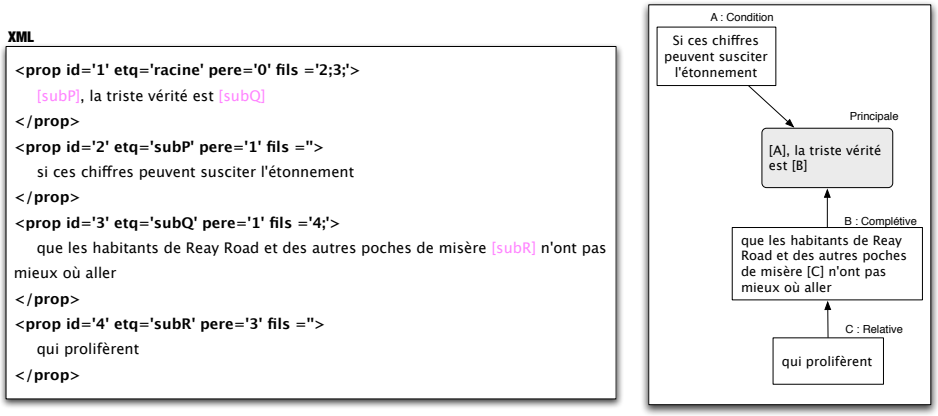


FIG. 13.2 – Résultat de la détection des propositions et arbre construit (FR)

quatre sous-catégories (pré-verbale, post-verbale, périphérique et déterminante). En plus de ces propositions, notre système extrait les éléments appelés extra-prédicatifs (e.g. introducteurs de cadre, constructions détachées ou thème) considérés comme extérieurs à la proposition.

Le résultat d'analyse fourni par le système est donc la liste des propositions détectées avec leurs relations de dépendance et leur étiquette correspondant à notre typologie (cf. figure 13.2).

Pour les textes japonais, nous disposons de résultats semblables à ceux du français (cf. figure 13.3 (voir page suivante)). Selon notre définition (cf. chapitre 7), il existe en japonais deux types de propositions, racine et subordonnée, et les sous-catégories des subordonnées sont : neutre, condition, déterminante, citation, agglutinante, conjonction. En plus de ces propositions, les éléments extérieurs (thème, éléments externes) sont également identifiés.

À partir de ces résultats de détection des propositions, nous construisons un arbre dépendanciel des propositions (« arbre des propositions » ci-après) pour chaque phrase (cf. figure 13.2 pour le français et figure 13.3 pour le japonais).

Nous réalisons ensuite l'alignement des propositions, non pas sur l'ensemble des textes mais sur les paires de phrases alignées une par une.

13.1.2 Fusion de plusieurs phrases en cas d'alignement des phrases non 1-1

À l'étape de création des perles (indexée 4 dans la figure 13.1 page précédente), nous constituons, à l'aide du résultat d'alignement des phrases, les perles (les paires des ensembles de phrases alignées) avec les phrases segmentées en propositions. Par exemple, si la phrase française 235 est alignée avec la phrase japonaise 251, les arbres des propositions de ces phrases constituent une perle et sont passés à la fonction réalisant l'alignement des propositions perle par perle.

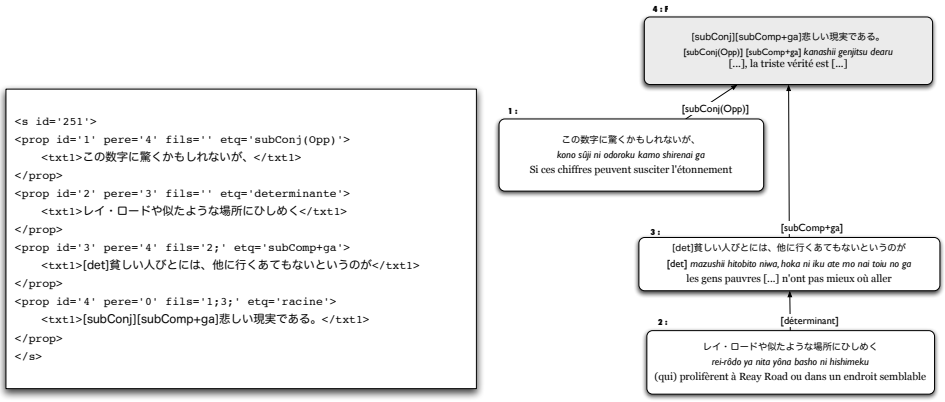


FIG. 13.3 – Résultat de la détection des propositions et arbre construit (JP)

Les paires de phrases alignées peuvent être constituées d'une phrase française et de plusieurs phrases japonaises ou inversement (modèle 1-n ou n-1), voire de plusieurs phrases françaises et japonaises (modèle n-m). Dans ce cas, il faut une opération supplémentaire de fusion des arbres des propositions. Par exemple, si la phrase française 235 est alignée avec l'ensemble des phrases japonaises 251 et 252, on réalise une fusion des arbres des propositions de ces deux phrases afin de constituer un seul arbre des propositions comportant toutes les propositions des phrases 251 et 252, pour lequel on va réaliser un alignement avec l'arbre des proposition de la phrase française 235.

Dans le cadre des présents travaux, nous considérons que la première phrase possède la proposition racine, les propositions racines de toutes les autres phrases (phrases subordonnées, ci-après) dépendant de la proposition racine de la phrase qui les précède directement. La relation syntaxique qu'entretient la proposition racine des phrases subordonnées avec la racine de la phrase précédente est considérée simplement comme une coordination sans chercher à réaliser une analyse fine qui constituerait un sujet de recherche à part entière. Ainsi, dans le cas de la fusion des phrases japonaises 251 et 252 dont nous avons parlé, la construction de l'arbre unique est réalisée de sorte que la proposition racine de la phrase 252 soit régie par la proposition racine de la phrase 251 qui constituera la racine du nouvel arbre fusionné.

13.2 Problèmes et solution adoptée

13.2.1 Difficultés d'appariement des propositions dues aux différences entre les langues

Dans l'article de Kashioka et al. (2003) (cf. § 12.3) présentant leur réalisation d'un alignement manuel des propositions d'un corpus parallèle anglais-japonais,

nous avons constaté deux points qui pouvaient être problématiques lors de la conception d'un système automatique : absence d'unité correspondante dans le texte anglais et beaucoup de croisements des alignements.

En effet, lors de leur expérience d'alignement manuel, les auteurs ont constaté que près de 10% des propositions japonaises n'avaient pas de segment anglais correspondant. Par ailleurs, ils ont signalé la présence de beaucoup de croisements des alignements : l'ordre des propositions japonaises était différent de celui des segments anglais correspondants.

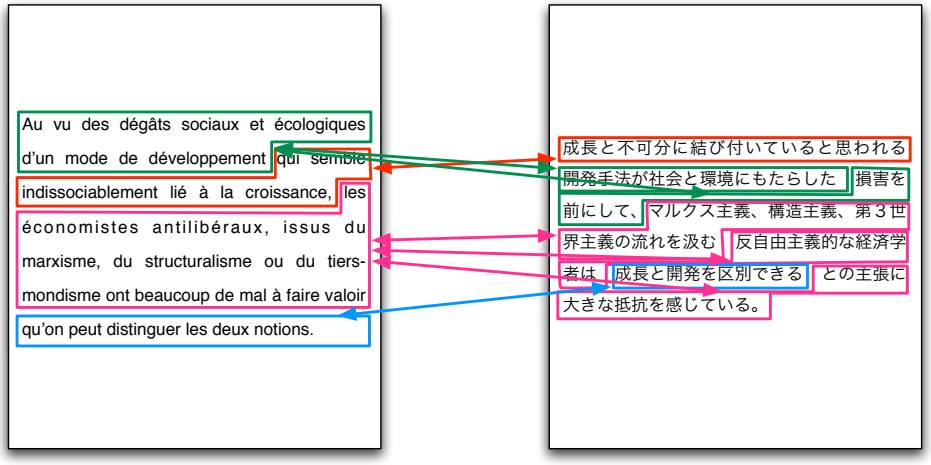


FIG. 13.4 – Exemple de non-parallélisme de l'alignement des propositions français-japonais

13.2.2 Éléments de solution

Notre définition de la proposition japonaise, différente de celle de Kashioka *et al.*, est telle que le premier problème d'absence d'élément correspondant décrit précédemment ne se pose pas – du moins de manière aussi gênante – dans nos travaux : les éléments dits de modalité situés en fin de phrase sont inclus dans la proposition formée par le syntagme prédicatif qui les précède directement.

Ce qui est plus problématique pour notre opération d'alignement est le caractère non-parallèle des propositions en relation de traduction (cf. figure 13.4). De cette observation, nous avons déduit que l'automatisation de cette tâche nécessiterait un algorithme utilisant une structure non linéaire mais à deux dimensions telle que les graphes, et nous avons posé comme hypothèse que les informations sur les relations entre les propositions seraient utiles pour l'alignement de ces unités (cf. figure 13.5 (voir page suivante)).

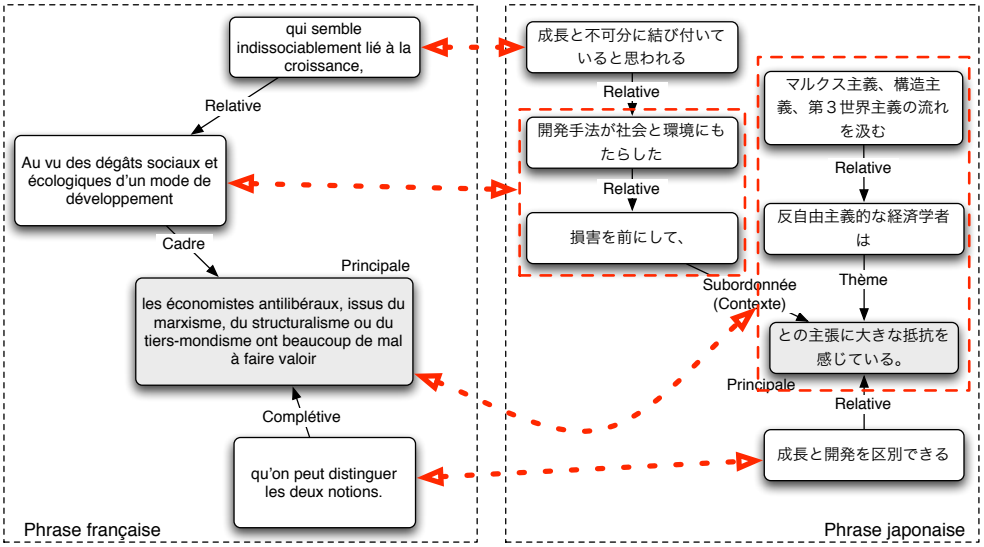


FIG. 13.5 – Alignement des propositions à l'aide de graphes

Notre approche est semblable à celle de Matsumoto et al. (1993) décrite dans la section 12.4. La difficulté est que la recherche de la meilleure décomposition des arbres pour obtenir les structures isomorphes permettant l'appariement maximal revient à un appariement *many-to-many* des graphes, qui est un problème de grande complexité algorithmique. Dans les travaux de Matsumoto, est retenue une stratégie d'amélioration par l'utilisation de la méthode du *branch-and-bound*.

Dans le cadre des présents travaux, nous avons choisi une solution basée sur une technique d'appariement des graphes. En effet, dans la théorie des graphes, il existe un ensemble de méthodes beaucoup plus économiques que les procédures de recherche combinatoire, généralement connues sous le nom de méthodes spectrales.

Néanmoins, cette méthode s'appuie essentiellement sur la topologie des graphes à apparier et n'est pas destinée à exploiter différentes informations disponibles, notamment les informations lexicales dans le cas de nos travaux. La dernière étape de la méthode spectrale, consistant en un regroupement des points projetés, nous a inspiré l'approche pour l'alignement par la classification ascendante hiérarchique (CAH). Celle-ci devant permettre de mieux profiter des informations lexicales tout en supportant les croisements des traductions.

Après examen de l'existant, nous avons réalisé deux méthodes d'alignement des propositions. L'une est basée sur les méthodes d'appariement des graphes – profitant pleinement des structures des arbres des propositions –, l'autre exploitant les informations lexicales et de longueur tout en étant robuste vis-à-vis des croisements de correspondance avec une méthode inspirée de la classification ascendante hiérarchique.

13.3 Méthodes basées sur l'approche spectrale

Dans la théorie des graphes, l'appariement des graphes par une approche spectrale vise à représenter et distinguer les propriétés structurales des graphes à l'aide des valeurs propres et des vecteurs propres de leurs matrices d'adjacence, et se base généralement sur une technique de décomposition spectrale.

L'algorithme sur lequel nous nous sommes plus particulièrement appuyés, celui proposé par Kosinov & Caelli (2002, 2004), est une amélioration des techniques existantes visant en particulier la réalisation d'appariements de graphes inexacts – c'est-à-dire la mise en correspondance des ensembles de nœuds d'une paire de graphes. Lerallut (2006) a ensuite amélioré cette méthode pour prendre en compte des informations supplémentaires en cas d'appariement de graphes valués.

Dans le cadre de notre alignement des propositions, la méthode de Kosinov est directement utilisée pour apparier les arbres des propositions. Afin d'exploiter au mieux les informations disponibles pour réaliser un meilleur appariement, nous avons également réalisé une adaptation de la méthode de Lerallut à notre opération d'alignement des propositions.

13.3.1 La méthode de Kosinov

La méthode d'appariement des graphes inexacts proposée par Kosinov & Caelli (2002, 2004) combine les avantages des techniques de décomposition spectrale, de projection et de classification (*clustering*).

Elle consiste, étant donné les matrices d'adjacence A_1 et A_2 créées à partir des graphes G_1 et G_2 respectivement :

(i) à calculer les valeurs propres et les vecteurs propres ;

(ii) à tronquer les matrices selon le nombre de dimensions choisies pour la projection ;

(iii) à normaliser les valeurs propres et les vecteurs propres pour projeter ensuite chaque graphe ;

(iv) à réaliser l'appariement par regroupement des nœuds projetés à l'aide d'un algorithme de classification.

Décomposition spectrale

Chaque matrice d'adjacence A créée à partir de chaque graphe est tout d'abord décomposée en produit des matrices de valeurs propres et de vecteurs propres comme :

$$A = VDV^T$$

où V est une matrice de vecteurs propres et D est une matrice diagonale de valeurs propres.

Les méthodes spectrales pures s'appuient uniquement sur les valeurs propres ainsi obtenues, mais ces informations ne sont pas suffisantes pour représenter

pleinement la variabilité des structures de graphes. Cependant ce problème peut être résolu par l'utilisation des valeurs propres avec les vecteurs propres associés.

Normalisation et projection

L'idée principale provient de la projection sur un sous-espace propre utilisée dans le domaine de l'Analyse en Composantes Principales. Les méthodes de projection sur un sous-espace propre sont destinées à réduire le volume des données en minimisant à la fois le nombre de dimensions et la perte d'informations. Les données originales sont projetées sur le sous-espace propre associé aux k valeurs propres les plus importantes comme spécifié dans l'équation suivante :

$$\widehat{x} = U_k^T x$$

où $\widehat{x}$ est la projection, U_k^T la matrice transposée des k vecteurs propres, et x un élément des données originales. Avec une approche semblable, on peut projeter les données relatives aux nœuds obtenues avec la matrice d'adjacence sur l'ensemble de ses vecteurs propres les plus importants, formant un sous-espace propre d'une dimension réduite du graphe. Dans ce sous-espace propre, des nœuds ou des ensembles de nœuds ayant des propriétés structurales semblables sont proches les uns des autres, permettant ainsi une comparaison et un appariement des graphes.

Néanmoins, étant donné que les graphes à aligner peuvent posséder un nombre différent de nœuds, une opération de normalisation est également nécessaire pour assurer de bonnes conditions de comparaison.

Les matrices de vecteurs propres V et de valeurs propres D sont donc tronquées selon le nombre de dimensions choisies ($k = 2$ dans notre cas) et la normalisation est réalisée comme suit :

$$V'_k = \frac{V_k}{||V_k||}$$

$$D'_k = \frac{D_k}{||D_k||}$$

et la projection de chaque nœud d'un graphe est calculée comme :

$$A' = D'_k (V'_k)^T$$

Il est à noter que pour aligner un ensemble de projections de nœuds d'un graphe avec l'autre ensemble, une correction des signes des coordonnées des projections est également réalisée. Pour chaque vecteur propre, les nombres d'éléments positifs et négatifs sont calculés. Si le nombre d'éléments négatifs est supérieur à celui d'éléments positifs, le vecteur propre est multiplié par -1.

La figure 13.7 page ci-contre montre le résultat de la projection des nœuds des graphes X et Y présentés dans la figure 13.6 page suivante.

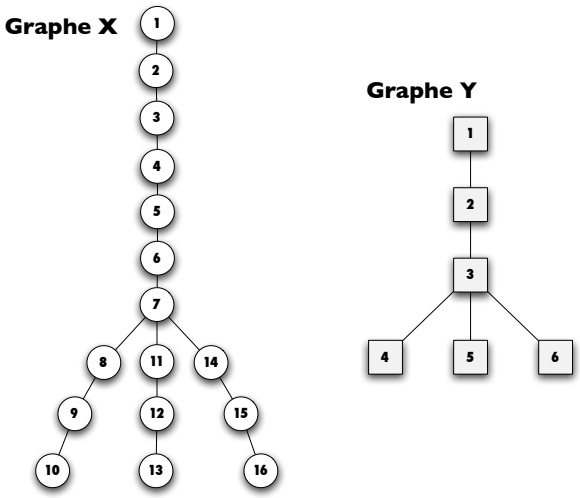


FIG. 13.6 – Deux graphes X et Y

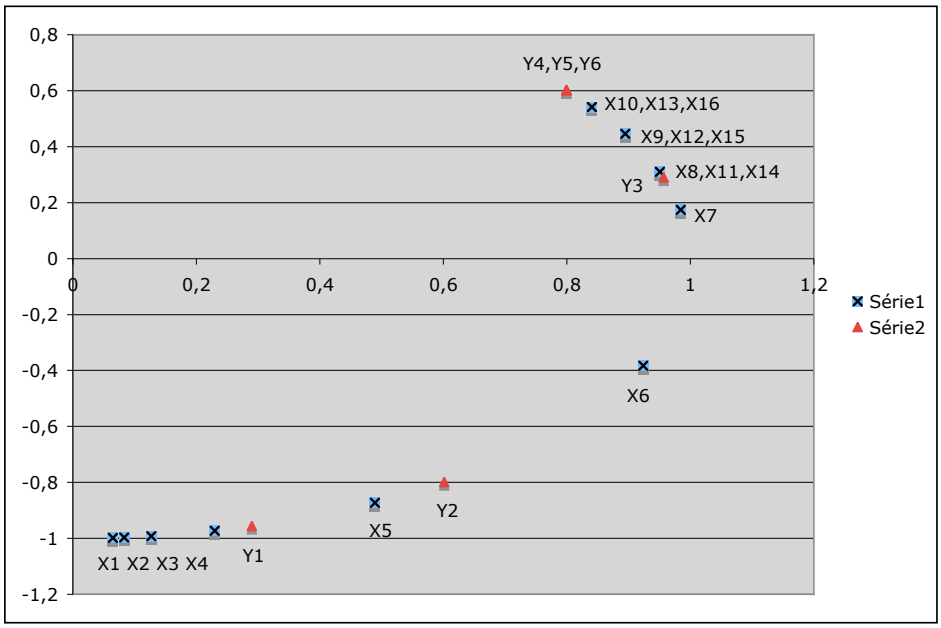


FIG. 13.7 – Projection des nœuds des deux graphes X et Y

Classification (*clustering*)

Par examen du positionnement de ces projections de nœuds, la mise en correspondance est maintenant possible. Le regroupement des points projetés par une méthode de classification ascendante hiérarchique permet de réaliser l'appariement des ensembles de nœuds entre les graphes.

13.3.2 Amélioration pour l'appariement des graphes valués

La méthode de Kosinov réalise la comparaison des graphes uniquement avec leurs caractères topologiques. Les travaux de Lerallut (2006), cherchant à l'appliquer à un traitement des images, proposent une amélioration permettant de prendre en compte des informations supplémentaires en cas d'appariement de graphes valués.

La méthode de Lerallut part du résultat obtenu avec la méthode de Kosinov, qui permet tout d'abord d'obtenir la matrice topologique contenant les distances euclidiennes entre toutes les projections dans le sous-espaces propre :

$$M_{ij}^{topo} = \text{dist}_{\text{géom}}(N_i, N_j), \forall N_i, N_j \in G_1 \cup G_2$$

Les graphes sont ensuite valués par l'affectation de couleurs à chaque nœud, et la matrice des distances de couleurs est calculée entre tous les nœuds des deux graphes :

$$M_{ij}^{couleur} = \text{dist}_{\text{couleur}}(N_i, N_j), \forall N_i, N_j \in G_1 \cup G_2$$

Après avoir normalisé ces deux matrices en les divisant par leur valeur maximum, on calcule une somme pondérée, le coefficient de pondération permettant de choisir l'influence relative de chacune des deux matrices :

$$M^{final} = \alpha \frac{M^{couleur}}{\max(M^{couleur})} + (1-\alpha)\frac{M^{topo}}{\max(M^{topo})}$$

Afin d'écarter les valeurs très distantes, une modification est enfin réalisée comme suit :

$$M_{ij}^{final} = \exp(-\frac{(M_{ij}^{final})^2}{2\sigma^2})$$

Un sous-espace propre de cette matrice est alors calculé afin d'y projeter tous les nœuds.

13.3.3 Application de la méthode spectrale à l'alignement des propositions

L'alignement des propositions réalisé par la méthode de Kosinov s'appuie uniquement sur la topologie des graphes. Toutefois, les arbres des propositions dont nous disposons comme entrée du système contiennent beaucoup plus d'informations qui pourraient être utilisées au profit d'un bon appariement.

Afin d'exploiter au mieux ces informations disponibles, nous avons tout d'abord tenté d'adapter la méthode de Lerallut de sorte que les graphes à apparier soient valués, non par l'affectation de couleurs, mais selon les types de propositions. Mais, afin de calculer la distance entre deux nœuds sur la base de leur type de proposition, il nous a d'abord fallu définir une distance entre chaque type de proposition.

Distances entre les types de nœuds

Nous avons tout d'abord pensé, naturellement, à l'utilisation de probabilités de correspondance. Néanmoins, nous ne disposons à l'heure actuelle d'aucun corpus *ad hoc* – c'est-à-dire de corpus parallèle français-japonais aligné au niveau des propositions – et ce en quantité suffisante pour le calcul de ces probabilités. Aussi, avons-nous choisi une méthode plus empirique, qui présente l'avantage de permettre de constituer un premier corpus pour des travaux futurs.

Nous avons d'abord mis en correspondance les types de propositions du français et du japonais, qui semblaient les plus proches sur le plan syntaxique. La classification des types utilisés dans les deux langues (cf. § 13.1) est définie dans le tableau 13.8.

Propositions françaises	Propositions japonaises
racine	racine
coordonnée	neutre sub. conjonctive
sub. post-nominale	déterminante
sub. post-verbale sub. à position SN	sub. agglutinante (une partie) sub. de citation
sub. périphérique incidente	sub. agglutinante sub. de condition
élé. extra-prédicatif	élé. externe
thème	thème

TAB. 13.8 – Classification des types de propositions communes aux français et japonais

Nous avons ensuite posé comme hypothèse qu'étant donné l'existence d'un lien non négligeable entre les fonctions syntaxiques et la place dans la phrase, il était possible de définir une distance entre chaque type de proposition sur la base de la topologie de la phrase. À cette fin, nous nous sommes appuyé sur la structure canonique de la phrase française (cf. figure 13.9 (voir page suivante)).

La racine est définie comme zéro, point central de la phrase. La zone du noyau syntaxique, constitué autour du prédicat, s'étendant à gauche est définie ensuite comme zone positive, par opposition à la partie initiale comportant des éléments

extra-prédicatifs formant la zone négative. Le principe de base est que la distance d'un type donné de proposition par rapport à la racine est définie par le nombre de propositions susceptibles d'apparaître entre elles, soit $\beta \times (n+1)$ où n est le nombre de propositions intermédiaires.

La distance entre la racine et la subordonnée post-nominale pouvant s'insérer à l'intérieur de cette première est définie comme $\frac{\beta}{2}$. Il en va de même pour celle entre l'élément extra-prédicatif et le thème. En effet, dans les travaux linguistiques sur le français, le thème est généralement classé dans la catégorie des éléments extra-prédicatifs, mais du fait de son statut central et particulier dans la phrase japonaise, ce type est défini à part dans le cadre de la présente étude. Par ailleurs, dans la phrase française, l'élément thématisé coïncide souvent avec le sujet, apparaissant ainsi à l'intérieur de la racine. La distance du thème par rapport à la racine est donc également définie comme un demi. Enfin, la distance entre la coordonnée et la subordonnée périphérique est elle-aussi définie comme un demi. Selon ces définitions, toutes les propositions x sont caractérisées par leur distance par rapport à la racine notée $\text{dist}(x)$ et la distance entre les deux propositions x et y est obtenue par $|\text{dist}(x) - \text{dist}(y)|$. Par exemple, avec $\beta = 10$, la distance entre le thème et la subordonnée post-verbale est $|-5-15| = 20$.

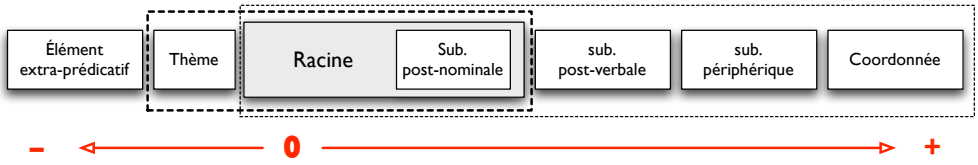


FIG. 13.9 – Structure canonique de la phrase française

Utilisation de la distance des types dans l'appariement des graphes

Nous avons utilisé les distances ainsi définies pour calculer la matrice de distances des couleurs et réalisé l'appariement des graphes avec la méthode de Lerallut.

Avec le poids $\alpha = 0,5$ comme coefficient de pondération, l'appariement ne reflétait pas bien les distances des types de propositions. Il arrivait souvent que des nœuds censés se rapprocher suite à l'introduction des distances des types de proposition s'éloignent même.

Avec un poids plus élevé comme $\alpha = 0,8$, les nœuds des types correspondants pouvaient se rapprocher, mais les relations de dépendance syntaxiques étaient mal conservées.

Amélioration du calcul de la matrice finale

Afin de mieux refléter les informations sur les types de nœuds tout en conservant la structure des arbres d'entrée (c'est-à-dire les relations entre les propositions), nous avons introduit une autre formule pour calculer la matrice finale obtenue par combinaison de ces deux informations.

Le principe du nouveau calcul consiste à prendre en compte des informations topologiques pour les relations entre les nœuds du même arbre et des informations sur les types pour les distances entre les nœuds des différents arbres.

Étant donné les deux graphes X et Y, la matrice finale de la méthode de Lerallut est une matrice M_{final} de $|X|+|Y| \times |X|+|Y|$, $M_{final}(i,j)$ correspondant à la somme des distances topologique et de type normalisées entre les nœuds i et j.

Nous décomposons cette matrice finale comme :

$$M_{final} = \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix}$$

de manière à obtenir les sous-matrices M_{11} comme une matrice $|X| \times |X|$, M_{12} comme $|X| \times |Y|$, M_{21} comme $|Y| \times |X|$ et M_{22} comme $|Y| \times |Y|$, où :

$$M_{11}(i,j) = \text{dist}_{\text{topo}}(X_i, X_j) \times (1-\alpha)$$
$$M_{12}(i,j) = \text{dist}_{\text{type}}(X_i, Y_j) \times \alpha$$
$$M_{21}(i,j) = \text{dist}_{\text{topo}}(X_j, Y_i) \times (1-\alpha)$$
$$M_{22}(i,j) = \text{dist}_{\text{type}}(Y_i, Y_j) \times \alpha$$

13.3.4 La méthode du *Clustering*

Kosinov explique que le regroupement des points projetés peut être réalisé par une méthode de classification ascendante hiérarchique classique[1]. Seules quelques modifications sont nécessaires pour favoriser le regroupement des points appartenant à des arbres différents plutôt que ceux du même arbre. Pour ce faire, il propose de mettre les distances entre deux points du même arbre à une valeur plus élevée (valeur 2 conseillée) que celles des distances entre deux points d'arbres différents.

Cependant, définir des distances élevées communes entre deux points du même arbre entraîne une perte d'information nécessaire à une classification adéquate.

Pour favoriser le regroupement des points appartenant à des arbres différents tout en conservant les informations pertinentes liées aux distances entre deux points du même arbre, nous avons pondéré les distances entre deux points du même arbre plutôt que de les fixer à une valeur donnée (poids $\alpha = 2$ dans nos travaux).

Par ailleurs, pour obtenir plusieurs petits groupes plutôt qu'une seule grande classe, nous avons voulu pénaliser les regroupements des ensembles déjà fusion-

[1] Les notions générales de la classification ascendante hiérarchique seront présentées dans § 13.4.1.

nés. À cet effet, nous avons choisi un indice d'agrégation défini par la méthode du diamètre (*complete linkage*, cf. § 13.4.1) avec laquelle les distances entre classes regroupées sont déterminées par la plus grande distance existant entre deux points de classes différentes (c'est-à-dire les voisins les plus éloignés).

13.4 Méthode inspirée de la classification ascendante hiérarchique (CAH)

La deuxième méthode que nous avons décidé d'étudier est basée sur la classification ascendante hiérarchique, celle-ci devant permettre de mieux profiter des informations lexicales tout en étant robuste vis-à-vis des croisements des traductions. En effet, nous considérons maintenant l'alignement, comme nous l'avons fait à la dernière étape dans les méthodes spectrales, comme le regroupement des points semblables appartenant à deux classes différentes. Le facteur supplémentaire à considérer est, tout comme dans les méthodes classiques (de programmation dynamique ou de chaînes de Markov), de favoriser la constitution de plusieurs petites perles plutôt qu'une seule grande.

Avant d'entrer dans la description de la méthode, nous présentons très brièvement la définition et le principe général des méthodes de CAH. L'exposé sur notre méthode commencera par la présentation de la procédure générale, suivie de la description détaillée des trois matrices utilisées.

13.4.1 Définition et principe général des méthodes de CAH

Notre étude sur les généralités des méthodes de CAH est essentiellement basée sur les ouvrages (Lebart et al., 2006 ; Carpentier, 2005).

Classification automatique

La classification automatique consiste à produire des classes d'objets à partir d'un certain nombre de variables ou de caractères et elle s'oppose aux techniques de « classement » visant à affecter des objets à des classes préalablement identifiées. Les techniques de classification recourent à une procédure algorithmique : une série d'opérations est définie de manière répétitive. Plusieurs types d'algorithmes de classification existent : les méthodes de partitionnement, les algorithmes ascendants et descendants.

Les algorithmes ascendants (ou agglomératifs) réalisent la construction des classes par agglomérations successives de façon ascendante des éléments deux à deux et fournissent une hiérarchie de partitions des objets.

Principe de CAH

Les grandes lignes de l'algorithme de CAH sont comme suit :

1. nombre initial de classes $k = n$ (n = nombre d'éléments à classer) ;

2. construction d'une première matrice de distances M de $n \times n$ telle que $M(i, j)$ contienne la distance entre les éléments i et j ;
3. répétition des opérations suivantes jusqu'à n'avoir plus qu'un seul élément regroupant tous les objets ($k = 1$) ;
   a) recherche dans la matrice de distances des deux éléments les plus proches, que l'on agrège en un nouvel objet : la nouvelle partition obtenue est à $k = k - 1$ classes ;
   b) construction d'une nouvelle matrice de distances de $k \times k$ en calculant les distances entre le nouveau groupe et les autres éléments (les autres distances restant à la même valeur).

Tous les éléments à classer doivent disposer de coordonnées permettant de calculer et recalculer leurs distances. Différentes mesures de cette distance (ou dissimilarité) existent : distance euclidienne, distance euclidienne au carré, distance à la puissance, etc.

Après le regroupement des deux objets (algo. 3a), il faut choisir une distance entre le nouveau groupe et les autres éléments (algo. 3b). Ce qui revient à définir les règles de calcul des distances entre des groupements disjoints d'individus, dites **critères d'agrégation**. Un grand nombre de solutions sont également proposées. Par exemple, soient x et y les éléments regroupés en une classe s, on définit la distance entre ce nouveau groupe et l'élément t par la plus petite distance existante entre différents éléments de ce premier et ce second :

$$d(s, t) = \min(d(x, t), d(y, t))$$

Cette distance, appelée saut minimum (*single linkage*), est un critère d'agrégation. Il existe également le saut maximum (*complete linkage*) ou diamètre – utilisant la plus grande distance –, celui basé sur la distance moyenne, ou encore la technique recourant à l'analyse de la variance.

Exemple d'application de la méthode CAH

Afin de mieux illustrer la description précédente, prenons comme exemple la dernière étape de l'appariement des graphes dans les méthodes spectrales décrites précédemment (cf. § 13.3.4) : la procédure de regroupement des nœuds projetés par la classification ascendante hiérarchique.

La figure 13.10 (voir page suivante) présente les points projetés correspondant aux nœuds des graphes ($1, 2, 3 \in G_1$ et $4, 5, 6, 7, 8, 9 \in G_2$) que nous allons regrouper avec une méthode de CAH. Les coordonnées exactes de ces points sont présentées dans la table 13.11 (voir page suivante).

La figure 13.12 page 429 montre la matrice initiale et toutes les matrices créées après agrégation des deux éléments (ou classes).

À l'étape initiale (étape 1 de la figure 13.12), il existe neuf classes (A~I) correspondant chacune à un point projeté (1~9). La matrice initiale est créée en calculant les distances entre tous les neuf points projetés à partir de leurs coordonnées :

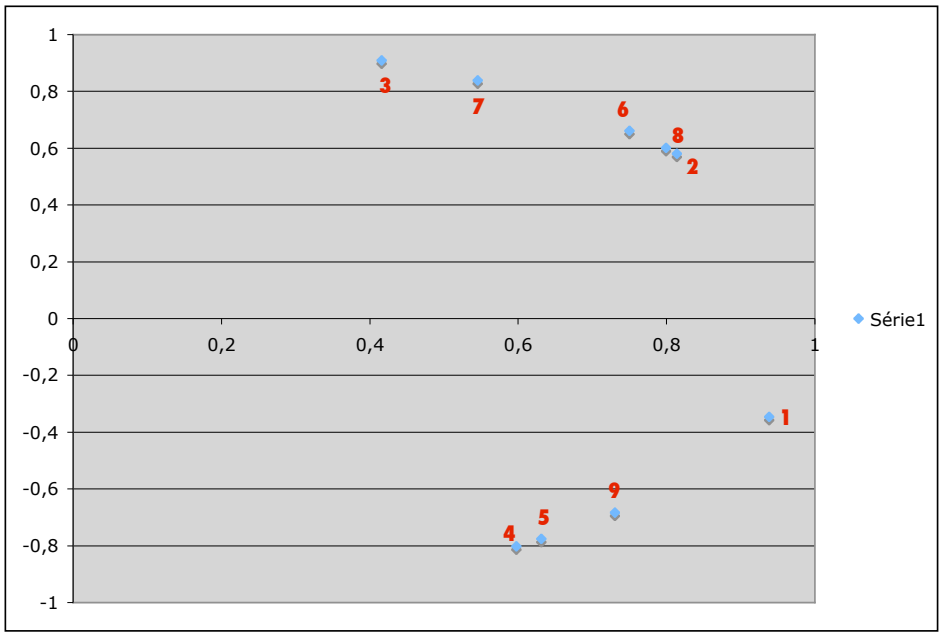
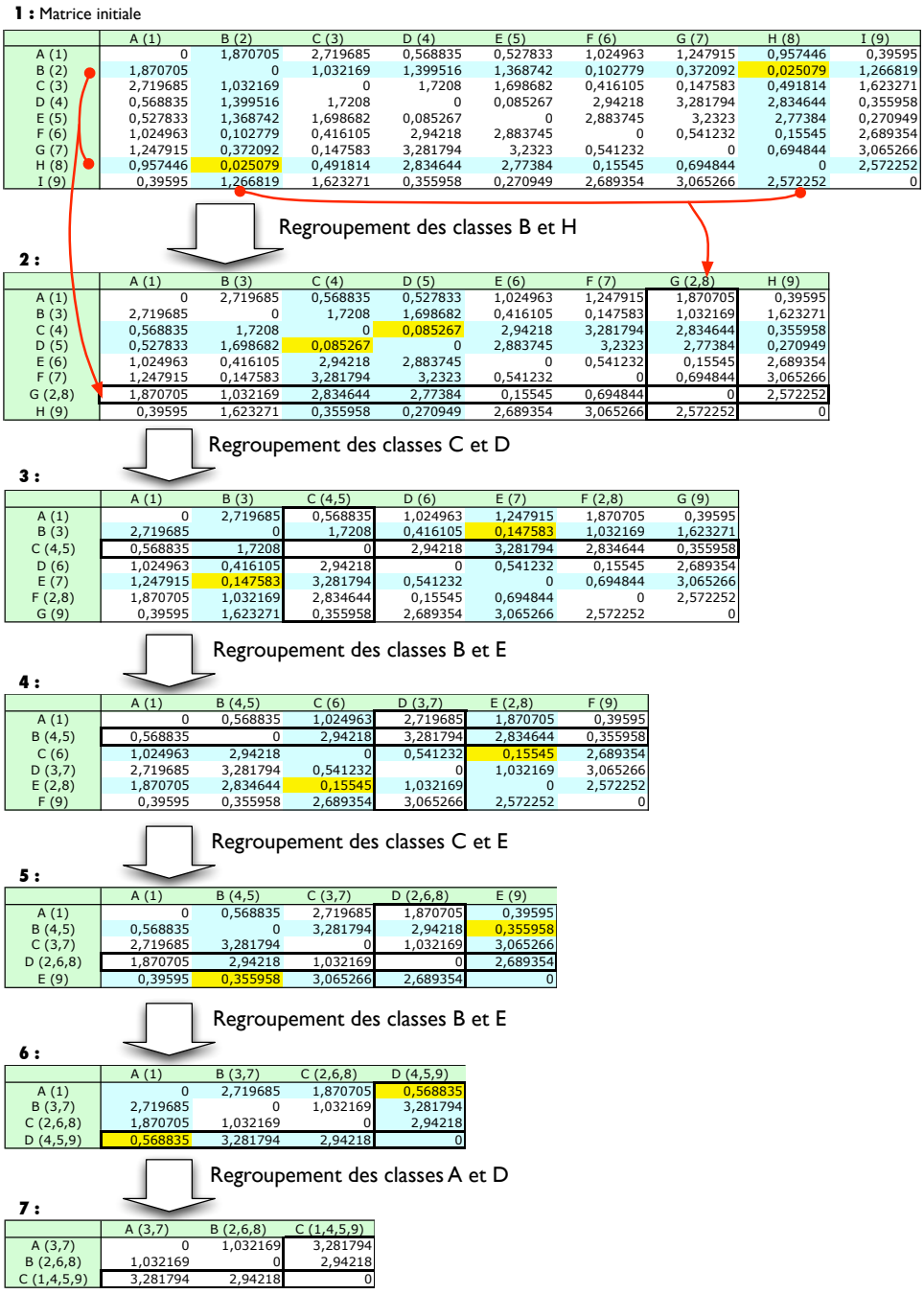


FIG. 13.10 – Nœuds projetés à regrouper

	X	Y
1	0,938137	-0,346263
2	0,814032	0,580819
3	0,416026	0,909353
4	0,597527	-0,801849
5	0,631162	-0,775651
6	0,750116	0,661306
7	0,545335	0,838218
8	0,799211	0,601050
9	0,730209	-0,683223

TAB. 13.11 – Coordonnées des points projetés



	A (1)	B (2)	C (3)	D (4)	E (5)	F (6)	G (7)	H (8)	I (9)
A (1)	0	1,870705	2,719685	0,568835	0,527833	1,024963	1,247915	0,957446	0,39595
B (2)	1,870705	0	1,032169	1,399516	1,368742	0,102779	0,372092	0,025079	1,266819
C (3)	2,719685	1,032169	0	1,7208	1,698682	0,416105	0,147583	0,491814	1,623271
D (4)	0,568835	1,399516	1,7208	0	0,085267	2,94218	3,281794	2,834644	0,355958
E (5)	0,527833	1,368742	1,698682	0,085267	0	2,883745	3,2323	2,77384	0,270949
F (6)	1,024963	0,102779	0,416105	2,94218	2,883745	0	0,541232	0,15545	2,689354
G (7)	1,247915	0,372092	0,147583	3,281794	3,2323	0,541232	0	0,694844	3,065266
H (8)	0,957446	0,025079	0,491814	2,834644	2,77384	0,15545	0,694844	0	2,572252
I (9)	0,39595	1,266819	1,623271	0,355958	0,270949	2,689354	3,065266	2,572252	0

	A (1)	B (3)	C (4)	D (5)	E (6)	F (7)	G (2,8)	H (9)
A (1)	0	2,719685	0,568835	0,527833	1,024963	1,247915	1,870705	0,39595
B (3)	2,719685	0	1,7208	1,698682	0,416105	0,147583	1,032169	1,623271
C (4)	0,568835	1,7208	0	0,085267	2,94218	3,281794	2,834644	0,355958
D (5)	0,527833	1,698682	0,085267	0	2,883745	3,2323	2,77384	0,270949
E (6)	1,024963	0,416105	2,94218	2,883745	0	0,541232	0,15545	2,689354
F (7)	1,247915	0,147583	3,281794	3,2323	0,541232	0	0,694844	3,065266
G (2,8)	1,870705	1,032169	2,834644	2,77384	0,15545	0,694844	0	2,572252
H (9)	0,39595	1,623271	0,355958	0,270949	2,689354	3,065266	2,572252	0

	A (1)	B (3)	C (4,5)	D (6)	E (7)	F (2,8)	G (9)
A (1)	0	2,719685	0,568835	1,024963	1,247915	1,870705	0,39595
B (3)	2,719685	0	1,7208	0,416105	0,147583	1,032169	1,623271
C (4,5)	0,568835	1,7208	0	2,94218	3,281794	2,834644	0,355958
D (6)	1,024963	0,416105	2,94218	0	0,541232	0,15545	2,689354
E (7)	1,247915	0,147583	3,281794	0,541232	0	0,694844	3,065266
F (2,8)	1,870705	1,032169	2,834644	0,15545	0,694844	0	2,572252
G (9)	0,39595	1,623271	0,355958	2,689354	3,065266	2,572252	0

	A (1)	B (4,5)	C (6)	D (3,7)	E (2,8)	F (9)
A (1)	0	0,568835	1,024963	2,719685	1,870705	0,39595
B (4,5)	0,568835	0	2,94218	3,281794	2,834644	0,355958
C (6)	1,024963	2,94218	0	0,541232	0,15545	2,689354
D (3,7)	2,719685	3,281794	0,541232	0	1,032169	3,065266
E (2,8)	1,870705	2,834644	0,15545	1,032169	0	2,572252
F (9)	0,39595	0,355958	2,689354	3,065266	2,572252	0

	A (1)	B (4,5)	C (3,7)	D (2,6,8)	E (9)
A (1)	0	0,568835	2,719685	1,870705	0,39595
B (4,5)	0,568835	0	3,281794	2,94218	0,355958
C (3,7)	2,719685	3,281794	0	1,032169	3,065266
D (2,6,8)	1,870705	2,94218	1,032169	0	2,689354
E (9)	0,39595	0,355958	3,065266	2,689354	0

	A (1)	B (3,7)	C (2,6,8)	D (4,5,9)
A (1)	0	2,719685	1,870705	0,568835
B (3,7)	2,719685	0	1,032169	3,281794
C (2,6,8)	1,870705	1,032169	0	2,94218
D (4,5,9)	0,568835	3,281794	2,94218	0

	A (3,7)	B (2,6,8)	C (1,4,5,9)
A (3,7)	0	1,032169	3,281794
B (2,6,8)	1,032169	0	2,94218
C (1,4,5,9)	3,281794	2,94218	0

FIG. 13.12 – Exemple de regroupement des nœuds projetés par la classification ascendante hiérarchique (CAH)

l'élément $M(i,j)$ de la matrice initiale est la distance entre le point x de la classe i et le point y de la classe j. Par exemple, l'élément $M(A,B)$ de la matrice initiale correspond à la distance entre le point 1 de la classe A et le point 2 de la classe B, soit 1,870705, calculé par

$$\sqrt{|0,938137-0,814032|^2+|-0,346263-0,580819|^2}\cdot 2$$

la distance étant pondérée (poids = 2) car les points 1 et 2 appartiennent au même graphe.

Dans cette matrice initiale, on cherche les deux points les plus proches : la valeur minimum (sur fond jaune) étant à 0,025079 pour les éléments $M(B,H)$ et $M(H,B)$, nous réalisons ensuite l'agrégation des voisins les plus proches, B et H.

Après la première agrégation (étape 2), les anciennes classes B et H sont regroupées et constituent maintenant une nouvelle classe G, et toutes les autres classes restent telles qu'elles étaient, quoique renommées. Le critère d'agrégation étant la méthode du diamètre (*complete linkage*) – afin de pénaliser le regroupement avec la classe résultant d'une agrégation antérieure pour obtenir le maximum de petites classes –, la distance entre la nouvelle classe G et une autre classe x est déterminée par la plus grande distance entre une des deux classes regroupées et la classe x. Par exemple, la distance entre la classe A et la nouvelle classe G est la plus grande des deux distances, distance entre la classe A et l'ancienne classe B et celle entre la classe A et l'ancienne classe H. La première étant plus grande que la seconde, la distance entre la classe A et la nouvelle classe G, $M(A,G)$ et $M(G,A)$, est à 1,870705.

Dans cette deuxième matrice, la valeur minimum étant à 0,085267 pour les éléments $M(C,D)$ et $M(D,C)$, nous recommençons l'agrégation des voisins les plus proches, C et D, qui constituent ensemble la nouvelle classe C. La valeur minimum de la troisième matrice nouvellement créée étant les éléments $M(B,E)$ et $M(E,B)$, les classes B et E sont regroupées et constituent dans la quatrième matrice la nouvelle classe D. Les classes C et E de cette quatrième constituent à leur tour la nouvelle classe D dans la cinquième matrice, puis les classes B et E forment la nouvelle classe D dans la sixième matrice, enfin l'agrégation des classes A et D entraîne la création de la nouvelle classe C dans la septième matrice.

Au bout de ces six agrégations, nous obtenons trois classes : la classe A regroupant les points 3 et 7, la classe B regroupant les points 2, 6 et 8, et la classe C regroupant les points 1, 4, 5 et 9. La procédure se termine à cet état, car dans cette application le contrôle d'arrêt est défini comme étant atteint lorsque tous les points sont regroupés avec au moins un point appartenant à l'autre graphe.

13.4.2 Procédure générale de l'alignement basé sur CAH

Nous décrivons maintenant le fonctionnement d'une méthode d'alignement inspirée de cet algorithme de CAH.

Soient deux (ensembles de) phrases dans deux langues différentes de m propositions et de n propositions, nous créons tout d'abord deux matrices de ($m+$

$n) \times (m+n)$: matrice de similarité ($M_{similarité}$) contenant les similarités de chaque paire de propositions, et matrice d'évolution du rapport des longueurs ($M_{raplong}$) pour stocker les valeurs indiquant l'évolution du rapport des longueurs entre les propositions de langues différentes. L'évolution du rapport des longueurs correspond au changement du rapport des longueurs entre les propositions de langues différentes, qui se produira si le regroupement des deux éléments considérés a lieu.

En combinant ces deux matrices de similarité et de rapport des longueurs, nous créons une troisième matrice, matrice courante ($M_{courante}$), dans laquelle nous cherchons la valeur minimum pour réaliser l'agrégation des deux éléments.

Après l'agrégation des deux éléments, nous recalculons la matrice de similarité selon le critère d'agrégation adopté. La matrice de rapport des longueurs est également recalculée, tenant compte du changement de longueurs des éléments regroupés. À partir de ces deux matrices nouvellement calculées, nous calculons à nouveau la matrice courante et recommençons les opérations d'agrégation tout comme la CAH.

À la différence de l'algorithme de CAH décrit précédemment, l'itération s'arrête dans notre opération dès que toutes les propositions sont regroupées avec au moins une proposition de l'autre langue.

13.4.3 Matrice de similarité

Structure générale

Étant donné les deux (ensembles de) phrases X de m propositions et Y de n propositions, la matrice de similarité M de $(m+n) \times (m+n)$ est définie comme :

$$M_{similarité} = \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix}$$

avec les sous-matrices M_{11} de $m \times m$, M_{12} de $m \times n$, M_{21} de $n \times m$ et M_{22} de $n \times n$, où :

$$M_{11}(i,j) = synt(X_i, X_j)$$
$$M_{12}(i,j) = simlex(X_i, Y_j)$$
$$M_{21}(i,j) = simlex(X_j, Y_i)$$
$$M_{22}(i,j) = synt(Y_i, Y_j)$$

$simlex(X_i, Y_j)$ est la similarité lexicale obtenue de manière classique telle qu'avec le coefficient de Dice (la définition exacte dans notre réalisation est présentée ci-après). Lorsque la similarité lexicale est nulle, on lui donne la valeur minimum α pour favoriser la fusion des éléments (propositions) appartenant à des classes différentes.

$synt(X_i, X_j)$ est obtenue de la même manière qu'une matrice d'adjacence, c'est-à-dire 0 s'il n'existe aucun arc entre les nœuds i et j dans l'arbre d'entrée, et β s'il en existe un. Ce mécanisme permet en fait, dans le cas du regroupement

d'éléments appartenant à la même classe, de réaliser l'agrégation entre deux éléments en relation syntaxique, plutôt qu'entre deux éléments qui n'en ont aucune.

Calcul de la similarité lexicale

Dans la présente réalisation, nous avons utilisé pour le calcul de la similarité lexicale une méthode basée sur le coefficient de Dice :

$$simlex(X_i, Y_j) = \frac{2 \cdot trad(X_i, Y_j)}{lg(X_i) + lg(Y_j)}$$

où

- $trad(P_1, P_2)$ est la fonction fournissant le nombre de couples de mots M_1 et M_2 ($M_1 \in P_1$ et $M_2 \in P_2$) en relation de traduction ;
- $lg(P)$ est la longueur de P en nombre de mots.

Pour le calcul de $trad(P_1, P_2)$, deux types de ressources sont utilisés : une liste de mots alignés au moment de l'alignement des phrases du même corpus (cf. ch. 3) et un dictionnaire bilingue.

La première ressource contient les paires de transfuges, de mots en *katakana* et de lemmes. La partie de liste correspondant aux paires de lemmes est vérifiée, bien que rapidement, manuellement car celle-ci contient plus de bruit que la partie correspondant aux deux premiers. L'avantage d'utiliser cette ressource est d'une part la fiabilité de leur relation traductionnelle, et d'autre part, surtout, le fait que les deux premiers types sont souvent absents des dictionnaires bilingues.

Le dictionnaire bilingue que nous utilisons est créé à partir de deux dictionnaires publiques existants : le dictionnaire japonais-français de Jean-Marc Desperrier (DicoJF ci-après[2]), réalisé à partir du dictionnaire japonais-anglais Edict compilé par Jim Breen ; le dictionnaire multilingue JMdict de Jim Breen[3], créé également à partir de Edict et dont une grande partie des définitions françaises proviennent du premier. Nous avons tout d'abord extrait du JMdict 52 367 entrées contenant au moins une définition française. Ces entrées ont ensuite été réorganisées de manière à obtenir des paires 1-1 des termes japonais et français. Ainsi, nous avons obtenu une liste de 96 210 paires de mots japonais-français en relation de traduction, dans laquelle nous avons finalement ajouté 97 paires extraites du DicoJF qui ne figuraient pas dans le JMdict.

Recherche des couples de mots en relation de traduction à l'aide des dictionnaires

Toutefois, cette liste n'était toujours pas adaptée à la recherche des couples de mots en relation de traduction. Le problème relevait de la différence de nature des unités japonaises et françaises que nous considérons comme constituants de la phrase. En effet, les unités que l'analyseur morphologique japonais fournit

[2] `http://dico.fj.free.fr/index.php`

[3] `http://www.csse.monash.edu.au/~jwb/edrdg/licence.html`

comme résultat de son analyse correspondent souvent aux unités du français plus petites que les mots, et les unités lexicales que nous utilisons pour la recherche des mots en relation de traduction correspondent souvent non pas aux unités lexicales du français mais seulement à leur radical[4].

Par exemple, la séquence japonaise 宗教の (*shûkyô - no*, religion - [*no*]) trouve généralement son unité correspondante, l'adjectif « religieux(se) », dans le texte français. L'analyseur morphologique segmente cette séquence japonaise en deux unités, l'unité lexicale 宗教 (*shûkyô*, religion) d'une part et la particule の (*no*) de l'autre. Dans le dictionnaire, figure le paire « 宗教 (*shûkyô*) - religion », or la simple consultation du dictionnaire ne permet pas de déduire le lien entre cette définition « 宗教 (*shûkyô*) - religion » et la paire des mots appartenant aux phrases alignées « 宗教 (*shûkyô*) - religieuse ». Le problème est identique pour les constituants déterminants des mots composées. En effet, en japonais, on peut créer des mots composés par juxtaposition de deux ou plusieurs mots en idéogrammes *kanji*. Dans ces mots composés, n'est nécessaire aucun élément morphologique ou lexical indiquant la fonction du déterminant des constituants antéposés. Ainsi, la juxtaposition de 経済 (*keizai*, économie) et 成長 (*seichô*, croissance) constitue le mot composé 経済成長 « croissance économique ». Lorsque la phrase initiale japonaise contient ce mot composé 経済成長 et la phrase française « croissance économique », nous avons au moment du calcul de similarité lexicale la liste des mots japonais contenant 経済 (*keizai*, économie) et 成長 (*seichô*, croissance) et la liste des mots français comportant « croissance » et « économique ». La paire « 成長 (*seichô*) - croissance » figure dans le dictionnaire mais la mise en relation des mots 経済 (*keizai*) et « économique » n'est pas aussi évidente.

Pour résoudre ce problème, nous avons introduit le calcul des similarités des chaînes lors de la consultation du dictionnaire. La procédure de recherche des couples de mots en relation de traduction à l'aide du dictionnaire se déroule alors comme suit (cf. figure 13.13 (voir page suivante)) :
Pour tous les mots J_x de la liste japonaise :

- consulter le dictionnaire afin de constituer la liste des mots traductions en français T_k ;
- calculer la similarité entre toutes les traductions obtenues T_k et tous les mots de la liste française F_j ;
- si la similarité d'un couple d'une traduction T_w et un mot de la liste française F_y dépasse le seuil prédéfini, alors la paire du mot japonais J_x et du mot français F_y est considérée comme un couple de mots en relation de traduction.

Pour le calcul de la similarité des chaînes X, Y, nous utilisons la formule suivante basée sur le coefficient de Dice :

$$simch(X,Y) = \frac{2 \cdot spc(X,Y)}{lg(X) + lg(Y)}$$

où

[4]Voir aussi § 3.2.

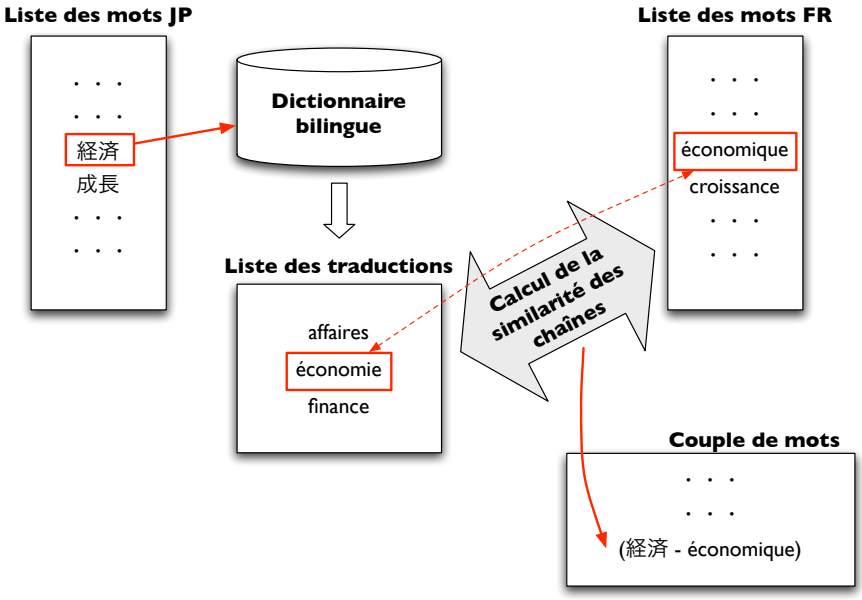


FIG. 13.13 – Recherche des couples de mots en relation de traduction à l'aide du dictionnaire

– $spc(M_F, M_T)$ est la fonction fournissant la longueur de la sous-chaîne préfixale commune des M_F et M_T (M_F appartenant à la liste des mots français et M_T à la liste des traductions obtenues par consultation du dictionnaire) ;
– $lg(M)$ est la longueur de la chaîne M en nombre de caractères.

Toutefois, afin d'éviter de favoriser les chaînes de traduction courtes, lorsque la chaîne de traduction M_T est plus courte que le mot de la liste française M_F, on ne tient pas compte de la longueur de cette première, et la similarité est obtenue par la formule :

$$simch(M_F, M_T) = \frac{spc(M_F, M_T)}{lg(M_F)}$$

Critère d'agrégation pour la matrice de similarité

Chaque fois qu'un regroupement de deux éléments est réalisé, la matrice de similarité est à nouveau calculée. Dans notre réalisation, les similarités liées à la classe nouvellement créée suite à l'agrégation sont obtenues en divisant la somme des similarités des éléments regroupés par la valeur v calculée sur la base du nombre de propositions faisant partie de cette nouvelle classe, tenant compte de la valeur maximum contenue dans la matrice de similarité.

La valeur v est définie plus précisément comme suit :
– si la valeur maximum de la matrice de similarité est strictement supérieure au premier seuil défini (0,3 dans notre réalisation), la valeur v est égale au

nombre de propositions puissance 2 ;
- si la valeur maximum est inférieure au premier seuil et strictement supérieure au second seuil défini (0,07 dans notre réalisation), la valeur v est égale au nombre de propositions ;
- sinon la valeur v est à 2 (c'est-à-dire la similarité résultant du regroupement est la similarité moyenne).

Ce mécanisme de dégradation a pour but de faire diminuer l'influence de la similarité dans le calcul final de la matrice courante, par réduction des écarts entre toutes les similarités lorsque les similarités deviennent toutes des valeurs faibles. La définition de toutes ces valeurs ayant été réalisée empiriquement, il pourrait être intéressant de les réexaminer, voire redéfinir, dans des expériences futures.

13.4.4 Matrice d'évolution du rapport des longueurs

La matrice d'évolution du rapport des longueurs $M_{raplong}$ est définie telle qu'à chacun de ses éléments $M_{raplong}(i,j)$ corresponde l'évolution pondérée du rapport des longueurs entre les propositions de langues différentes, qui se produira si le regroupement des deux éléments considérés, i et j, a lieu :

$$M_{raplong}(i,j) = \Big(rap\big(F(i,j),J(i,j)\big) - \min\big(rap\big(F(i),J(i)\big),rap\big(F(j),J(j)\big)\big)\Big)\cdot a$$

où
- $rap(x,y)$ est le rapport des longueurs normalisées des éléments (ou des classes) x et y ;
- $F(x)$ (resp. $J(x)$) est la longueur de (l'ensemble des) proposition(s) française(s) (resp. japonaise(s)), constituant l'élément (ou la classe) x ;
- $F(x,y)$ (resp. $J(x,y)$) est la longueur de l'ensemble des propositions françaises (resp. japonaises) constituant la classe regroupant les éléments (ou les classes) x et y ;
- a est le poids défini comme le logarithme de la moyenne des deux longueurs normalisées $\alpha = \log\Big(\frac{((F(i)+F(j))\times R)+(J(i)+J(j))}{2}\Big)$.

Le rapport des longueurs est calculé à partir de valeurs stockées dans le tableau des longueurs, puis la matrice d'évolution du rapport des longueurs et le tableau des longueurs sont recalculés, toujours de la même manière, après chaque agrégation de deux éléments.

Tableau des longueurs

Le tableau des longueurs T_{lg} contient des paires de valeurs, l'une $F(i)$ correspondant à la somme des longueurs des propositions françaises et l'autre $J(i)$ à celle des propositions japonaises :

$$T_{lg}(i) = (F(i),J(i))$$

Notons que nous parlerons souvent de ces deux éléments constituant un élément du tableau séparément et noterons tout simplement $F(i)$ et $J(i)$ pour faire référence aux valeur telles que $T_{lg}(i) = (F(i), J(i))$.

La valeur $T_{lg}(s)$ de l'élément s représentant la classe constituée de l'ensemble des u propositions françaises PF_i et de l'ensemble des w propositions japonaises PJ_j est :

$$T_{lg}(s) = (F(s), J(s))$$

avec

$$F(s) = \sum_{i=1}^{u} PF_i$$

$$J(s) = \sum_{j=1}^{w} PJ_j$$

À l'étape initiale de l'alignement de m propositions françaises (PF_i, $1 \leq i \leq m$) et n propositions japonaises (PJ_j, $1 \leq j \leq n$), le tableau des longueurs T_{lg} est un tableau de $m+n$ éléments, et tous les éléments ne comportent qu'une seule longueur positive car avant la première agrégation, toutes les classes sont constituées d'une seule proposition d'une langue. Formellement, le tableau des longueurs initial est :

$$\begin{array}{lll} T_{lg}(i) = (lg(PF_i), 0) & \text{si} & 1 \leq i \leq m \\ T_{lg}(i) = (0, lg(PJ_{(i-m)})) & \text{si} & m+1 \leq i \leq m+n \end{array}$$

Après la première agrégation de deux éléments, la valeur de la nouvelle classe des éléments regroupés est mise à jour et le tableau des longueurs devient à $(m+n)-1$ éléments. Cette opération de recalcul est réalisée après chaque agrégation. La valeur de la nouvelle classe r constituée après le regroupement des éléments s et t est :

$$T_{lg}(r) = (F(s) + F(t), J(s) + J(t))$$

Nous parlons également des longueurs de l'ensemble des propositions françaises et japonaises de la conjonction des deux classes, notées :

$$T_{lg}(i, j) = (F(i, j), J(i, j))$$

et calculées comme :

$$F(i, j) = F(i) + F(j)$$

$$J(i, j) = J(i) + J(j)$$

Rapport des longueurs

À l'aide de ce tableau des longueurs, le rapport des longueurs de l'élément s tel que $T(s) = (F(s), J(s))$ est calculé comme :

$$rap(s) = \frac{\max\left((F(s) \times R), J(s)\right)}{\min\left((F(s) \times R), J(s)\right)}$$

où R est le ratio entre les sommes des longueurs de toutes les propositions françaises et de celles japonaises à aligner obtenu par :

$$R = \frac{\sum_{i=1}^{n} lg(J_i)}{\sum_{j=1}^{m} lg(F_j)}$$

Notre hypothèse est que plus le rapport des longueurs entre les propositions françaises et les propositions japonaises constituant l'élément considéré est proche du rapport de base 1, plus la probabilité que les propositions de l'élément considéré constituent la perle de l'alignement est grande.

Le rapport des longueurs d'un regroupement de deux éléments, tel que $rap(s,t)$, revient exactement au calcul du rapport des longueurs de l'élément regroupé r constitué de s et de t, dont les longueurs sont calculées de la manière décrite précédemment.

À l'état initial, une des longueurs de tout élément étant à 0 – puisque chaque élément correspond précisément à une et une seule proposition d'une langue –, le rapport des longueurs est non pas à 0 mais à β pour tout élément. Cette valeur par défaut est définie dans le but de favoriser le premier regroupement des éléments. Dans notre réalisation, β est défini de manière empirique à 2. Là encore, il pourrait être intéressant de la réétudier dans des expériences futures.

Évolution du rapport des longueurs

L'évolution des rapports des longueurs résultant du regroupement des éléments i et j est calculée à partir des rapports des longueurs ainsi obtenus :

$$évolution(i,j) = rap(i,j) - \min(rap(i), rap(j)).$$

Si le regroupement considéré (en l'occurrence le regroupement des éléments i et j) est favorable, la valeur $M_{raplong}(i,j)$ sera négative, et dans le cas inverse, elle sera positive.

Considérons un regroupement des éléments a et b tels que $T_{lg}(a) = (8,1)$, $T_{lg}(b) = (2,4)$, avec $R = 0,5$. Le rapport des longueurs actuel de l'élément a est $\frac{\max(8\times0,5;1)}{\min(8\times0,5;1)} = 4$, et celui de l'élément b, $\frac{\max(2\times0,5;4)}{\min(2\times0,5;4)} = 4$. Le rapport de l'élément regroupant a et b est :

$$rap(a,b) = \frac{\max((8+2)\times0,5;(1+4))}{\min((8+2)\times0,5;(1+4))} = 1$$

L'évolution du rapport des longueurs suite au regroupement des éléments a et b est donc $1-4=-3$. Ce regroupement entraînant une évolution négative serait considéré comme favorable.

Considérons encore un autre regroupement des éléments c et d tels que $T_{lg}(c) = (4,4)$, $T_{lg}(d) = (2,5)$, avec $R = 0,5$. Le rapport des longueurs actuel de l'élément c est $\frac{\max(4\times0,5;4)}{\min(4\times0,5;4)} = 2$, et celui de l'élément d, $\frac{\max(2\times0,5;4)}{\min(2\times0,5;5)} = 5$. Le rapport

de l'élément regroupant a et b est :

$$rap(c,d) = \frac{\max((4+2)\times 0,5;(4+5))}{\min((4+2)\times 0,5;(4+5))} = 3$$

L'évolution du rapport des longueurs suite au regroupement des éléments c et d est donc $3-2=1$. Ce regroupement entraînant une évolution positive serait considéré comme défavorable.

L'évolution du rapport des longueurs ainsi obtenue ne permet cependant pas de discriminer les regroupements entraînant une perle d'un petit nombre de propositions par rapport à ceux qui produisent une grosse perle de propositions. Afin de favoriser la création d'un grand nombre de petites perles plutôt qu'une ou quelques grosses perles, ces valeurs indiquant l'évolution du rapport sont pondérées. Dans notre réalisation, le poids est défini comme le logarithme de la moyenne de la somme des longueurs normalisées des propositions françaises et celle des longueurs des propositions japonaises. Pour le calcul de l'évolution du rapport des longueurs résultant du regroupement des éléments i et j, le poid α est :

$$\alpha = \log\left(\frac{\big((F(i)+F(j))\times R\big)+(J(i)+J(j))}{2}\right)$$

13.4.5 Matrice courante

En combinant ces deux matrices, de similarité et de rapport des longueurs, une troisième matrice, matrice courante, est calculée et recalculée après chaque agrégation de deux éléments.

La matrice courante est définie comme :

$$M_{courante}(i,j) = \frac{M_{raplong}(i,j)}{M_{similarité}(i,j)}$$

Dans cette matrice courante, nous cherchons la valeur minimum pour réaliser l'agrégation de deux éléments, puis on calcule à nouveau les trois matrices et recommençons les opérations d'agrégation jusqu'à ce que toutes les propositions soient regroupées avec au moins une proposition de l'autre langue.

13.5 Évaluation des méthodes

Nous avons réalisé une évaluation des méthodes proposées avec quatre corpus parallèles[5] de natures diverses et de langues originaires différentes (1, 2 en français et 3, 4 en japonais) : (1) corpus LMD et LMDJP, constitués d'articles du Monde Diplomatique, (2) corpus BRVF et BRVFJP, composés de deux brevets techniques et (3) BRVJ et BRVJJP, composés d'un brevet technique, (4) corpus FdT et FdTJP, un extrait du roman « La fin des temps » de Haruki MURAKAMI. Comme

[5] Pour le contenu détaillé de chaque corpus, voir la Liste des corpus utilisés (page 547).

nous l'avons déjà présenté, le corpus est d'abord aligné au niveau des phrases par notre système d'alignement des phrases (cf. ch. 3) et le résultat est vérifié manuellement. Puis, pour chaque phrase, la détection des propositions est réalisée à l'aide de nos détecteurs de propositions du français (cf. ch. 9) et du japonais (cf. ch. 11) et le résultat d'analyse est également corrigé manuellement.

13.5.1 Description du corpus

	(A/B)	(C)	(D)	(E)	(F)
	Perles	Fr	Jp	Prop.	Prop./Perle
LMD	222/500	644	1026	583	2,626
BRVF	161/339	447	854	444	2,758
BRVJ	44/66	146	280	141	3,205
FdT	99/200	286	428	251	2,535

TAB. 13.14 – Description des corpus de l'évaluation

Nous avons utilisé au total 1105 paires (ou perles) de phrases alignées (détails pour chaque corpus indiqués dans (B) du tableau 13.14). Parmi celles-ci, nous n'avons pris en compte dans nos résultats d'évaluation que les paires comportant plus d'une proposition dans chaque langue, soit 526 paires de phrases (A), qui représentent 1523 propositions françaises (C) et 2588 propositions japonaises (D), composant 1419 paires de propositions en relation de traduction (E). Le nombre moyen de paires de propositions par perle s'étend de 2,5 à 3,2 (F), avec une moyenne de 2,781 pour l'ensemble des corpus.

Nous pouvons constater que le nombre de propositions japonaises est au moins 50% plus élevé que celui des propositions françaises. Cela implique que le modèle de traduction 1-1 (modèle pour la paire en relation de traduction constituée d'une unité dans une langue avec une unité de l'autre langue) est beaucoup moins courant que dans le cas de l'alignement des phrases. La figure 13.15 (voir page suivante) présente la répartition par modèle de traduction de chaque paire de propositions. En effet, les paires 1-1 représentent moins de 50% et le nombre d'alignements d'une proposition française avec de 2 à plus de 4 propositions japonaises s'élève à environ 40%. Ce type de paire complexe est une source de perturbation pour les méthodes d'alignement des phrases classiques.

Par ailleurs, le nombre moyen de paires de propositions par perle est particulièrement élevé pour le corpus BRVJ. Cela reflète bien le style des phrases des brevets qualifiées souvent de « très longues ».

13.5.2 Résultats

Dans le tableau 13.16, est présenté le résultat de notre évaluation des trois méthodes : méthode des graphes uniquement topologique (M1), méthode des

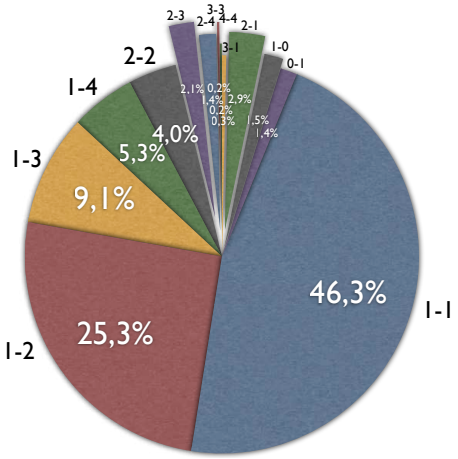


FIG. 13.15 – Répartition des modèles de traduction

graphes avec type de propositions (M2) et méthode avec classification ascendante hiérarchique (M3). La zone marquée « Partiel » (F) indique la proportion de paires partiellement correctes parmi l'ensemble des paires effectivement alignées, et la zone marquée « Exact » (G), celle des paires exactement alignées correctement. Enfin, la zone marquée « Paires créées » (H), correspond à la proportion du nombre de paires créées par rapport au nombre correct de paires.

	Partiel (F)			Exact (G)			Paires créées (H)		
	M1	M2	M3	M1	M2	M3	M1	M2	M3
LMD	0,643	0,784	0,951	0,127	0,200	0,591	0,813	0,746	0,918
BRVF	0,619	0,705	0,977	0,081	0,158	0,706	0,750	0,757	0,867
BRVJ	0,663	0,689	0,990	0,048	0,078	0,537	0,738	0,638	0,674
FdT	0,670	0,659	0,932	0,138	0,151	0,464	0,892	0,817	0,936

TAB. 13.16 – Résultats de l'alignement par les trois méthodes

Remarques générales

Le plus grand atout des algorithmes basés sur les méthodes spectrales (M1, M2) est la rapidité de calcul : l'alignement du corpus LMD (plus de 200 paires de phrases) est réalisé en moins d'une seconde contre 14 pour la méthode CAH

(M3)[6].

Mais, les résultats d'alignement des méthodes spectrales sont loin d'un niveau satisfaisant. Quoique les résultats montrent que nous avons réussi à améliorer la méthode de Kosinov (M1) avec l'introduction des types de propositions (M2), les chiffres obtenus ne sont pas encore satisfaisants. Beaucoup de phrases auraient nécessité plus d'informations et leur alignement n'a été amélioré qu'avec la méthode à classification ascendante hiérarchique (M3) basée sur la similarité lexicale.

Le résultat particulièrement médiocre pour le corpus BRVJ s'explique sans doute – du moins partiellement – par la diversité de ses modèles de traduction. En effet, il comporte beaucoup de modèles extrêmement complexes, à savoir des paires constituées d'une proposition française et de sept jusqu'à onze propositions japonaises. Ceci est dû à la différence de définition de la proposition entre le français et le japonais. Comme nous l'avons vu dans les études linguistiques, du fait de l'absence d'opposition sur la forme, nous ne pouvons pas faire de distinction entre emplois fini et infini des mots variables en japonais. De plus, tout complément étant susceptible d'être omis, le repérage de la proposition dans la phrase japonaise se base essentiellement sur la présence d'un prédicat. Les propositions japonaises ainsi définies ne correspondent pas toujours aux propositions françaises définies sur la base de l'opposition sujet-prédicat. Beaucoup ont comme éléments équivalents en français, des syntagmes participiaux. Dans le style particulier des brevets techniques, les syntagmes participiaux sont utilisés de manière beaucoup plus importante que dans d'autres types de textes, d'où cette différence considérable des nombres de propositions entre le français et le japonais, constituant les unités en relation de traduction.

En plus de cette diversité des modèles de traduction utilisés, comme nous l'avons déjà fait remarquer dans la description des corpus, ce corpus est également caractérisé par un nombre moyen élevé de propositions par perle. Ce qui n'est probablement pas un facteur facilitant l'opération d'alignement.

Nous allons maintenant examiner d'une manière plus détaillée quelques résultats de chaque méthode et les cas d'exemples dans lesquels l'alignement des propositions est fondamentalement difficile.

Méthodes spectrales (M1, M2)

Comme on peut le déduire d'après les résultats, les méthodes spectrales se sont montrées peu efficaces pour notre opération d'alignement. Une des principales causes de cet échec est, vraisemblablement, la taille réduite des graphes que nous traitons dans nos travaux : lorsque les graphes à apparier ne comportent que trois ou quatre nœuds, les informations topologiques nécessaires à leur appariement sont également restreintes.

[6]Les tests ont été effectués sur la configuration suivante : MacBookPro, 2,33 GHz, 2 Go de RAM, Mac OS X 10.4.

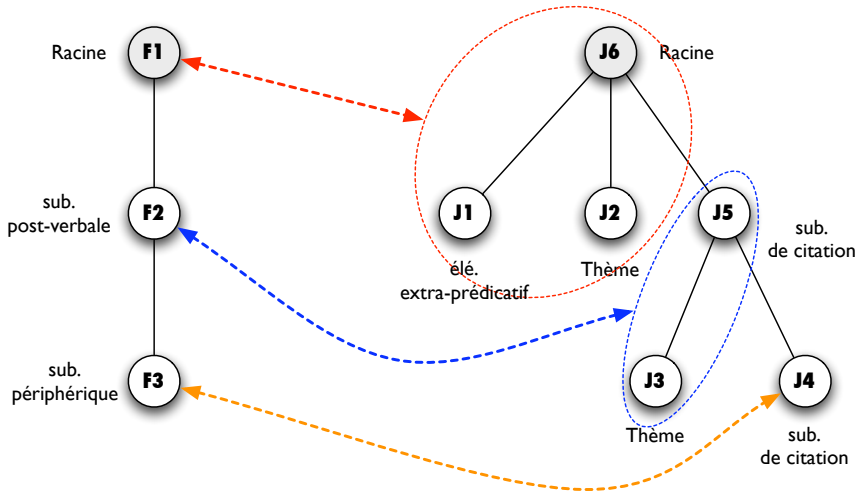


FIG. 13.17 – Arbres des propositions d'entrée et appariement correct de leurs nœuds

Le problème de la méthode de Kosinov (M1) est double dans notre application : elle ne tient compte que de la topologie et elle ne vise que l'appariement des graphes non-orientés. Cette deuxième propriété a finalement eu une influence cruciale sur l'appariement des arbres des propositions. Beaucoup d'arbres des propositions non symétriques sont interprétés comme des graphes non-orientés symétriques.

Considérons le cas des phrases parallèles suivantes (‖ indique les frontières des propositions) :

Phrase française :
($F1$) Paris avait estimé, à l'époque ‖ ($F2$) , qu'une référence aux valeurs religieuses n'était pas acceptable ‖ ($F3$) car elle soulevait des problèmes politiques et constitutionnels en France.

Phrase japonaise :
($J1$) 当時、(*tôji,* à l'époque) ‖ ($J2$)フランスは、 (*furansu wa,* La France) ‖ 宗教的価値への言及は (*shûkyôteki kachi eno genkyû wa,* une référence aux valeurs religieuses) ‖ ($J3$) 国内で政治上、憲法上の問題を引き起こすがゆえに (*kokunai de seijijô, kenpôjô no mondai wo hikiokosuga yueni,* car [elle] soulève des problèmes politiques et constitutionnels dans le pays) ‖ ($J4$) 認められないとの (*mitomerarenai tono,* qui dit que ce n'était pas acceptable) ‖ ($J5$)姿勢をとった。 (*shisei wo totta,* [La France] a pris la position)

La figure 13.17 montre les arbres des propositions de ces phrases et la figure 13.18 page suivante montre un résultat de la projection de ces arbres des pro-

positions. Les projections sont symétriques, alors que les arbres des propositions d'entrée ne le sont pas.

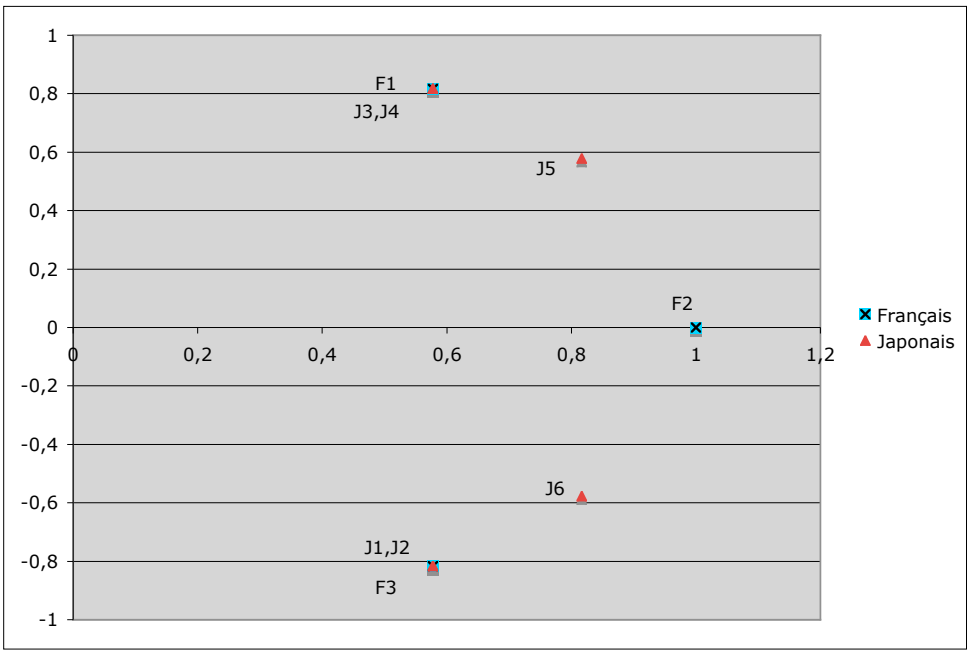


FIG. 13.18 – Résultat de la projection avec la méthode topologique (Kosinov)

Dans les cas comme cet exemple, l'introduction des informations sur le type de proposition (M2) a permis d'améliorer le résultat et de fournir l'alignement correcte.

La figure 13.19 (voir page suivante) montre le résultat de la projection avec la méthode (M2) prenant en compte les types de propositions pour les même arbres des propositions. On peut y constater le rapprochement des nœuds du même type tels que les racines des deux arbres, situées dans des positions éloignées dans le résultat de la méthode topologique (cf. figure 13.18).

Toutefois, il est difficile de trouver une formule permettant de refléter les informations supplémentaires tout en conservant les relations topologiques. De plus, beaucoup de phrases auraient nécessité encore plus d'informations et leur alignement n'a été amélioré qu'avec la méthode à classification ascendante hiérarchique (M3) basée sur la similarité lexicale.

Méthode basée sur la CAH (M3)

L'introduction des informations lexicales a permis d'aligner correctement des phrases pour lesquelles la topologie et les informations sur les types des propositions ne suffisaient pas. Les figures 13.20 page 445 (exemple I) et 13.21 page 446

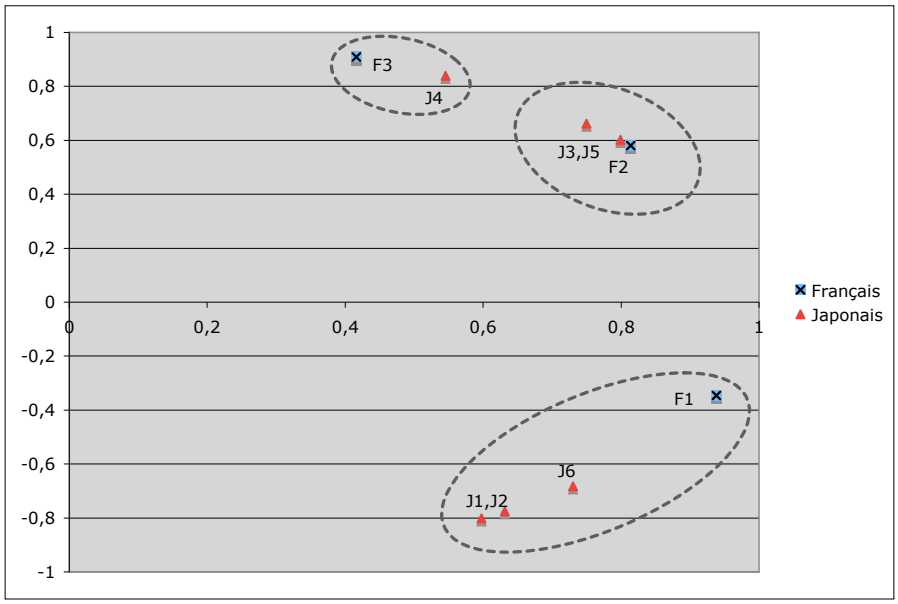


FIG. 13.19 – Résultat de la projection avec la méthode améliorée utilisant les distances des types de propositions

(exemple II) présentent deux exemples de phrases correctement alignées par la méthode CAH (M3), pour lesquels aucune des méthodes spectrales (M1, M2) n'a pu fournir de résultat correct.

Exemple I L'élément français de la perle est constitué de deux phrases, la première contenant une relative sous forme insérée et détachée par des tirets. L'élément japonais de la perle est composé de trois phrases contenant neuf propositions. Chaque phrase japonaise correspond à une des trois propositions françaises (2 propositions racine et relative de la première phrase + 1 proposition de la seconde phrase). La relative enchâssée dans la première phrase française est traduite dans le japonais par une phrase indépendante, qui apparaît après la traduction de la seconde phrase. Ce changement d'ordre est d'autant plus défavorable pour l'alignement basé sur la topologie qu'il a également provoqué une modification des relations syntaxiques, puisque nous considérons que dans le cas de la fusion de plusieurs phrases, les phrases sauf la première sont régies syntaxiquement par celle qui les précéde directement. De plus, les relations entre les phrases fusionnées étant définies comme une coordination, l'introduction des informations sur les types de proposition n'a pas permis d'améliorer le résultat d'alignement.

Exemple II L'élément français de la perle est constitué d'une phrase contenant, en plus de la proposition racine (F1), une complétive (F2) qui comporte elle-même

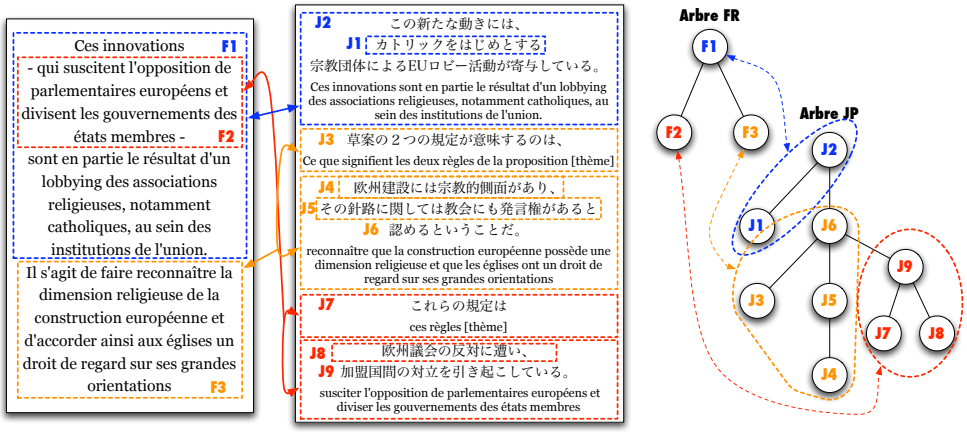


FIG. 13.20 – Exemple de phrases correctement alignées par la méthode M3 (I)

deux relatives (F3, F6) et deux circonstancielles coordonnées (F4, F5). L'élément japonais de la perle est également composé d'une phrase contenant, elle, huit propositions. En dépit de la différence de nombre de propositions, la structure est quasi-identique. La phrase japonaise constituée d'un thème (J1) et d'une racine (J8) et contient une subordonnée à connecteur agglutinant *koto* (J4), correspondant aux complétives, qui comporte elle-même deux relatives (J3, J7) – dont l'une (J3) possède une subordonnée (J2) – ainsi qu'une subordonnée à conjonction de condition (J6) précédée par une subordonnée à forme neutre (J5), considérée comme coordonnée. La difficulté de l'alignement basé sur les relations syntaxiques de dépendance pour cette paire de phrases est liée à la structure de coordination. Comme on peut le constater sur les arbres des propositions de la figure, les rôles des deux circonstancielles coordonnées sont inversées dans les phrases française et japonaise : dans la phrase française, c'est la proposition F4 comportant le connecteur qui entre en relation avec sa régissante alors que dans la phrase japonaise, sa traduction J5 à la forme neutre entre en relation avec la proposition J6 qui la suit, et c'est cette subordonnée J6 à conjonction de condition qui entre en relation avec leur proposition régissante. Cette inversion des relations syntaxiques due à la coordination est très fréquente du fait du choix des traducteurs qui préfèrent, vraisemblablement, conserver l'ordre d'apparition des éléments coordonnées. L'alignement correct de ces éléments inversés est impossible à l'aide uniquement des informations syntaxiques. Mais, avec la méthode basée sur les informations lexicales, l'inversion des relations syntaxique ne pose pas de problème particulier dès lors que les éléments inversés ont une similarité lexicale suffisante avec leur correspondant.

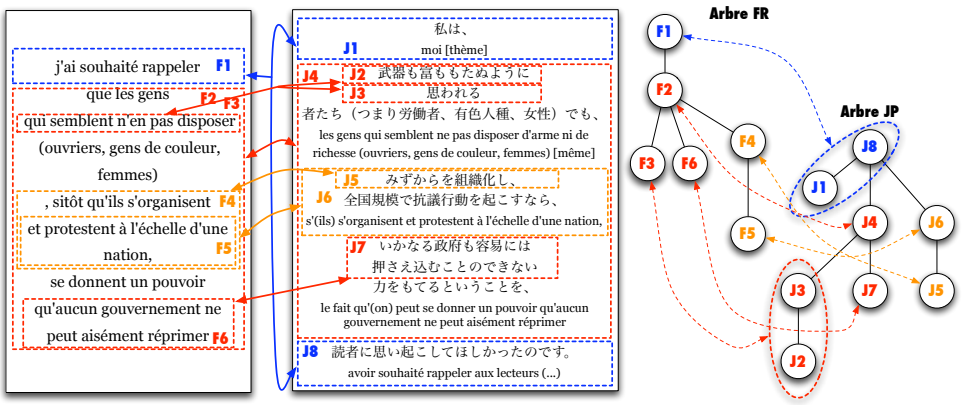


FIG. 13.21 – Exemple de phrases correctement alignées par la méthode M3 (II)

Problèmes liés au calcul de la similarité lexicale Le tableau 13.22 présente les caractéristiques lexicales des corpus (a, b, c) et les résultats de la recherche des mots en relation de traduction réalisée dans la procédure de la méthode M3 (d, e). La colonne (a) indique le nombre de perles (= (A) du tableau 13.14 page 439), les colonnes (b) et (c) présentent le nombre de mots lexicaux, respectivement français et japonais, avec leur moyenne par perle entre parenthèses. La colonne (d) correspond au nombre de paires de mots lexicaux trouvés dans la liste de mots alignés obtenue avec le système d'alignement des phrases AlALeR et la colonne (e) à celui obtenu avec la consultation du dictionnaire bilingue, avec les moyennes par perle entre parenthèses. La colonne (f) représente le nombre de paires de mots lexicaux obtenu suite au calcul de similarité des chaînes, parmi les valeurs de (e).

	(a)	(b)	(c)	(d)	(e)	(f)
	Perles	Mots FR (/perle)	Mots JP (/perle)	AlALeR (/perle)	Dico (/perle)	Dico' (/perle)
LMD	222	3024 (13,7)	3756 (16,9)	395 (1,8)	989 (4,5)	158 (0,7)
BRVF	161	3010 (18,7)	3198 (19,9)	1268 (7,9)	802 (5,0)	203 (1,3)
BRVJ	44	1151 (26,2)	1049 (23,8)	170 (3,9)	190 (4,3)	53 (1,2)
FdT	99	1082 (10,9)	1802 (10,9)	44 (0,4)	351 (3,5)	39 (0,4)

TAB. 13.22 – Description des corpus de l'évaluation (II) et résultats de la recherche des mots en relation de traduction

Nous avons abordé, dans la description de la méthode (cf. § 13.4.3), le problème lié à la mise en correspondance des mots avec dictionnaire, dû à la différence des unités japonaises traitées par le dictionnaire et par l'analyseur morphologique. Afin d'optimiser l'exploitation des données contenues dans le dictionnaire, nous avons introduit un calcul de similarité des chaînes au moment de la consultation du dictionnaire. Les résultats obtenus par cette amélioration (f) re-

présentent 453 paires sur 2332, soit 20% du résultat total. Ce résultat est encourageant mais non satisfaisant, car globalement, l'utilisation du dictionnaire bilingue n'a permis qu'un alignement d'à peine un tiers, voire moins, des mots lexicaux.

Remarques générales sur la méthode Cette méthode basée sur la CAH possède également encore d'autres points potentiels d'amélioration (comme la désambiguïsation lexicale par exemple), mais la capacité d'alignement avec des croisements est un atout crucial. De plus, comme nous le savons bien, la méthode de classification nous permet de définir nous-même la fin du développement des fusions. Par ce mécanisme, nous pourrions obtenir un résultat moins robuste mais plus fiable.

Cas de l'alignement automatique fondamentalement difficile

Dans les résultats des trois méthodes, l'échec provient également de la différence considérable entre le français et le japonais. Cette différence réside sur différents plans – lexical, syntaxique ou encore rhétorique – si bien que nous avons rencontré des constructions très différentes de diverses natures. Ces exemples, pour lesquels un appariement même manuel est souvent très difficile, sont constatés plus particulièrement dans le corpus littéraire FdT.

Différences sur le plan lexical Considérons les phrases parallèles suivantes :

> **Phrase française :**
> [F1$_{racine}$ À tel point que je ne savais plus] [F2$_{subQ}$ s'il progressait ou non]
> **Phrase japonaise :**
> [J1$_{subAgg}$ 停まっているのか(*tomatteiru no ka*, s'il est arrêté) [J2$_{subAgg}$ 動いているのかも(*ugoiteiru no ka mo*, s'il est en train de bouger)] [J3$_{racine}$ わからないくらいだった。(*wakaranai kurai datta*, À tel point que je ne savais pas ...)]

Dans cet exemple, la séquence japonaise « *tomatteiru no ka, ugoiteiru no ka mo* (s'il est arrêté, s'il est en train de bouger) » est traduite en français par « s'il progressait ou non » avec un seul verbe « progresser » alors que dans la phrase japonaise, deux verbes « *ugoku* (bouger) » et « *tomaru* (s'arrêter) » sont mis en contraste et que chacun constitue une proposition indépendante.

Différences sur le plan syntaxique

> **Phrase française :**
> [F1$_{racine}$ c'est notamment lors des débats sur les programmes d'aide aux pays du sud] [F2$_{postN}$ que les questions de la contraception et du statut de la famille sont abordées]
> **Phrase japonaise :**
> [J1$_{theme}$ 避妊と家族の地位という問題は、(*hinin to kazoku no chii*

toiu mondai wa, les questions de la contraception et du statut de la famille [thème])] [J2$_{racine}$ 特に開発途上国援助プログラムをめぐる議論の中で大きく取り上げられた。(*tokuni kaihatsu tojô koku enjo puroguramu wo meguru giron no nakade ôkiku toriagerareta*, être abordé, notamment lors des débats sur les programmes d'aide aux pays en voie de développement)]

La phrase japonaise est constituée du thème (J1) et de la racine (J2). Le thème correspond au sujet de F2 « les questions de la contraception et du statut de la famille », et la racine comporte l'élément mis en relief « notamment lors des débats sur les programmes d'aide aux pays du sud » et le prédicat de F2 « sont abordées ». En effet, la mise en focus d'un syntagme du type « c'est ... que » peut être réalisée en japonais par la simple utilisation de particules dites casuelles. Il faut donc fusionner les deux propositions dans les deux langues pour établir la correspondance. Ce genre d'alignement, impossible à réaliser automatiquement avec l'utilisation des seuls relations de dépendance et types de propositions, reste très complexe même avec l'utilisation des informations lexicales.

Différences sur le plan rhétorique L'alignement automatique correct des unités concernées par ce type de différence semble particulièrement difficile.

Phrase française :
[F1$_{racine}$ Je n'arrivais pas à croire] [F2$_{subQ}$ que c'était moi] [F3$_{subQ}$ qui avais émis ce bruit]
Phrase japonaise :
[J1$_{racine}$ 私には (*watashi ni wa*, À moi) [J2$_{subCit}$ それが(*sore ga*, ceci [*ga*]) [J3$_{subAgg}$ 自分の体から発せられた(*jibun no karada kara hasserareta*, émis de mon corps)] 音だとは(*oto da to wa*, être un son / bruit)] どうしても思えなかった。(*dôshitemo omoenakatta*, je n'arrivais pas croire ...)]

Tandis que dans la phrase française, l'actant « moi » est mis en relief, dans la phrase japonaise, aucune construction syntaxique de mise en relief n'est utilisée. La traduction littérale est : « je n'arrivais pas à croire que c'était un bruit émis de mon corps ». La mise en correspondance de chaque proposition française avec des propositions japonaises est dans ce cas impossible et l'alignement nécessite la fusion des deux propositions F2 et F3, entraînant un modèle complexe 2-3, à savoir la paire constituée des deux propositions françaises F2 et F3 et des trois propositions japonaises J2, J3, J4.

L'exemple suivant est un cas encore plus difficile.

Phrase française :
[F1$_{ED}$ En y réfléchissant,] [F2$_{ED}$ les trucages,] [F3$_{racine}$ je n'étais pas près de les découvrir :] [F4$_{ED}$ déjà,] [F5$_{propcrd}$ je ne savais pas] [F6$_{subQ}$ si l'ascenseur marchait ou non]
Phrase japonaise :
[J1$_{subCond}$ 考えてみれば (*kangaete mireba*, si (je) réfléchis)] [J2$_{racine}$ たねどころか私には(*tane dokoroka watashi ni wa*, sans aller jusqu'aux trucages,

à moi) [J3$_{subAgg}$ エレベーターが動いているのか(*erebêtâ ga ugoiteiru no ka*, si l'ascenseur est en train de bouger)] [J4$_{subAgg}$ 停まっているのかさえ(*tomatteiru no ka sae*, s'il est arrêté)] わからないのだ。(*wakaranai kurai datta*, je ne sais pas ...)]

La traduction littérale française de l'original est : « si je réfléchis bien, je ne sais, sans aller jusqu'aux trucages, même pas si l'ascenseur est en train de bouger ou s'il est arrêté ». Le syntagme nominal japonais « *tane dokoroka* (sans aller jusqu'aux trucages) » a été traduit par une proposition quasi-autonome avec un thème. Le système n'a pas, bien entendu, réussi à mettre correctement en relation ces propositions. À la première lecture, nous avons mis, nous-mêmes, du temps pour établir toutes les correspondances.

13.6 Conclusion

Nous avons présenté deux approches pour l'alignement des propositions des textes parallèles français-japonais. L'une s'appuie sur des méthodes d'appariement de graphes consistant à projeter les nœuds sur un sous-espace propre. L'autre est inspirée de la classification ascendante hiérarchique. Les deux approches sont caractérisées par leur capacité d'alignement des traductions croisées quant à l'ordre d'apparition, ce qui était impossible pour beaucoup de méthodes classiques d'alignement des phrases.

Les résultat obtenus avec les méthodes spectrales n'étaient pas satisfaisants. Il est en effet difficile de trouver une formule permettant de refléter les informations supplémentaires autres que la topologie. Une très récente étude de Fraikin et al. (2006) propose une amélioration visant le traitement des graphes orientés. Néanmoins, du fait des différences considérables de structures, l'application de cette méthode à l'alignement de langues très différentes semble difficile.

En revanche, l'alignement basé sur la méthode de classification ascendante hiérarchique est prometteur dans la mesure où cette technique permet d'exploiter plus efficacement différentes informations sans être perturbée par les croisements de traductions. Malgré cet intérêt, les résultas obtenus avec la méthode inspirée de la CAH ne sont pas encore tout à fait satisfaisants. La principale cause d'échec provient des mauvais résultats de la mise en correspondance des mots. Seulement 20 à 30% des mots sont mis en relation par la consultation du dictionnaire. Une des raisons de cette difficulté est la différence de catégorie entre les mots correspondants en français et en japonais. Il faut donc soit déterminer l'unité la plus adéquate afin de réorganiser spécifiquement un dictionnaire bilingue pour mieux l'adapter à notre tâche, soit utiliser une méthode de mise en correspondance des mots complètement différente telle que celles utilisées dans la traduction automatique statistique.

À travers cette expérience, nous avons également rencontré beaucoup de constructions pour lesquelles un appariement même manuel était très difficile. Ces exemples sont, pour nous, non seulement des indicateurs de futurs obstacles

à franchir, mais aussi très enrichissants du point de vue de l'étude contrastive sur les structures syntaxiques des phrases française et japonaise.

CONCLUSION

これは終わりではない。
これは終わりの始まりですらない。
しかし、あるいは、始まりの終わりかも知れない。

Ce n'est pas la fin,
ni même le commencement de la fin ;
mais c'est peut-être la fin du commencement.

This is not the end.
It is not even the beginning of the end.
But it is, perhaps, the end of the beginning.
— Winston Churchill

Nous avons présenté l'ensemble de nos travaux sur l'alignement de textes parallèles français-japonais comportant aussi bien des réalisations informatiques – une série de systèmes réalisant ensemble au final l'alignement des propositions de textes parallèles français-japonais – que des études linguistiques, qui ont constitué la base de ces développements.

La détermination – ou même la définition – de l'unité « proposition » étant en réalité une tâche beaucoup plus complexe qu'il peut sembler au premier abord, le développement des systèmes a nécessité un fort investissement dans les études linguistiques afin de fournir un fondement solide à la réalisation. Ce besoin nous a finalement fourni l'occasion de pousser nos réflexions sur un large domaine couvrant les sujets connexes.

Toutefois, dans les travaux sur la linguistique japonaise, nos études comportent, comme nous l'avons déjà dit dans la conclusion de la partie concernée, encore un grand nombre de questions de détail en suspens, du fait de l'absence partielle (voire totale) d'études linguistiques antérieures.

Il y a plus de cinquante ans, Mikami a publié un livre (Mikami, 1953) qu'il a qualifié de « premier pas dans le domaine de la syntaxe ». Il expose dans ce livre différentes questions syntaxiques et écrit dans la postface :

> « Si d'importants efforts ne sont pas consacrés aux questions similaires à celles que j'ai présentées vers la fin du livre – questions que j'ai beaucoup de mal à résoudre –, nous ne pourrons jamais voir la constitution de la grammaire japonaise. »

Ses questions sont encore loin d'être toutes résolues et nous étudions toujours à l'école la grammaire basée sur la théorie de Hashimoto.

Tout traitement automatique des langues nécessite – ou doit nécessiter – des fondements théoriques solides construits grâce aux recherches linguistiques. Il est donc important et indispensable de favoriser les progrès de la linguistique japonaise, en particulier dans le domaine de la syntaxe.

Les besoins spécifiques du TAL peuvent non seulement s'appuyer sur les fruits des recherches linguistiques mais aussi leur apporter de nouveaux regards, permettant des découvertes inattendues.

Nous espérons que nos travaux linguistiques réalisés avec une approche TAL contribueront aux progrès des recherches syntaxiques et même de la linguistique en général.

Avant de clôturer la discussion, nous allons présenter les perspectives des présents travaux.

* * *

Nos travaux possèdent deux possibilités de développement de nature différente : l'une concerne la mise à profit des données résultantes, alignées au niveau des propositions ; l'autre concerne l'amélioration de l'opération d'alignement elle-même.

Mise à profit des corpus alignés au niveau de la proposition

Nous abordons ici deux types de travaux possibles : l'enrichissement des corpus parallèles alignés et l'exploitation des données alignées, en particulier en vue de recherches en linguistique contrastive.

Enrichissement des corpus parallèles alignés

Du fait de la présence d'éléments dépendant du cotexte (e.g. anaphores, éléments elliptiques), les paires de propositions alignées n'ont pas d'équivalence lorsqu'elles sont traitées isolées. Cette non-équivalence est plus forte dans le cas des propositions que des phrases, du fait, par exemple, du partage fréquent des mêmes compléments par plusieurs propositions de la même phrase (cf. § 4.8.4 et § 7.13).

Pour compenser ce défaut des données alignées au niveau des propositions, quelques opérations supplémentaires sont souvent nécessaires et utiles, comme par exemple, la résolution des anaphores ou la restitution des éléments elliptiques.

Dans les textes japonais, deux opérations supplémentaires sont envisageables : la résolution des fonctions cumulatives du syntagme thématisé et la restitution des compléments omis.

Résolution des fonctions cumulatives du syntagme thématisé

Comme nous l'avons déjà vu dans nos études linguistiques, le syntagme thématisé assure souvent une ou même plusieurs fonctions syntaxiques vis-à-vis des prédicats constituant le rhème.

Considérons les phrases suivantes, traduction l'une de l'autre.

Exemple 1 BRVJ[7]

> L'invention concerne les dispositifs dont le système est constitué de plusieurs modules tels que les séquenceurs, et concerne en particulier une amélioration permettant d'assurer de façon certaine la mise à la terre à l'intérieur des modules.
>
> 本発明はシーケンサ等の多数のモジュールによりシステムが構成される装置に係り、特にモジュール内部での接地を確実に行う改良に関する。

La phrase japonaise est d'abord séparée en deux parties : d'un côté,

> 本発明は
> (*hon hatsumei - wa*)
> (présente invention - [*wa*])

et de l'autre,

> シーケンサ等の多数のモジュールによりシステムが構成される装置に係り、特にモジュール内部での接地を確実に行う改良に関する。

Le syntagme isolé introduit par la particule « は » (*wa*) dans la phrase japonaise est le thème. Le thème cumule souvent, en plus de son rôle thématique, la fonction intra-prédicative des propositions qui constituent la partie rhématique, et dans le cas de la phrase d'exemple, le syntagme thématisé occupe le cas *ga* pour les deux propositions qui constituent le rhème (cf. figure page suivante).

Cette configuration où le syntagme thématisé assure le cas *ga* est très fréquente, et dans ce cas, la distinction en deux couches (la couche thème/rhème et celle qui s'organise autour d'un prédicat) peut paraître redondante – ce qui empêche d'ailleurs la prise de conscience de l'existence même de cette différence. Dans le cas de l'exemple, on pourrait dire qu'il suffirait d'inclure le thème dans la première proposition pour obtenir l'équivalence.

Afin de montrer que le problème n'est pas toujours aussi simple, considérons deux autres exemples.

Exemple 2 LMD

> La fondation Rhin-Danube, la fondation Limat et ICU coopèrent à l'échelon international, notamment aux Philippines, où elles ont créé

[7] Pour le contenu détaillé de chaque corpus utilisé ici, voir la Liste des corpus utilisés (page 547).

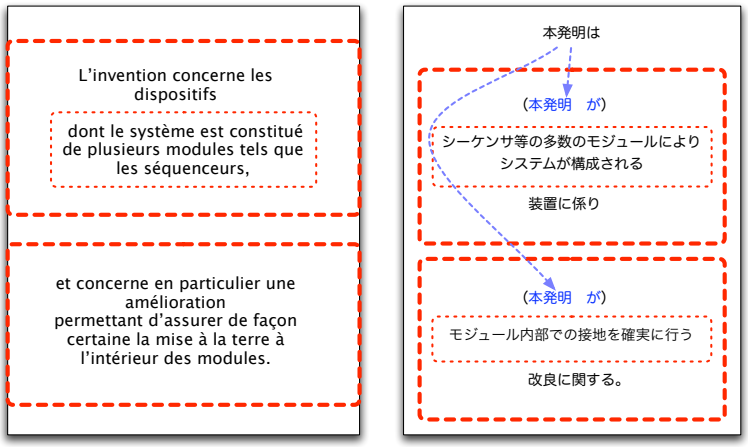


FIG. – Fonction cumulative du syntagme thématisé dans la phrase japonaise

en 1995 l'Université d'Asie et du Pacifique (University of Asia and the Pacific / UA & P).

ライン・ドナウ財団、リマト財団とICUは国際的規模で、とりわけフィリピンで協力関係を築き、1995年にはアジア太平洋大学（UA&P）を開校している。

La structure de la phrase japonaise est identique à celle de l'exemple précédent (cf. figure ci-dessous). Mais, la structure de la phrase française étant diffé-

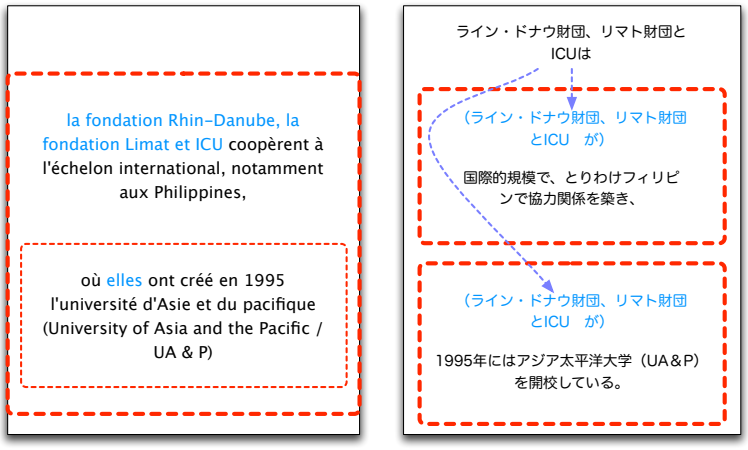


FIG. – Fonction cumulative du syntagme thématisé dans la phrase japonaise II

rente – elle comporte une relative et pas de coordonnée –, il est déjà moins facile de défendre l'équivalence entre la relative française et la seconde proposition ja-

ponaise sans restituer le complément en *ga* – joué par le syntagme thématisé – dont l'élément correspondant français, le sujet, est cette fois bien présent, bien que sous la forme d'un pronom.

Le dernier exemple montre le cas d'une phrase française comportant un thème (cf. figure page suivante).

Exemple 3 FdT

> Les pièces de un et de cinq yen, je les mets dans ma poche revolver, mais en principe je ne m'en sers pas dans les calculs.
>
> 一円玉と五円玉はヒップ・ポケットに入れるが原則として計算には使わない。

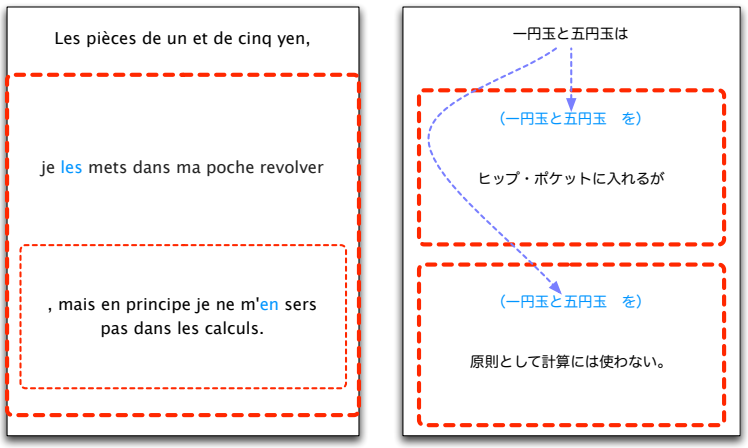


FIG. – Fonction cumulative du syntagme thématisé dans la phrase japonaise III

Le thème japonais a trois éléments correspondants dans la phrase française : le thème en prolepse « Les pièces de un et de cinq yen », le pronom clitique objet « les » dans la première proposition (racine), enfin le pronom clitique objet « en » dans la seconde proposition coordonnée. Sans restituer aucun complément joué par le syntagme thématisé dans le cadre de ses fonctions cumulatives, il est difficile de parler d'équivalence des propositions alignées.

La détermination de la fonction cumulative du thème est une opération capitale, non seulement pour l'alignement mais pour tout traitement automatique du japonais. L'automatisation de cette tâche est, malgré son importance, un sujet peu (voire pas du tout) étudié. La résolution de la fonction cumulative du thème pourrait, pourtant, probablement être réalisée avec des éléments linguistiques que nous sommes déjà capables de manipuler.

Nous considérons que l'absence d'introduction de cette distinction était, et est encore aujourd'hui, une des sources principales d'échec dans les travaux de TAL

en japonais. Nous considérons donc que non seulement les études sur l'automatisation de la détermination de la fonction cumulative du thème sont intéressantes, mais aussi que l'ajout d'informations liées aux fonctions cumulatives du thème dans le corpus pourrait constituer des données intéressantes pour travailler sur ce sujet trop peu étudié.

Restitution des compléments omis dans la phrase japonaise

Dans la phrase japonaise, l'omission d'éléments – aussi bien le syntagme en *ga* que d'autres compléments – est extrêmement fréquente. Les compléments implicites dus à la fonction cumulative du thème que nous venons de voir n'en constituent qu'un type. En effet, le japonais est une langue dépendant fortement non seulement du cotexte mais aussi du contexte extra-linguistique.

Cette caractéristique a une grande influence sur la traduction bien entendu, mais aussi sur l'apprentissage des langues étrangères par les Japonais, et est étudiée dans le domaine de l'enseignement de la langue. Takagaki présente cette particularité par une comparaison avec le français dans l'article (Takagaki, 2001) consacré aux problèmes de structuration des textes français pour les apprenants japonais. D'après l'auteur, pour s'exprimer, les locuteurs français ont tendance à constituer un monde fermé à l'espace intra-linguistique, tandis que les productions des locuteurs japonais dépendent plus fortement du monde extra-linguistique. Elle reconnaît ensuite dans cette différence la cause du problème des apprenants japonais qui produisent des phrases trop dépendantes de l'interprétation que peut en faire l'interlocuteur, pour être reconnues comme des phrases françaises « complètes ».

Cette particularité de la phrase japonaise qui est assez incomplète intrinsèquement, a bien évidemment des influences sur nos travaux, en particulier sur la qualité des données que nous fournissons en sortie de nos systèmes. Une paire de propositions, bien qu'alignées, risque de contenir moins d'informations dans sa partie japonaise que dans sa partie française. Il serait donc intéressant de compléter les éléments omis des phrases japonaises – ne serait-ce que des éléments déductibles par le cotexte – pour rendre encore plus utiles nos résultats d'alignement.

Exploitation des données : contribution aux travaux de linguistique contrastive

Afin de mieux mettre en valeur ces données alignées et surtout de faciliter le travail des linguistes, il est indispensable de concevoir un outil spécifique à cet emploi de nos données. La figure page 457 représente un exemple d'interface de ce type d'outil, un concordancier bilingue. La fenêtre gauche est la fenêtre principale interactive où l'utilisateur saisit un mot (« quand » dans la figure). Le résultat de la requête est affiché à l'intérieur de cette fenêtre principale : les propositions contenant le mot demandé sont alignées de sorte que les occurrences du mot ap-

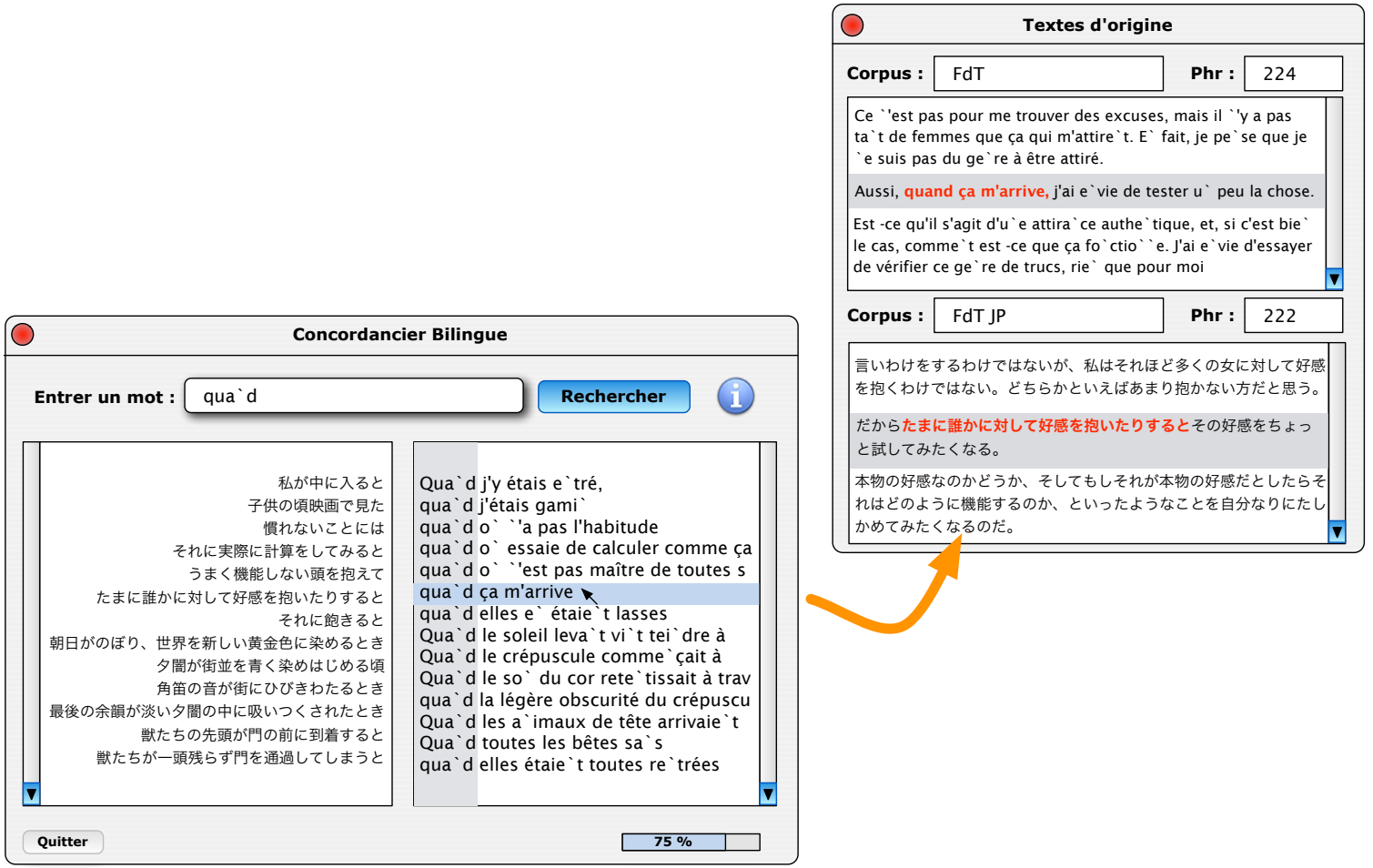


FIG. – Exploitation des données alignées par un concordancier bilingue

paraissent sur fond coloré ; les propositions japonaises alignées avec ces propositions françaises sont affichées dans la fenêtre de gauche sur la même ligne que leur correspondant français. Si l'utilisateur s'intéresse à un exemple particulier, la sélection de cet exemple entraîne l'ouverture d'une autre fenêtre (à droite sur la figure) sur laquelle l'exemple est affiché avec son cotexte large extrait du texte source, et ce, pour les deux langues étudiées.

La réalisation de ce type d'outil favoriserait les recherches en linguistique contrastive, prérequis indispensable aux progrès du multilinguisme dans le domaine du TAL.

Amélioration de l'alignement : de la proposition aux unités sous-phrastiques diverses

Plusieurs possibilités existent pour améliorer l'alignement des propositions. Nous abordons ici non pas l'aspect algorithmique – que nous avons déjà traité dans le dernier chapitre consacré à la réalisation du système – mais l'aspect linguistique, et plus précisément la possibilité d'amélioration par un ré-examen de la définition des unités à aligner.

Problème de la proposition comme unité de l'alignement

En effet, comme nous l'avons déjà vu dans l'analyse des résultats de notre système d'alignement (§ 13.5), les définitions de la proposition que nous avons adoptées pour le français et pour le japonais ont impliqué, notamment dans les textes de brevets techniques, une différence considérable des nombres de propositions constituant les unités correspondantes entre le japonais et le français.

Afin d'illustrer nos propos, considérons ces deux phrases française et japonaise, qui sont mutuellement traductions l'une de l'autre.

Exemple 4 LMD0704

> C'est une bouée de sauvetage à laquelle se raccrochent les gouvernements fervents partisans et pourvoyeurs de l'agriculture intensive, les chefs d'entreprises multinationales gaspillant les ressources, déversant sans vergogne leurs déchets et affrétant des bateaux-poubelles, les organisations non gouvernementales ne sachant plus que faire et les économistes pris en flagrant délit d'ignorance des contraintes naturelles.
>
> それは、集約農業を熱心に信奉し、整備する諸国政府や、資源を浪費し、恥ずかしげもなく廃棄物を投棄し、ぼろぼろの老朽船をチャーターする多国籍企業の経営者たち、もはや何をすればよいのか分からないNGO、そして自然環境の課す制約に無知なことを暴かれた経済学者たちがしがみついている救命ブイである。

Si on réalise la détection des propositions de ces deux phrases, nous obtenons seulement deux propositions dans la phrase française mais douze (dont un thème) dans la phrase japonaise.

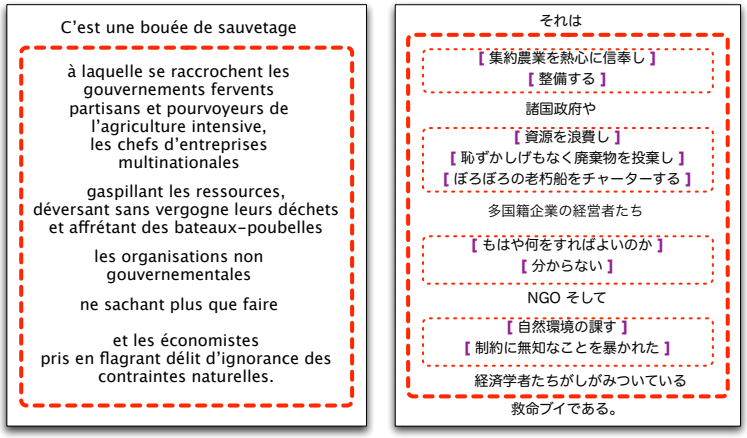


FIG. – Exemple de propositions alignées I

Si maintenant nous décidons de considérer comme des propositions tous les syntagmes ayant un verbe, même à la forme participiale, le nombre d'unités reconnues devient plus proche (bien qu'encore assez différent) à savoir sept dans la phrase française et toujours douze dans la phrase japonaise.

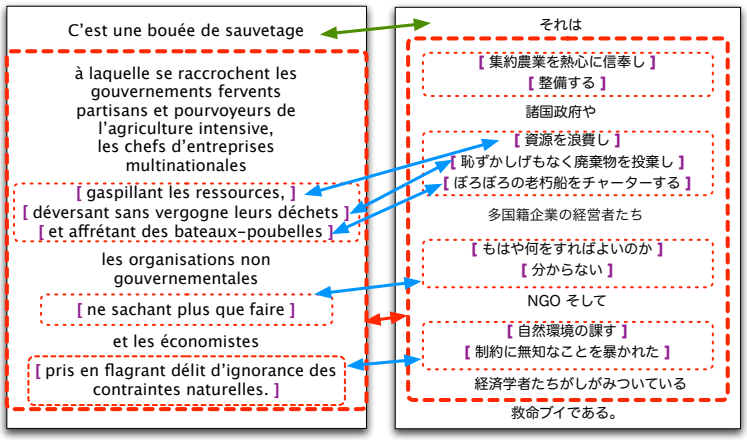


FIG. – Exemple de propositions alignées II

Nous avons donc deux choix pour tenter de rapprocher les unités à aligner. Le premier est de garder la définition actuelle de la proposition française et de déterminer les types – en supposant qu'ils existent – de propositions constituant des

unités correspondant mieux aux propositions françaises. Le second est d'extraire de la phrase française non pas des propositions, mais des syntagmes verbaux qui semblent avoir une nature plus proche de la proposition japonaise.

Cette dernière solution nous semble plus utile du point de vue de la constitution de bases de données, mais l'identification de ce type d'unité entraîne des problèmes très délicats liés à la définition même de l'unité. En effet, nous ne savons pas exactement tracer de frontière entre les adjectifs et les formes participiales des verbes.

Cette non-univocité des propositions française et japonaise telles que nous les avons définies, conjuguée avec la difficulté même de leur définition et leur détection automatique, nous oriente finalement vers un autre piste pour nos recherches sur l'alignement : l'alignement d'autres unités sous-phrastiques.

Alignement d'unités sous-phrastiques

De nos jours, les recherches sur l'alignement d'unités sous-phrastiques sont abondantes, celles-ci donnant de meilleurs résultats dans la traduction automatique que les mots alignés originellement employés dans ce domaine. Dans la section 12.4, nous avons déjà abordé quelques travaux de ce type qui traitent le japonais. La particularité de ces nouveaux travaux est que les unités à aligner ne sont pas préalablement fixées et que l'on cherche, plutôt qu'à aligner certaines unités préalablement déterminées, à mettre en relation différentes structures de tout niveau afin d'obtenir un maximum de patrons parallèles.

Contrairement aux travaux réalisés pour le traitement du japonais que nous avons présentés et qui utilisent les résultats de l'analyse syntaxique, certains travaux tels que Simard et al. (2005) ou Chiang (2005) se passent de cette opération préparatoire qui génère souvent beaucoup de bruit et se basent essentiellement sur l'alignement de mots de type « plusieurs-à-plusieurs ». La thèse de Chenon (2005) propose également un alignement hiérarchique de ce type sans analyse syntaxique. Le point qui différencie ses travaux des autres est que l'auteur ne s'appuie pas sur un alignement de type plusieurs-à-plusieurs pour repérer les unités à aligner, mais qu'il construit d'abord un arbre binaire de phrase sur la base des indices de « sécabilité » des séparateurs. Ce concept de sécabilité provient du constat que « certains mots sont plus soudés entre eux que certains autres ». L'alignement est ensuite réalisé entre les éléments de tout niveau des arbres construits pour chaque langue.

Le principal avantage de ce type d'alignement est qu'il ne nécessite aucun traitement préparatoire constituant lui-même un sujet de recherche très complexe, tel que l'analyse syntaxique.

Nous pourrions donc envisager également l'alignement non pas d'une unité spécifique telle que la proposition mais de diverses structures de tout niveau, tout en profitant des connaissances accumulées tout au long de la réalisation de la présente thèse.

Annexes

ANNEXE : ALALER

A.1 Algorithme de segmentation à l'aide de *trie*

Algorithme 2 Segmentation des séquences constituées entièrement de *kanji* à l'aide de *trie*

♢ Données :

1. entrées :
   - `lmotsg` : liste de mots graphiques (éventuellement à segmenter) ;
2. sorties :
   - `lemmes` : liste de lemmes ;
3. données locales :
   - `arbpre` : arbre (*trie*) de vérification des sous-chaînes préfixales ; sur chaque nœud est inscrit le nombre de mots y passant ;
   - `arbsuf` : arbre (*trie*) de vérification des sous-chaînes suffixales ; sur chaque nœud est inscrit le nombre de mots y passant ;
   - `mcourant` : mot en cours de traitement constitué de n caractères ;
   - `tpre` : tableau de longueur n, chaque case i contenant le nombre de mots ayant la même sous-chaîne préfixale que la sous-chaîne de 0 à i de `mcourant` ;
   - `tsuf` : tableau de longueur n, chaque case i contenant le nombre de mots ayant la même sous-chaîne suffixale que la sous-chaîne de i à $n-1$ de `mcourant` ;
   - `tprobpre` : tableau de longueur n, chaque case i contenant la probabilité que la frontière se trouve entre le $i^{\text{ème}}$ caractère et le $i+1^{\text{ème}}$ caractère, calculée à partir de `tpre` ;

– `tprobsuf` : tableau de longueur n, chaque case i contenant la probabilité que la frontière se trouve entre le $i^{\text{ème}}$ caractère et le $i+1^{\text{ème}}$ caractère, calculée à partir de `tsuf` ;

♢ Procédure :

1. Construction de `arbpre` à partir de `lmotsg`.
2. Construction de `arbsuf` à partir de `lmotsg`.
3. Pour tous les mots de `lmotsg` de longueur n, réaliser les opérations suivantes :
   a) remplir le tableau `tpre` à l'aide de `arbpre` ;
   b) remplir le tableau `tsuf` à l'aide de `arbsuf` ;
   c) (Recherche de frontières)
   Si $n \leq 2$, alors `mcourant` est ajouté tel quel dans `lemmes` et le traitement est terminé,
   Sinon :
      i. Calcul des probabilités de frontières :
      `tprobpre`[i] = `tpre`[i] - `tpre`[$i+1$]
      `tprobsuf`[i] = `tsuf`[i] - `tpre`[$i-1$]
      ii. Si le dernier caractère est un morphème grammatical (`tprobsuf`[$n-1$] > 10), alors il est supprimé et $n := n-1$;
      iii. Examen des deux premiers caractères : si `tprobpre`[1] > 0, alors la sous-chaîne préfixale constituée des deux premiers caractères est un lemme et est enregistrée dans `lemmes` ;
      iv. Si $n > 4$, alors examen des deux derniers caractères : si `tprobsuf`[$n-2$] > 0, alors la sous-chaîne suffixale constituée des deux derniers caractères est un lemme et est enregistrée dans `lemmes` ;
      v. S'il reste une sous-chaîne intermédiaire, alors toute la partie restante est considérée comme un lemme et est enregistrée dans `lemmes` ;
4. (Comparaison de lemmes)
   Pour tous les mots de `lemmes` de longueur n, réaliser les opérations suivantes :
   a) **Si** $n \leq 2$, alors le traitement est terminé,
   Sinon :
      $i := 1$;
      i. Si $i \geq n-1$, alors le traitement est terminé ;
      ii. Si la sous-chaîne préfixale de `motcourant` de 0 à i est semblable à un des lemmes enregistrés dans `lemmes`, alors elle est supprimée et $n := n-(i+1)$, $i := 0$;

iii. $i := i + 1$;

iv. Retourner en 4(a)i ;

b) Si un ou plusieurs lemmes sont reconnus dans `motcourant`, et que la partie restante a une longueur supérieure ou égale à 2, elle est considérée comme un nouveau lemme et est stockée à la fin de la liste `lemmes`.

A.2 Grammaire de retranscription des *katakana*

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%
% Règles de retranscription
% des katakana en alphabet
%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%
% Dans le cas où le choix de l'état suivant n'est pas unique :
%    Priorité : ant > post
%    Ex.)
%       Si (A, x) --> B | C | D,
%       alors, on vérifie d'abord la possibilité pour B
%
% X = l'ensemble d'états
%
%


t(X, ア ; a)                          t(X, ク ; ku)
s(a ; a)                              s(ku ; ku, cu)

t(X, イ ; i)                          t(X, ケ ; ke)
s(i ; i, y)                           s(ke ; ke)

t(X, ウ ; u, w)                       t(X, コ ; ko)
s(u ; u, w)                           s(ko ; ko, co)
s(w ; w)
                                      t(X, ガ ; ga)
t(X, エ ; e)                          s(ga ; ga)
s(e ; e)
                                      t(X, ギ ; gi, g)
t(X, オ ; o)                          s(gi ; gi)
s(o ; o)                              s(g ; g)

t(X, カ ; ka)                         t(X, グ ; gu)
s(ka ; ka, ca)                        s(gu ; gu)

t(X, キ ; ki, k)                      t(X, ゲ ; ge)
s(ki ; ki)                            s(ge ; ge)
s(k ; k)
                                      t(X, ゴ ; go)
```

```
s(go ; go)

t(X, サ ; sa)
s(sa ; sa)

t(X, シ ; si, s)
s(si ; si)
s(s ; sh, ti)

t(X, ス ; su)
s(su ; su, ce)

t(X, セ ; se)
s(se ; se, ce)

t(X, ソ ; so)
s(so ; so)

t(X, ザ ; za)
s(za ; za)

t(X, ジ ; zi, z)
s(zi ; zi, ji)
s(z ; z, j)

t(X, ズ ; zu)
s(zu ; zu)

t(X, ゼ ; ze)
s(ze ; ze)

t(X, ゾ ; zo)
s(zo ; zo)

t(X, タ ; ta)
s(ta ; ta)

t(X, チ ; ti, ch)
s(ti ; ti, chi)
s(ch, ch)

t(X, ツ ; tu)
s(tu ; tsu, tu)

t(X, ッ ; 0)
s(0 ; 0)

t(X, テ ; te, t)
s(te ; te)
s(t ; t)

t(X, ト ; to)
s(to ; to)

t(X, ダ ; da)
s(da ; da)

t(X, ヂ ; zi, z)

t(X, ヅ ; zu)

t(X, デ ; de, d)
s(de ; de)
s(d ; d)

t(X, ド ; do)
s(do ; do)

t(X, ナ ; na)
s(na ; na)

t(X, ニ ; ni, n)
s(ni ; ni)
s(n ; n)

t(X, ヌ ; nu)
s(nu ; nu)

t(X, ネ ; ne)
s(ne ; ne)

t(X, ノ ; no)
s(no ; no)

t(X, ハ ; ha)
s(ha ; ha)

t(X, ヒ ; hi, h)
s(hi ; hi)
s(h ; h)

t(X, フ ; fu, f)
s(fu ; fu)
s(f ; f, ph)

t(X, ヘ ; he)
s(he ; he)

t(X, ホ ; ho)
s(ho ; ho)

t(X, バ ; ba)
s(ba ; ba, va)

t(X, ビ ; bi, b)
s(bi ; bi, vi)
s(b ; b)
```

```
t(X, ブ ; bu)
s(bu ; bu, vu)

t(X, ベ ; be)
s(be ; be, ve)

t(X, ボ ; bo)
s(bo ; bo, vo)

t(X, パ ; pa)
s(pa ; pa)

t(X, ピ ; pi, p)
s(pi ; pi)
s(p ; p)

t(X, プ ; pu)
s(pu ; pu)

t(X, ペ ; pe)
s(pe ; pe)

t(X, ポ ; po)
s(po ; po)

t(X, マ ; ma)
s(ma ; ma)

t(X, ミ ; mi, m)
s(mi ; mi)
s(m ; m)

t(X, ム ; mu)
s(mu ; mu)

t(X, メ ; me)
s(me ; me)

t(X, モ ; mo)
s(mo ; mo)
```

```
t(X, ヤ ; ya)
s(ya ; ya)

t(X, ユ ; yu)
s(yu ; yu)

t(X, ヨ ; yo)
s(yo ; yo)

t(X, ラ ; ra)
s(ra ; ra, la)

t(X, リ ; ri, r)
s(ri ; ri, li)
s(r ; r, l)

t(X, ル ; ru)
s(ru ; ru, lu)

t(X, レ ; re)
s(re ; re, le)

t(X, ロ ; ro)
s(ro ; ro, lo)

t(X, ワ ; wa)
s(wa ; wa)

t(X, ヲ ; wo)
s(wo ; wo)

t(X, ン ; N)
s(N ; n, m)

t(X, ヴ ; v)
s(v ; v)

t(X, ー ; R)
s(R ; r, 0)
```

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
```

```
t(k, ヤ ; ya)
t(g, ヤ ; ya)
t(s, ヤ ; a)
t(ch, ヤ ; a)
t(z, ヤ ; ya)
t(n, ヤ ; ya)
t(h, ヤ ; ya)
```

```
t(b, ヤ ; ya)
t(p, ヤ ; ya)
t(m, ヤ ; ya)
t(r, ヤ ; ya)
t(v, ヤ ; ya)

t(k, ユ ; yu)
```

```
t(g, ュ ; yu)
t(s, ュ ; u)
t(z, ュ ; yu)
t(ch, ュ ; u)
t(d, ュ ; yu)
t(n, ュ ; yu)
t(h, ュ ; yu)
t(b, ュ ; yu)
t(p, ュ ; yu)
t(m, ュ ; yu)
t(r, ュ ; yu)
t(v, ュ ; yu)

t(k, ョ ; yo)
t(g, ョ ; yo)
t(s, ョ ; o)
t(z, ョ ; yo)
t(ch, ョ ; o)
t(z, ョ ; yo)
t(n, ョ ; yo)
t(h, ョ ; yo)
t(b, ョ ; yo)
t(p, ョ ; yo)
t(m, ョ ; yo)
t(r, ョ ; yo)
t(v, ョ ; yo)

t(ku, ア ; a)
t(tu, ア ; a)
t(f, ア ; a)
t(gu, ア ; a)
t(v, ア ; a)

t(w, イ ; i)
t(t, イ ; i)
t(f, イ ; i)
t(d, イ ; i)
t(v, イ ; i)

t(X, ウ ; u)

t(i, エ ; e)
t(w, エ ; e)
t(s, エ ; e)
t(ch, エ ; e)
t(tu, エ ; e)
t(f, エ ; e)
t(z, エ ; e)
t(v, エ ; e)

t(i, オ ; o)
t(w, オ ; o)
t(s, オ ; o)
t(tu, オ ; o)
t(f, オ ; o)
t(z, オ ; o)
t(v, オ ; o)
```

A.3 Algorithme de retranscription par notre transducteur

Algorithme 3 Retranscription des mots en *katakana* par transducteur

♢ Données :

- `transducteur` : transducteur créé à partir de la grammaire préalablement définie ;
- `entrée` : séquence en katakana à retranscrire ;
- `symboleCourant` : symbole de `entrée` en cours de traitement ;
- `étatCourant` : état courant ;
- `étatsSuivants` : liste des candidats états suivants ;
- `étatVide` : constante indiquant l'état auquel aucun symbole de sortie n'est lié ;
- `sorties` : liste de chaînes retranscrites en alphabet de la séquence entrée

en katakana.

◊ Procédure :

1. `étatCourant` := 0 ;
2. `étatsSuivants` := 0 ;
3. Pour tous les caractères de `entrée` (du premier, 1, à l'avant-dernier, m), réaliser les opérations suivantes ;
   a) `symboleCourant` := `entrée`[i] ;
   b) (Recherche du nouvel état courant parmi les candidats stockés dans `étatsSuivants`)
      Si `étatsSuivants` n'est pas vide, alors ;
      i. Si `étatsSuivants` ne contient qu'un seul élément, alors `étatcourant` := `étatsSuivants` ;
      ii. Sinon, pour chaque élément de `étatsSuivants` (du premier, 1, au dernier, n), réaliser les opérations suivantes ;
         A. `étatcourant` := `étatsSuivants`[i] ;
         B. Si dans le `transducteur`, il existe un chemin partant de `étatcourant` et étiqueté par `symboleCourant`, alors fin d'opération.
         C. Sinon ;
            α. S'il y a encore des éléments à traiter dans `étatsSuivants`, alors continuer ;
            β. Sinon, `étatcourant` := `étatsSuivants`[1], et fin d'opération.
      iii. vider `étatsSuivants` ;
   c) (Initialisation de la liste `étatsSuivants`)
      i. Si dans le `transducteur`, il n'existe aucun chemin partant de `étatCourant` et étiqueté par `symboleCourant`, alors `étatsSuivants` := `étatVide` ;
      ii. Sinon, stocker tous les chemins partant de `étatCourant` et étiquetés par `symboleCourant` dans `étatsSuivants`.
   d) Si `étatCourant` = 0, alors passer au prochain symbole d'entrée et continuer.
   e) Stockage du symbole de sortie lié à `étatCourant` dans `sorties`
   f) Passer au prochain symbole d'entrée et continuer.
4. (Stockage du symbole de sortie lié à l'état suivant)
   a) Stockage du symbole de sortie lié à `étatsSuivants`[1] dans `sorties`.

A.4 Exemples de retranscription à l'aide du transducteur

Nous présentons dans cette section quatre exemples de retranscription, chacun mettant en lumière une particularité présentée dans le second paragraphe de la section 3.3.1 page 101.

L'exemple 3 montre un cas concret d'un état à plusieurs symboles de sortie et l'exemple 4 page suivante celui du passage par l'état vide. Enfin, les deux derniers exemples (ex. 5 page 474 et 6 page 476) présentent des procédures de traitement contenant un choix de transition non déterministe.

Exemple 3 (Pluralité des symboles de sortie liés à un état)
Considérons la séquence d'entrée :

パリ (*pari*, « Paris »)

avec les règles suivantes :

1. `t(X, パ ; pa)`
2. `t(X, リ ; ri, r)`
3. `s(pa ; pa)`
4. `s(ri ; ri, li)`
5. `s(r ; r, l)`

La figure A.1 représente une partie du transducteur créé à partir de cette grammaire, partie concernée par le traitement de la séquence d'entrée considérée. La règle de transition `t(X, リ ; ri, r)` indique qu'au caractère d'entrée リ correspondent deux possibilités d'états suivants, *ri* et *r*. Tenant compte de l'ordre représentant la priorité, le chemin qui amène à l'état *ri* est étiqueté par $リ_1$ et celui qui amène à l'état *r*, par $リ_2$.

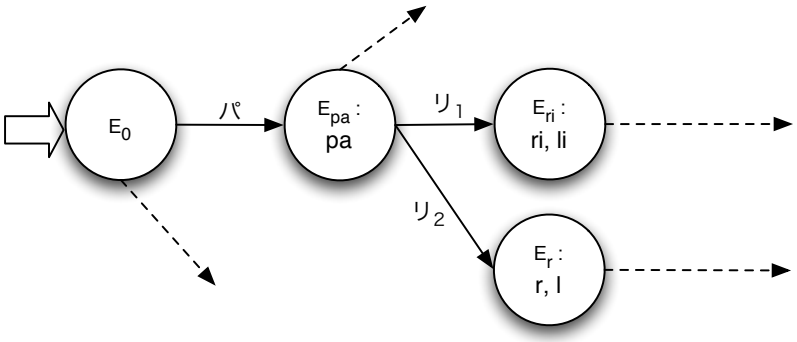


FIG. A.1 – Retranscription du mot en *katakana PARI* (« Paris »)

Initialement, le pointeur pointe sur le premier caractère パ et l'état courant est E_0. La liste des états suivants et la séquence de sortie sont pour l'instant vides.

Séquence d'entrée	État courant	États suivants	Séquence de sortie
⇓ パピリ	E_0		

On initialise la liste des états suivants : on stocke l'état E_{pa} auquel amène le chemin étiqueté par パ.

⇓ パピリ	E_0	E_{pa}	

Comme l'état courant est encore l'état initial E_0, on passe tout de suite au caractère suivant.

パ⇓ ピリ	E_0	E_{pa}	

On met l'état suivant à l'état courant.

パ⇓ ピリ	E_{pa}	E_{pa}	

On vérifie s'il existe un chemin partant de l'état courant et étiqueté par le caractère d'entrée pointé. Comme il en existe un, on vide la liste des états suivants. Puis on stocke le symbole de sortie lié à l'état courant en fin de séquence de sortie.

パ⇓ ピリ	E_{pa}		pa

On met à jours la liste des états suivants. Comme il existe deux chemins partant de l'état courant et étiquetés par リ (caractère courant), on stocke les deux états suivants, d'abord E_{ri} auquel ammène le chemin étiqueté par $リ_1$, puis E_r auquel amène le chemin étiqueté par $リ_2$.

パ⇓ ピリ	E_{pa}	E_{ri}, E_r	pa

Comme tous les caractères ont déjà été traités, on stocke le symbole de sortie lié au premier état appartenant à la liste des états suivants, E_{ri}, en fin de séquence de sortie. Mais, Comme E_{ri} a deux symboles de sortie, *ri* et *li*, qui lui sont liés, on crée deux séquences de sortie, chacune avec un des deux symboles de sortie.

パ⇓ ピリ	E_{pa}	E_{ri}, E_r	pari/pali

Exemple 4 (État vide)
Considérons la séquence d'entrée :

バゲット (*bagetto*, « baguette »)

avec les règles suivantes :

1. `t(X, バ ; ba)`
2. `t(X, ゲ ; ge)`
3. `t(X, ッ ; 0)`
4. `t(X, ト ; to)`

5. `s(ba ; ba, va)`
6. `s(ge ; ge)`
7. `s(0 ; 0)`
8. `s(to ; to)`

La figure A.2 représente une partie du transducteur créé à partir de cette grammaire, partie concernée par le traitement de la séquence d'entrée considérée.

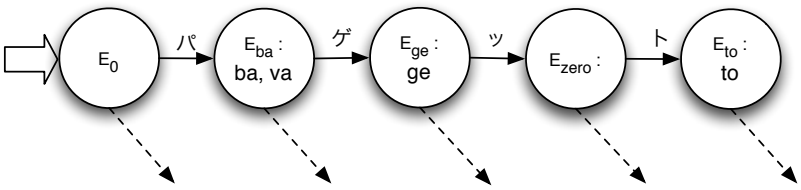


FIG. A.2 – Retranscription du mot en *katakana BAGETTO* (« baguette »)

Initialement, le pointeur pointe sur le premier caractère バ et l'état courant est E_0. La liste des états suivants et la séquence de sortie sont pour l'instant vides.

Séquence d'entrée	État courant	États suivants	Séquence de sortie
⇓ バゲットト	E_0		

On initialise la liste des états suivants : on stocke l'état E_{ba} auquel amène le chemin étiqueté par バ.

⇓ バゲットト	E_0	E_{ba}	

Comme l'état courant est encore l'état initial E_0, on passe tout de suite au caractère suivant.

バ ⇓ ゲットト	E_0	E_{ba}	

On met l'état suivant à l'état courant.

バ ⇓ ゲットト	E_{ba}	E_{ba}	

On vérifie s'il existe un chemin partant de l'état courant et étiqueté par le caractère d'entrée pointé. Comme il en existe un, on vide la liste des états suivants. Puis on stocke le symbole de sortie lié à l'état courant en fin de séquence de sortie. Mais, comme E_{ba} a deux symboles de sortie, *ba* et *va*, qui lui sont liés, on crée deux séquences de sortie, chacune avec un des deux symboles de sortie.

バ ⇓ ゲットト	E_{ba}		ba/va

On met à jours la liste des états suivants : on stocke E_{ge} auquel amène le chemin étiqueté par ゲ.

バ$\overset{\Downarrow}{\text{ゲ}}$ット	E_{ba}	E_{ge}	ba/va

On passe au caractère suivant.

バゲ$\overset{\Downarrow}{\text{ッ}}$ト	E_{ba}	E_{ge}	ba/va

On met l'état suivant à l'état courant.

バゲ$\overset{\Downarrow}{\text{ッ}}$ト	E_{ge}	E_{ge}	ba/va

On vérifie s'il existe un chemin partant de l'état courant et étiqueté par le caractère d'entrée pointé. Comme il en existe un, on vide la liste des états suivants. Puis on stocke le symbole de sortie lié à l'état courant à la fin de chacune des séquences de sortie.

バゲ$\overset{\Downarrow}{\text{ッ}}$ト	E_{ge}		bage/vage

On met à jour la liste des états suivants : on stocke E_{zero} auquel amène le chemin étiqueté par ッ.

バゲ$\overset{\Downarrow}{\text{ッ}}$ト	E_{ge}	E_{zero}	bage/vage

On passe au caractère suivant.

バゲッ$\overset{\Downarrow}{\text{ト}}$	E_{ge}	E_{zero}	bage/vage

On met l'état suivant à l'état courant.

バゲッ$\overset{\Downarrow}{\text{ト}}$	E_{zero}	E_{zero}	bage/vage

On vérifie s'il existe un chemin partant de l'état courant et étiqueté par le caractère d'entrée pointé. Comme il en existe un, on vide la liste des états suivants. Comme aucun symbole de sortie n'est lié à l'état courant, il n'y a pas d'opération de stockage des symboles de sortie.

バゲッ$\overset{\Downarrow}{\text{ト}}$	E_{zero}		bage/vage

On met à jour la liste des états suivants : on stocke E_{to} auquel amène le chemin étiqueté par ト.

バゲッ$\overset{\Downarrow}{\text{ト}}$	E_{zero}	E_{to}	bage/vage

Comme tous les caractères ont déjà été traités, on stocke le symbole de sortie lié à l'état appartenant à la liste des états suivants, E_{to}, à la fin de chacune des séquences de sortie.

バゲッ$\overset{\Downarrow}{\text{ト}}$	E_{zero}	E_{to}	bageto/vageto

Exemple 5 (Transducteur non déterministe)
Considérons la séquence d'entrée :

ミラノ (*mirano*, « Milan »)

avec les règles suivantes :

1. `t(X, ミ ; mi, m)`
2. `t(X, ラ ; ra)`
3. `t(X, ノ ; no)`
4. `s(mi ; mi)`
5. `s(m ; m)`
6. `s(ra ; ra, la)`
7. `s(no ; no)`

La figure A.3 représente une partie du transducteur créé à partir de cette grammaire, partie concernée par le traitement de la séquence d'entrée considérée. La règle de transition `t(X, ミ ; mi, m)` indique qu'au caractère d'entrée ミ correspondent deux possibilités d'états suivants, mi et m. Tenant compte de l'ordre représentant la priorité, le chemin qui amène à l'état mi est étiqueté par ミ$_1$ et celui qui amène à l'état m, par ミ$_2$.

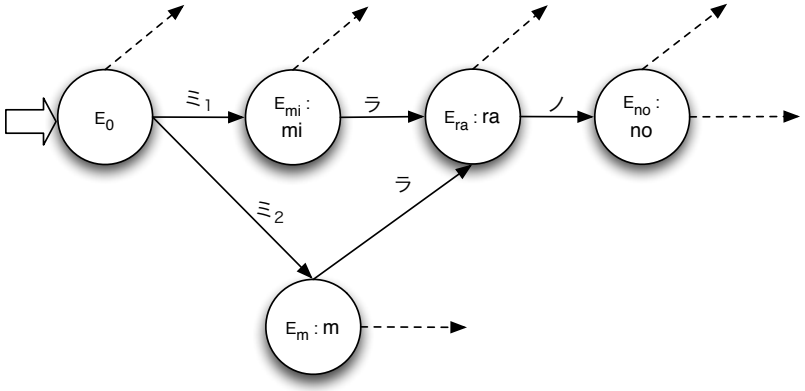


FIG. A.3 – Retranscription du mot en *katakana MIRANO* (« Milan »)

Initialement, le pointeur pointe sur le premier caractère ミ et l'état courant est E_0. La liste des états suivants et la séquence de sortie sont pour l'instant vides.

Séquence d'entrée	État courant	États suivants	Séquence de sortie
⇓ ミラノ	E_0		

On initialise la liste des états suivants. Comme il existe deux chemins étiquetés par ミ (caractère courant), on stocke les deux états suivants, d'abord E_{mi} auquel amène le chemin étiqueté par ミ$_1$, puis E_m auquel amène le chemin étiqueté par ミ$_2$.

ミ⇓ラノ	E_0	E_{mi}, E_m

On passe au caractère suivant.

ミ⇓ラノ	E_0	E_{mi}, E_m

On affecte le premier état suivant à l'état courant.

ミ⇓ラノ	E_{mi}	E_{mi}, E_m

On vérifie s'il existe un chemin partant de l'état courant et étiqueté par le caractère d'entrée pointé. Comme il en existe un, on vide la liste des états suivants. Puis on stocke le symbole de sortie lié à l'état courant à la fin de chacune des séquences de sortie.

ミ⇓ラノ	E_{mi}	mi

On met à jour la liste des états suivants : on stocke E_{ra} auquel amène le chemin étiqueté par ラ.

ミ⇓ラノ	E_{mi}	E_{ra}	mi

On passe au caractère suivant.

ミラ⇓ノ	E_{mi}	E_{ra}	mi

On affecte l'état de la liste des états suivants à l'état courant.

ミラ⇓ノ	E_{ra}	E_{ra}	mi

On vérifie s'il existe un chemin partant de l'état courant et étiqueté par le caractère d'entrée pointé. Comme il en existe un, on vide la liste des états suivants. Puis on stocke le symbole de sortie lié à l'état courant en fin de séquence de sortie. Mais, comme E_{ra} a deux symboles de sortie, *ra* et *la*, qui lui sont liés, on duplique la séquence de sortie pour en créer deux, chacune avec un des deux symboles de sortie à sa fin.

ミラ⇓ノ	E_{ra}	mira/mila

On met à jour la liste des états suivants : on stocke E_{no} auquel amène le chemin étiqueté par ノ.

ミラ⇓ノ	E_{ra}	E_{no}	mira/mila

Comme tous les caractères ont déjà été traités, on stocke le symbole de sortie lié à l'état appartenant à la liste des états suivants, E_{no}, à la fin de chacune des séquences de sortie.

ミラ$\overset{\Downarrow}{ノ}$	E_{ra}	E_{no}	mirano/milano

Exemple 6 (Transducteur non déterministe)
Considérons la séquence d'entrée :

バリュー (*baryû*, « value » ang.)

avec les règles suivantes :

1. `t(X, バ ; ba)`
2. `t(X, リ ; ri, r)`
3. `t(r, ュ ; yu)`
4. `t(X, ー ; R)`
5. `s(ba ; ba, va)`
6. `s(ri ; r, r)`
7. `s(r ; r, l)`
8. `s(yu ; yu)`
9. `s(R ; r, 0)`

La figure A.4 représente une partie du transducteur créé à partir de cette grammaire, partie concernée par le traitement de la séquence d'entrée considérée. La règle de transition `t(X, リ ; ri, r)` indique qu'au caractère d'entrée リ correspondent deux possibilités d'états suivants, ri et r. Tenant compte de l'ordre représentant la priorité, le chemin qui amène à l'état ri est étiqueté par リ$_1$ et celui qui amène à l'état r, par リ$_2$.

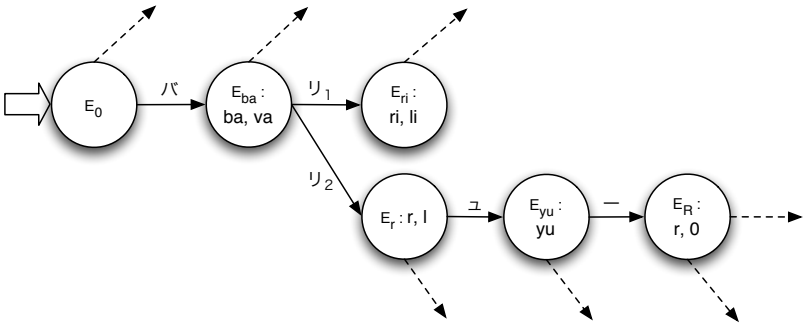


FIG. A.4 – Retranscription du mot en *katakana BARYÛ* (« value » ang.)

Initialement, le pointeur pointe sur le premier caractère バ et l'état courant est E_0. La liste des états suivants et la séquence de sortie sont pour l'instant vides.

Séquence d'entrée	État courant	États suivants	Séquence de sortie
バ$\overset{\Downarrow}{リ}$ュー	E_0		

On initialise la liste des états suivants : on stocke l'état E_{ba} auquel amène le chemin étiqueté par バ.

⇓ バリュー	E_0	E_{ba}

Comme l'état courant est encore l'état initial E_0, on passe tout de suite au caractère suivant.

バ⇓リュー	E_0	E_{ba}

On met l'état suivant à l'état courant.

バ⇓リュー	E_{ba}	E_{ba}

On vérifie s'il existe un chemin partant de l'état courant et étiqueté par le caractère d'entrée pointé. Comme il en existe un, on vide la liste des états suivants. Puis on stocke le symbole de sortie lié à l'état courant en fin de séquence de sortie. Mais, comme E_{ba} a deux symboles de sortie, *ba* et *va*, qui lui sont liés, on crée deux séquences de sortie, chacune avec un des deux symboles de sortie.

バ⇓リュー	E_{ba}	ba/va

On met à jour la liste des états suivants. Comme il existe deux chemins partant de l'état courant et étiquetés par リ (caractère courant), on stocke les deux états suivants, d'abord E_{ri} auquel amène le chemin étiqueté par $リ_1$, puis E_r auquel amène le chemin étiqueté par $リ_2$.

バ⇓リュー	E_{ba}	E_{ri}, E_r	ba/va

On passe au caractère suivant.

バリ⇓ュー	E_{ba}	E_{ri}, E_r	ba/va

On met le premier état appartenant à la liste des états suivants à l'état courant.

バリ⇓ュー	E_{ri}	E_{ri}, E_r	ba/va

On vérifie s'il existe un chemin partant de l'état courant et étiqueté par le caractère d'entrée pointé. Comme il n'en existe aucun, on met le second état appartenant à la liste des états suivants à l'état courant.

バリ⇓ュー	E_r	E_{ri}, E_r	ba/va

On vérifie s'il existe un chemin partant de l'état courant et étiqueté par le caractère d'entrée pointé. Comme il en existe un, on vide la liste des états suivants. Puis on stocke le symbole de sortie lié à l'état courant en fin de séquence de sortie. Mais, comme E_r a deux symboles de sortie, *r* et *l*, qui lui sont liés, on duplique la séquence de sortie stockée (ba/va) pour créer quatre combinaisons avec ces deux symboles de sortie.

バリ⇓ュー	E_r	bar/var/bal/val

On met à jour la liste des états suivants. On stocke E_{yu} auquel amène le chemin étiqueté par ュ.

バリ $\overset{\Downarrow}{\text{ュ}}$ −	E_r	E_{yu}	bar/var/bal/val

On passe au caractère suivant.

バリュ $\overset{\Downarrow}{}$	E_r	E_{yu}	bar/var/bal/val

On met l'état suivant à l'état courant.

バリュ $\overset{\Downarrow}{}$	E_{yu}	E_{yu}	bar/var/bal/val

On vérifie s'il existe un chemin partant de l'état courant et étiqueté par le caractère d'entrée pointé. Comme il en existe un, on vide la liste des états suivants. Puis on stocke le symbole de sortie lié à l'état courant à la fin de chacune des séquences de sortie.

バリュ $\overset{\Downarrow}{}$	E_{yu}		baryu/varyu /balyu/valyu

On met à jour la liste des états suivants. On stocke E_R auquel amène le chemin étiqueté par −.

バリュ $\overset{\Downarrow}{}$	E_{yu}	E_R	baryu/varyu /balyu/valyu

Comme tous les caractères ont déjà été traités, on stocke le symbole de sortie lié à l'état appartenant à la liste des états suivants, E_R, à la fin de chacune des séquences de sortie. Mais, comme E_R a deux symboles de sortie, *r* et un vide, qui lui sont liés, on duplique la séquence de sortie stockée (baryu/varyu/balyu/valyu) pour créer huit combinaisons avec ces deux symboles de sortie.

バリュ $\overset{\Downarrow}{}$	E_{yu}	E_R	baryur/varyur /balyur/valyur /baryu/varyu /balyu/valyu

A.5 Résultat de la retranscription

1. アフリカ (katakana)　: freq = 10 / 8
   - afurika　- afurica
   - afulika　- afulica

2. イニシアティブ (katakana)　: freq = 1 / 1
   - inisiatibu　- inisiativu
   - ynisiatibu　- ynisiativu
   - inisiatybu　- inisiatyvu
   - ynisiatybu　- ynisiatyvu

3. インフラ (katakana)　: freq = 1 / 1
   - infura　- infula
   - ynfura　- ynfula
   - imfura　- imfula
   - ymfura　- ymfula

4. エイズ (katakana)　: freq = 2 / 2
   - eizu　- eyzu

5. カナナスキス (katakana)　: freq = 2 / 2
   - kananasukisu　- kananasukice
   - cananasukisu　- cananasukice
   - kananacekisu　- kananacekice
   - cananacekisu　- cananacekice

6. グループ (katakana)　: freq = 1 / 1
   - gururpu　- gurupu
   - gulurpu　- gulupu

7. コンタクト (katakana)　: freq = 1 / 1
   - kontakuto　- kontacuto
   - contakuto　- contacuto
   - komtakuto　- komtacuto
   - comtakuto　- comtacuto

8. サブサハラ (katakana)　: freq = 2 / 2
   - sabusahara　- sabusahala
   - savusahara　- savusahala

9. システム (katakana)　: freq = 2 / 1
   - sisutemu　- taipu
   - sicetemu　- taypu

10. タイミング (katakana) : freq = 1 / 1
   - taimingu　- taimimgu
   - taymingu　- taymimgu

11. テント (katakana) : freq = 1 / 1
    - tento - temto

12. トン (katakana) : freq = 1 / 1
    - ton - tom

13. ドル (katakana) : freq = 4 / 2
    - doru - dolu

14. ニーズ (katakana) : freq = 10 / 10
    - nirzu - nizu

15. バイオテクノロジー (katakana) : freq = 1 / 1
    - baiotekunorozir - baiotekunorozi
    - vaiotekunorozir - vaiotekunorozi
    - bayotekunorozir - bayotekunorozi
    - vayotekunorozir - vayotekunorozi
    - baiotecunorozir - baiotecunorozi
    - vaiotecunorozir - vaiotecunorozi
    - bayotecunorozir - bayotecunorozi
    - vayotecunorozir - vayotecunorozi
    - baiotekunolozir - baiotekunolozi
    - vaiotekunolozir - vaiotekunolozi
    - bayotekunolozir - bayotekunolozi
    - vayotekunolozir - vayotekunolozi
    - baiotecunolozir - baiotecunolozi
    - vaiotecunolozir - vaiotecunolozi
    - bayotecunolozir - bayotecunolozi
    - vayotecunolozir - vayotecunolozi
    - baiotekunorojir - baiotekunoroji
    - vaiotekunorojir - vaiotekunoroji
    - bayotekunorojir - bayotekunoroji
    - vayotekunorojir - vayotekunoroji
    - baiotecunorojir - baiotecunoroji
    - vaiotecunorojir - vaiotecunoroji
    - bayotecunorojir - bayotecunoroji
    - vayotecunorojir - vayotecunoroji
    - baiotekunolojir - baiotekunoloji
    - vaiotekunolojir - vaiotekunoloji
    - bayotekunolojir - bayotekunoloji
    - vayotekunolojir - vayotekunoloji
    - baiotecunolojir - baiotecunoloji
    - vaiotecunolojir - vaiotecunoloji
    - bayotecunolojir - bayotecunoloji
    - vayotecunolojir - vayotecunoloji

16. パートナーシップ (katakana) : freq = 1 / 1
    - partonarsipu - partonasipu
    - patonarsipu - patonasipu

17. フォーラム (katakana) : freq = 2 / 2

- forramu	- forlamu
- phorramu	- phorlamu
- foramu	- folamu
- phoramu	- pholamu

18. プログラム (katakana) : freq = 2 / 2

- puroguramu	- purogulamu
- puloguramu	- pulogulamu

19. レベル (katakana) : freq = 2 / 2

- reberu	- rebelu
- leberu	- lebelu
- reveru	- revelu
- leveru	- levelu

20. ワクチン (katakana) : freq = 1 / 1

- wakutin	- wakutim
- wacutin	- wacutim
- wakuchin	- wakuchim
- wacuchin	- wacuchim

A.6 Résultat du calcul de la similarité entre les retranscriptions et les mots français

```
0. afurika ---> 1. afrique [0.222299]
1. afulika --->
2. afurica ---> 1. africains [0.466891]
3. afulica --->
4. inisiatibu --->
5. ynisiatibu --->
6. inisiatybu --->
7. ynisiatybu --->
8. inisiativu --->
9. ynisiativu --->
10. inisiatyvu --->
11. ynisiatyvu --->
12. infura --->
13. ynfura --->
14. imfura --->
15. ymfura --->
16. infula --->
17. ynfula --->
18. imfula --->
19. ymfula --->
20. eizu --->
21. eyzu --->
22. kananasukisu --->
        1. kananaskis [1.000000]
23. cananasukisu --->
24. kananacekisu --->
25. cananacekisu --->
26. kananasukice --->
27. cananasukice --->
28. kananacekice --->
29. cananacekice --->
30. gururpu --->
31. gulurpu --->
32. gurupu ---> 1. groupe,
        2. groupes [0.535164]
33. gulupu --->
34. kontakuto --->
35. contakuto --->
36. komtakuto --->
37. comtakuto --->
38. kontacuto --->
39. contacuto --->
        1. contact [0.788758]
40. komtacuto --->
41. comtacuto --->
42. sabusahara --->
        1. subsaharienne [0.448158]
43. savusahara --->
44. sabusahala --->
45. savusahala --->
46. sisutemu ---> 1. systèmes [0.399411]
47. sicetemu ---> 1. secteurs [0.310653]
48. taipu --->
49. taypu ---> 1. types [0.272641]
50. taimingu --->
51. taymingu --->
52. taimimgu --->
53. taymimgu --->
54. tento ---> 1. tentes [0.375310]
55. temto --->
56. ton --->
57. tom --->
58. doru --->
59. dolu --->
60. nirzu --->
61. nizu --->
62. baiotekunorozir --->
63. vaiotekunorozir --->
64. bayotekunorozir --->
65. vayotekunorozir --->
66. baiotecunorozir --->
67. vaiotecunorozir --->
68. bayotecunorozir --->
69. vayotecunorozir --->
70. baiotekunolozir --->
71. vaiotekunolozir --->
72. bayotekunolozir --->
73. vayotekunolozir --->
74. baiotecunolozir --->
75. vaiotecunolozir --->
76. bayotecunolozir --->
77. vayotecunolozir --->
78. baiotekunorojir --->
79. vaiotekunorojir --->
80. bayotekunorojir --->
81. vayotekunorojir --->
82. baiotecunorojir --->
83. vaiotecunorojir --->
84. bayotecunorojir --->
85. vayotecunorojir --->
86. baiotekunolojir --->
87. vaiotekunolojir --->
88. bayotekunolojir --->
89. vayotekunolojir --->
90. baiotecunolojir --->
91. vaiotecunolojir --->
92. bayotecunolojir --->
93. vayotecunolojir --->
```

```
94. baiotekunorozi --->
95. vaiotekunorozi --->
96. bayotekunorozi --->
97. vayotekunorozi --->
98. baiotecunorozi --->
99. vaiotecunorozi --->
100. bayotecunorozi --->
101. vayotecunorozi --->
102. baiotekunolozi --->
103. vaiotekunolozi --->
104. bayotekunolozi --->
105. vayotekunolozi --->
106. baiotecunolozi --->
       1. biotechnologies [0.510204]
107. vaiotecunolozi --->
108. bayotecunolozi --->
109. vayotecunolozi --->
110. baiotekunoroji --->
111. vaiotekunoroji --->
112. bayotekunoroji --->
113. vayotekunoroji --->
114. baiotecunoroji --->
115. vaiotecunoroji --->
116. bayotecunoroji --->
117. vayotecunoroji --->
118. baiotekunoloji --->
119. vaiotekunoloji --->
120. bayotekunoloji --->
121. vayotekunoloji --->
122. baiotecunoloji --->
123. vaiotecunoloji --->
124. bayotecunoloji --->
125. vayotecunoloji --->
126. partonarsipu ---> 1. partenariat [0.505225]
127. patonarsipu --->
128. partonasipu --->
129. patonasipu --->
130. forramu --->
131. phorramu --->
132. foramu --->
133. phoramu ---> 1. promouvoir [0.200687]
134. forlamu --->
135. phorlamu --->
136. folamu --->
137. pholamu --->
138. puroguramu --->
       1. programme, 2. programmes [0.672237]
139. puloguramu --->
140. purogulamu --->
141. pulogulamu --->
142. reberu --->
143. leberu --->
144. reveru --->
145. leveru --->
146. rebelu --->
147. lebelu --->
148. revelu --->
149. levelu --->
150. wakutin --->
151. wacutin --->
152. wakuchin --->
153. wacuchin --->
154. wakutim --->
155. wacutim --->
156. wakuchim --->
157. wacuchim --->
```

A.7 Problèmes liés à l'encodage dans le traitement multilingue

L'encodage est une question assez pénible lors de la conception d'un système de traitement automatique des langues. Sans parler des textes aux formats propres aux logiciels, même deux textes français au format pur « texte » peuvent avoir un encodage différent si l'un est créé sur un Macintosh et que l'autre est un fichier Windows. De même pour les textes japonais, l'encodage des textes créés sur Mac, Windows ou Linux diffère, car il dépend du système d'exploitation.

Pour qu'un système de TAL supporte différents encodages, il faut une phase préparatoire de transcodage des textes pour les traiter correctement. Il va de soi que l'encodage pose un problème sérieux lorsqu'il s'agit du traitement de deux langues d'écritures très différentes telles que le français et le japonais.

Les jeux de caractères utilisés pour les fichiers français peuvent coder correctement un texte français, ce qui n'était pas le cas quand on ne possédait comme moyen de codage que l'ASCII. Cependant, ils sont incapables de coder ne serait-ce qu'un caractère japonais. Les encodages utilisés par les Japonais peuvent coder correctement le japonais et l'anglais simultanément, mais il leur est impossible de traiter à la fois le japonais, le chinois, l'arabe et le français.

Unicode a été créé pour résoudre cette situation dans laquelle l'encodage dépend beaucoup trop du système et du logiciel, et où rien ne permet de coder correctement toutes les langues connues.

A.7.1 Qu'est-ce qu'Unicode ?

« Unicode » a vu le jour en 1989 à travers un consortium de constructeurs de logiciels multilingues. Le système Unicode est basé sur un codage 16 bits, capable de contenir 65 536 caractères.

En 1991, Unicode a été intégré à la norme ISO 10646 (Kuhn, 1999) sous le nom ISO-10646-UCS-2. UCS, abréviation de *Universal Character Set*, désigne le jeu de caractères universel défini par le standard international ISO 10646, qui est en fait une table de codage sur 4 octets, incluant tous les caractères nécessaires pour représenter toutes les langues connues dans le monde, y compris les langues mortes que nous ne savons pas encore déchiffrer.

Ces deux organisations, le consortium Unicode et *International Organization for Standardization* (ISO), qui travaillaient au départ séparément, ont donc uni leurs efforts pour la création d'une table de codage unique et avancée dont nous avions vraiment besoin. Aujourd'hui, ces deux organisations existent et publient indépendamment leur standard respectif, mais ils restent et resteront toujours compatibles.

Unicode ou UCS assigne à chaque caractère un code numérique et un nom officiel. Les caractères les plus utilisés sont placés dans l'un des premiers 65 534

emplacements (U+0000 À U+FFFD[1]). Ce sous-ensemble de 2 octets est appelé *Basic Multilingual Plane* (BMP) ou *Plane 0*. Les caractères U+0000 à U+007F sont identiques aux caractères ASCII et la colonne U+0000 à U+00FF correspond à ISO 8859-1 (Latin 1).

A.7.2 Encodages d'Unicode

Unicode ou UCS ne sont cependant que des tables de codages, qui attribuent à chaque caractère un numéro. Mais comment transposer ces numéros en séquences de bits ? Il existe différentes alternatives.

Différentes possibilités d'encodage

ISO 10646-1 définit les encodages UCS-2 et UCS-4 :
- UCS-2 : chaque caractère est représenté par deux octets. Cet encodage peut représenter seulement les 65 536 premiers caractères de BMP.
- UCS-4 : chaque caractère est représenté par 4 octets. Il est donc possible de représenter tous les caractères de UCS ou Unicode.

Unicode ne possédant, au départ, que les caractères appartenant au BMP, utilisait UCS-2. Mais lorsqu'Unicode a étendu son ensemble en définissant des caractères en dehors de BMP, les encodages UTF-16 et UTF-32 ont également été définis :
- UTF-16 : les 65 536 premiers caractères sont représentés par deux octets, les autres par quatre.
- UTF-32 : les caractères sont représentés par quatre octets, ce qui est identique à UCS-4.

Outre les encodages présentés précédemment, un autre encodage appelé UTF-8 a été introduit dans UCS et Unicode pour fournir un encodage multioctet (*multibyte*) compatible ASCII.
- UTF-8 : 128 caractères sont encodés en utilisant 1 octet : les caractères ASCII. 1920 caractères sont encodés en utilisant deux octets : le latin, le grec, le cyrillique, le copte, l'arménien, l'hébreu et les caractères arabes. 63 488 caractères sont encodés en utilisant 3 octets, le chinois et le japonais entre autres. Les 2 147 418 112 caractères restants (non encore assignés) peuvent être encodés en utilisant 4, 5 ou 6 octets.

Encodage et implémentation en C++

Nous avons adopté l'encodage UTF-8. À l'intérieur du programme, les caractères sont stockés dans des objets de type `wchar_t` (*wide character*, caractère étendu). Ce type est aujourd'hui officiellement destiné à être utilisé pour les va-

[1] Un nombre hexadécimal représentant un code UCS ou Unicode est conventionnellement précédé par « U+ » tel que dans « U+0041 ».

leurs sur 32 bits de la norme ISO 10646, indépendamment du paramètre *local*[2] courant utilisé. Différentes fonctions de conversion introduites par ISO C interviendront lors de l'exécution pour convertir en wchar_t les caractères lus en encodage multioctet UTF-8.

A.7.3 Problèmes d'Unicode liés au traitement du japonais

Malgré les avantages que nous venons de citer, Unicode possède également quelques inconvénients. Nous présentons dans cette section trois principaux problèmes d'Unicode qui se posent notamment lors du traitement du japonais.

Un code pour plusieurs caractères : unification *Han*

C'est le problème le plus général des trois concernant non seulement le japonais, mais aussi le chinois et le coréen.

Afin de permettre un traitement informatique des idéogrammes suffisamment simple pour réaliser l'échange et le traitement de données numérisées entre différents pays, plus de 120 000 idéogrammes différents ont été regroupés et simplifiés selon un schéma appelé « Unification Han » pour être ramenés à 20 992 signes. Par cette opération, deux idéogrammes ayant deux formes abstraites semblables dans des jeux de caractères nationaux différents peuvent être unifiés. La figure A.5 page ci-contre en montre un exemple, où trois idéogrammes sont représentés par un seul et même caracère Unicode.

Ce choix a été fortement critiqué notamment par les Japonais qui le considéraient comme non respectueux des cultures concernées. Toutefois, depuis la définition du supplément aux idéogrammes unifiés dans la version 3.0 d'Unicode, ce problème commence à être résolu.

Plusieurs codes pour un caractère

Dans le système d'écriture du japonais, il existe, comme pour le français, des caractères syllabiques à signe diacritique. Deux signes existent : deux traits obliques (*dakuten*) et un rond (*han dakuten*).

Placés en haut à droite d'un caractère de base, le *dakuten* (à gauche dans la figure A.6 page suivante) indique la sonorisation de la consonne constituant le caractère de base. Ainsi, avec *dakuten*, le caractère « *ka* » devient « *ga* », « *ki* » devient « *gi* », ainsi de suite.

Le *han dakuten* (à droite dans la figure A.6 page ci-contre) peut être ajouté uniquement aux caractères *ha, hi, fu, he* et *ho*, pour représenter respectivement *pa, pi, pu, pe* et *po*.

Ces diacritiques peuvent être codés avec Unicode de deux manières différentes, comme représenté figure A.7. En effet, les signes diacritiques possédant

[2] Ensemble de variables systèmes définissant les propriétés propres à chaque langue/pays, telles que le format de date, le séparateur décimal, le symbole monétaire, etc.

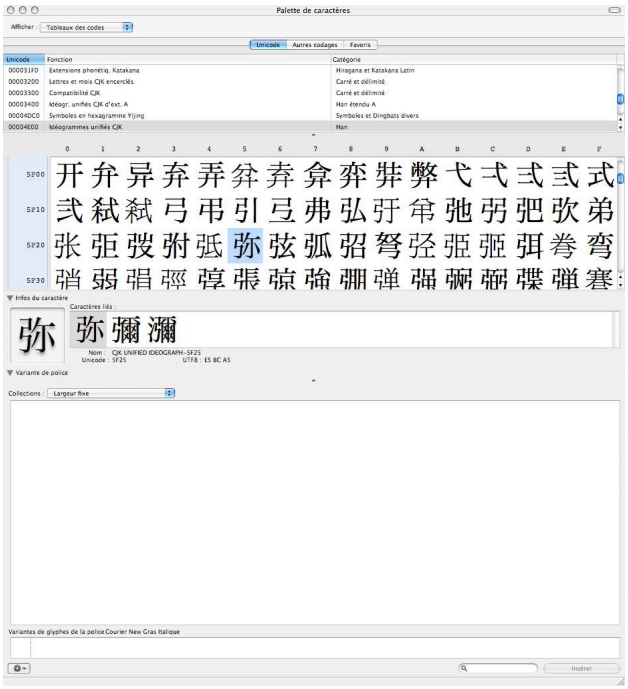


FIG. A.5 – Exemple du code 5F25 représentant trois caractères

FIG. A.6 – Ajout des signes diacrités (*dakuten* à gauche, *han dakuten* à droite)

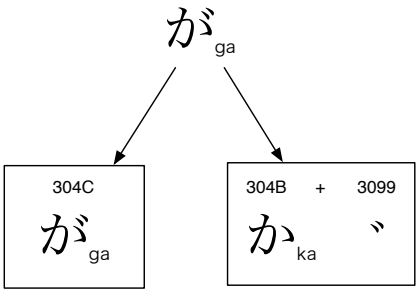


FIG. A.7 – Deux possibilités pour coder le caractère diacritique *ga* avec Unicode

eux-mêmes leur propre code (3099 pour *dakuten* et 309A pour *han dakuten*), un caractère à signe diacritique est codé soit par son propre code, soit par le code de son caractère de base + le code du signe diacritique utilisé. Étant donné que nous ne pouvons pas connaître le codage utilisé par le texte d'entrée *a priori*, le système doit savoir traiter correctement ces caractères quelle que soit la manière dont ils sont codés.

Caractères en largeur pleine et en demi-largeur

À l'aube de l'informatique, les Japonais utilisaient seulement les caractères codés dans la zone dîte ASCII étendu en 8 bits, définie par JIS X 0201 du *Japanese Industrial Standard*. Cette liste, extrêmement restreinte par rapport au système d'écriture réel, ne contenait donc ni *hiragana* ni, bien entendu, aucun idéogramme, mais seulement les *katakana* et les symboles de ponctuation japonais.

Plus tard, le standard JIS X 0208 a été défini pour tous les caractères japonais codés sur deux octets. À l'époque, les caractères définis sur deux octets étaient affichés (ou imprimés) avec une largeur double de celle des caractères à un seul octet, d'où l'appellation de *zenkaku* (*fullwidth*) pour les premiers et *hankaku* (*halfwidth*) pour les seconds.

Tenant compte de cet historique, Unicode consacre la zone de FF00 à FFEF aux « *Halfwidth and Fullwidth Forms* » regroupant les chiffres, l'alphabet romain et différents symboles en largeur pleine, ainsi que les symboles de ponctuation et les *katakana* – définis par JIS X 0201 – en demi-largeur.

Avec le développement des échanges de données numérisées, les *halfwidth katakana* étaient considérés comme une source de problème pour le décodage. Leur utilisation est ainsi aujourd'hui très réduite. Néanmoins, pour les symboles et, surtout, pour les chiffres, les deux formes sont encore largement employées, ce qui pose des problèmes non négligeables lors du traitement des textes japonais. Les systèmes traitant le japonais doivent ainsi reconnaître que « 2005 » a la même signification que « ２００５ » qui sont, en terme de codage Unicode, complètement différents, à savoir 0032-0030-0030-0035 et FF12-FF10-FF10-FF15 (sans parler des deux manières d'écrire 2005 en idéogrammes : « 二〇〇五 » et « 二千五 » !).

A.8 Liste des mots grammaticaux

Marques de négation

ne, n', pas, non.

Conjonctions de coordination

et, ni, ou, mais, car, or.

Conjonctions de subordination

comme, lorsque, lorsqu', puisque, quand, que, qu', quoique, si, s'.

Articles

le, la, les, l', au, aux, du, des,
un, une, des, de, d', du, de la, de l'.

Prépositions

à, après, avant, avec, chez, concernant, contre, dans, de, depuis, derrière, dès, devant, durant, en, entre, envers, hormis, hors, jusque, malgré, moyennant, nonobstant, outre, par, parmi, pendant, pour, près, sans, sauf, selon, sous, suivant, sur, touchant, vers, via.

Adjectifs possessifs et démonstratifs

mon, ma, ton, ta, son, sa, notre, votre, leur,
mes, tes, ses, nos, vos, leurs,
ce, cet, cette, ces.

Pronoms

je, j', tu, il, elle, nous, vous, ils, elles, on,
me, m', te, t', le, l', la, lui, se, s', nous, vous, les, leur, en, y,
moi, toi, lui, elle, soi, nous, vous, eux, elles, soi,
mien, tien, sien, nôtre, vôtre, leur,
mienne, tienne, sienne, nôtre, vôtre, leur,
miens, tiens, siens, nôtres, vôtres, leurs,
miennes, tiennes, siennes,
ce, c', ceci, cela, ça, ç',
celui, celui-ci, celui-là, celle, celle-ci, celle-là,
ceux, ceux-ci, ceux-là, celles, celles-ci, celles-là,
qui, que, qu', quoi, dont, où,
lequel, laquelle, lesquels, lesquelles,
duquel, de laquelle, desquels, desquelles,
auquel, à laquelle, auxquels, auxquelles.

Verbes auxiliaires

ai, as, a, avons, avez, ont,
avais, avais, avait, avions, aviez, avaient,
aurai, auras, aura, aurions, auriez, auront,
aurais, aurais, aurait, aurions, auriez, auraient,
aie, aies, ait, ayons, ayez, aient,

eus, eus, eut, eûmes, eûtes, eurent,
suis, es, est, sommes, êtes, sont,
étais, étais, était, étions, étiez, étaient,
serai, seras, sera, serons, serez, seront,
serais, serais, serait, serions, seriez, seraient,
sois, sois, soit, soyons, soyez, soient,
fus, fus, fut, fûmes, fûtes, furent.

Semi-auxiliaires ou verbes supports

vais, vas, va, allons, allez, vont,
allais, allais, allait, allions, alliez, allaient,
irai, iras, ira, irons, irez, iront,
irais, irais, irait, irions, iriez, iraient,
aille, ailles, aille, allions, alliez, aillent,
allai, allas, alla, allâmes, allâtes allèrent,
arrête, arrêtes, arrête, arrêtons, arrêtez, arrêtent,
arrêtais, arrêtais, arrêtait, arrêtions, arrêtiez, arrêtaient,
arrêterai, arrêteras, arrêtera, arrêterons, arrêterez, arrêteront,
arrêterais, arrêterais, arrêterait, arrêterions, arrêteriez, arrêteraient,
arrête, arrêtes, arrête, arrêtions, arrêtiez, arrêtent,
arrêtai, arrêtas, arrêta, arrêtâmes, arrêtâtes, arrêtèrent,
dois, dois, doit, devons, devez, doivent,
devais, devais, devait, devions, deviez, devaient,
devrai, devras, devra, devrons, devrez, devront,
devrais, devrais, devrait, devrions, devriez, devraient,
doive, doives, doive, devions, deviez, doivent,
dus, dus, dut, dûmes, dûtes, durent,
faut, fallait, faudra, faudrait, faille, fallut,
laisse, laisses, laisse, laissons, laissez, laissent,
laissais, laissais, laissait, laissions, laissiez, laissaient,
laisserai, laisseras, laissera, laisserons, laisserez, laisseront,
laisserais, laisserais, laisserait, laisserions, laisseriez, laisseraient,
laisse, laisses, laisse, laissions, laissiez, laissent,
laissai, laissas, laissa, laissâmes, laissâtes, laissèrent,
parais, parais, paraît, paraissons, paraissez, paraissent,
paraissais, paraissais, paraissait, paraissions, paraissiez, paraissaient,
paraîtrai, paraîtras, paraîtra, paraîtrons, paraîtrez, paraîtront,
paraîtrais, paraîtrais, paraîtrait, paraîtrions, paraîtriez, paraîtraient,
paraisse, paraisses, paraisse, paraissions, paraissiez, paraissent,
parus, parus, parut, parûmes, parûtes, parurent,
semble, sembles, semble, semblons, semblez, semblent,
semblais, semblais, semblait, semblions, sembliez, semblaient,
semblerai, sembleras, semblera, semblerons, semblerez, sembleront,

semblerais, semblerais, semblerait, semblerions, sembleriez, semble-
raient,
semble, sembles, semble, semblions, sembliez, semblent,
semblai, semblas, sembla, semblâmes, semblâtes, semblèrent,
peux, peux, peut, pouvons, pouvez, peuvent,
pouvais, pouvais, pouvait, pouvions, pouviez, pouvaient,
pourrai, pourras, pourra, pourrons, pourrez, pourront,
pourrais, pourrais, pourrait, pourrions, pourriez, pourraient,
puisse, puisses, puisse, puissions, puissiez, puissent,
pus, pus, put, pûmes, pûtes, purent,
sais, sais, sait, savons, savez, savent,
savais, savais, savait, savons, savez, savent,
saurai, sauras, saura, saurons, saurez, sauront,
saurais, saurais, saurait, saurions, sauriez, sauraient,
sache, saches, sache, sachions, sachiez, sachent,
sus, sus, sut, sûmes, sûtes, surent,
veux, veux, veut, voulons, voulez, veulent,
voulais, voulais, voulait, voulions, vouliez, voulaient,
voudrai, voudras, voudra, voudrons, voudrez, voudront,
voudrais, voudrais, voudrait, voudrions, voudriez, voudraient,
veuille, veuilles, veuille, voulions, vouliez, veuillent,
voulus, voulus, voulut, voulûmes, voulûtes, voulurent,
manques, manque, manquons, manquez, manquent,
manquais, manquais, manquait, manquions, manquiez, manquaient,
manquerai, manqueras, manquera, manquerons, manquerez, man-
queront,
manquerais, manquerais, manquerait, manquerions, manqueriez,
manqueraient,
manque, manques, manquions, manquiez, manquent,
manquai, manquas, manqua, manquâmes, manquâtes, manquèrent,
risques, risque, risquons, risquez, risquent,
risquais, risquais, risquait, risquions, risquiez, risquaient,
risquerai, risqueras, risquera, risquerons, risquerez, risqueront,
risquerais, risquerais, risquerait, risquerions, risqueriez, risqueraient,
risque, risques, risquions, risquiez, risquent,
risquai, risquas, risqua, risquâmes, risquâtes, risquèrent,
fais, fais, fait, faisons, faites, font,
faisais, faisais, faisait, faisons, faisez, faisaient,
ferai, feras, fera, ferons, ferez, feront,
ferais, ferais, ferait, ferions, feriez, feraient,
fasses, fasses, fasse, fassions, fassiez, fassent,
fis, fis, fit, fîmes, fîtes, firent.

Adverbes

tout, tous, toute, toutes, très.

ANNEXE B

ANNEXE : GRAMMAIRE POUR LA DÉTECTION DES PROPOSITIONS DU FRANÇAIS

B.1 Trois éléments primaires

Afin de définir les règles permettant la reconnaissance des subordonnées – qui peuvent apparaître à différentes positions – de manière économique (pour des raisons pratiques) à partir des résultats de *chunking*, nous avons d'abord défini trois éléments primaires de la phrase : le syntagme verbal (sv), le syntagme nominal (sn) et une dernière catégorie qui regroupe d'autres compléments (cmp).

B.2 Définition de la phrase

Nous avons défini formellement une phrase comme suit[1] :

$$\text{phrase} \rightarrow \text{[open-g], phrase.} \tag{B.1}$$
$$\text{phrase} \rightarrow \text{sep2, phrase.} \tag{B.2}$$
$$\text{phrase} \rightarrow \text{(sip | scamb | sque), proposition.} \tag{B.3}$$
$$\text{phrase} \rightarrow \text{(cmp, [vrg])?, proposition, coordonnee?.} \tag{B.4}$$
$$\text{phrase} \rightarrow \text{sub.} \tag{B.5}$$
$$\text{phrase} \rightarrow \text{sn.} \tag{B.6}$$
$$\text{phrase} \rightarrow \text{cmp.} \tag{B.7}$$

[1] Les éléments entre crochets sont des terminaux.

La règle B.1 permet de traiter les phrases commençant par des guillemets ouvrants, et la règle B.2, celles commençant par une conjonction de coordination. La règle B.3 définit les phrases interrogatives et exclamatives commençant par un marqueur donné.

La règle B.4, règle principale, définit que la phrase est constituée d'une proposition, éventuellement précédée par un cmp détaché, et éventuellement suivie d'une proposition de coordination.

La règle B.5 est la règle dédiée au traitement des subordonnées constituant toutes seules la phrase et les règles B.6 et B.7, celles pour reconnaître les phrases sans verbe constituées respectivement d'un sn et d'un cmp. Ces règles, en contradiction avec notre définition de la phrase basée sur l'opposition sujet-prédicat, sont d'autant plus importantes que les phrases graphiques ne correspondent pas toujours aux unités que nous souhaitons appeler phrases.

B.3 Définition des connecteurs

Avant d'étudier les règles définissant les sous-phrases, examinons celles définissant différents connecteurs. Nous avons réalisé les études linguistiques sur la typologie des connecteurs dans la section 4.7. Nous allons ici les récapituler.

B.3.1 Typologie des connecteurs

En nous basant sur l'étude des positions d'apparition des subordonnées, nous avons réalisé une classification des connecteurs, mots en « qu- », et nous avons défini quatre types de connecteurs :

1. Qui, Que, Dont, Où : connecteurs isolés

   (respectivement) **qui, que, dont, où**
   comportement particulier ;

2. Camb : connecteurs ambigus

   quand, comme, si
   apparaissant en position post-verbale, en positions initiale/finale et en position post-nominale ;

3. IP : indicateurs de propositions

   quel (et ses formes fléchies), **combien, comment, pourquoi**
   apparaissant seulement en position post-verbale ;

4. Rel : connecteurs relatifs

   quoi, lequel (et ses formes fléchies)
   apparaissant en position post-verbale et en position post-nominale.

B.3.2 Règles des connecteurs

Les règles des connecteurs définissent les constituants de phrase qui comportent comme élément central un connecteur de proposition, et qui constituent avec la proposition qui les suit une subordonnée.

Connecteur squi C'est le connecteur constitué avec le pronom « qui ».

$$\text{squi} \rightarrow [\text{open\_g}], \text{squi}. \tag{B.8}$$
$$\text{squi} \rightarrow [\text{qui}], (\text{cmpinc} \mid \text{subdt}). \tag{B.9}$$
$$\text{squi} \rightarrow [\text{qui}]. \tag{B.10}$$

La règle B.10 est la règle basique définissant le connecteur squi constitué du seul mot « qui ». La règle B.8 permet de traiter les subordonnées introduites par « qui » précédées par des guillemets ouvrants. La règle B.9 est dédiée au traitement des syntagmes ou des propositions détachées-insérées ou circonstancielles juste après le pronom (ex. « ... le symbole qui, d'après l'auteur, représentait ... » ou « ... le symbole qui, dit l'auteur, représentait ... »).

Connecteur sque C'est le connecteur constitué avec le pronom « que ».

$$\text{sque} \rightarrow [\text{open\_g}], \text{sque}. \tag{B.11}$$
$$\text{sque} \rightarrow [\text{que}], (\text{cmpinc} \mid \text{subdt}). \tag{B.12}$$
$$\text{sque} \rightarrow [\text{que}]. \tag{B.13}$$

Connecteur scamb C'est le connecteur créé à partir des mots camb – à savoir « quand », « comme » et « si ».

$$\text{scamb} \rightarrow [\text{open\_g}], \text{scamb}. \tag{B.14}$$
$$\text{scamb} \rightarrow [\text{p}], [\text{camb}], (\text{cmpinc} \mid \text{subinc}). \tag{B.15}$$
$$\text{scamb} \rightarrow [\text{camb}], (\text{cmpinc} \mid \text{subinc}). \tag{B.16}$$
$$\text{scamb} \rightarrow [\text{p}], [\text{camb}]. \tag{B.17}$$
$$\text{scamb} \rightarrow [\text{camb}]. \tag{B.18}$$

Les règles B.15 et B.17 définissent les connecteurs constitués d'un camb précédé d'une proposition. Ce sont par exemple « d'où » ou « pour quand », etc.

Connecteur sip C'est le connecteur créé à partir de mots ip. Les mots étiquetés ip sont « pourquoi », « comment », « combien », « quel(les) », ainsi que les syntagmes que ces deux derniers constituent avec le syntagme prépositionnel ou nominal qui les suit, tels que « combien de temps », « quelle chance ». Le pronom « qui »

précédé par une préposition constitue également un connecteur sip.

sip → [open_g], sip. (B.19)
sip → [p], [ip], (cmpinc | subdt). (B.20)
sip → [p], [qui], (cmpinc | subdt). (B.21)
sip → [p], [où], (cmpinc | subdt). (B.22)
sip → [ip], (cmpinc | subdt). (B.23)
sip → [où], (cmpinc | subdt). (B.24)
sip → [p], [ip]. (B.25)
sip → [p], [qui]. (B.26)
sip → [p], [où]. (B.27)
sip → [ip]. (B.28)
sip → [où]. (B.29)

Les règles B.15 et B.18 définissent, comme nous l'avons vu pour les règles de camb, les connecteurs constitués d'un ip précédé par une proposition. Ce sont par exemple « par quel (droit) » ou « pour combien », etc.

Connecteur scs C'est le connecteur créé à partir de mots cs. Les cs sont des locutions conjonctives, appelées traditionnellement « conjonction de subordination », telles que « alors que », « tandis que », « pourvu que », « d'autant que », « parce que », etc.

scs → [open_g], scs. (B.30)
scs → [cs], (cmpinc | subinc). (B.31)
scs → [cs]. (B.32)

Connecteur srel C'est le connecteur créé à partir de mots renommés par la grammaire srel0. rel0 sont les mots rel – « lequel » (et ses formes fléchies) et « quoi » – éventuellement précédés par une préposition, « que », « dont » ainsi que « qui »

précédé par une préposition (règles B.36 à B.41).

srel → [open_g], srel. (B.33)
srel → srel0, (cmpinc | subdt). (B.34)
srel → srel0. (B.35)
srel0 → [p], [rel]. (B.36)
srel0 → [p], [qui]. (B.37)
srel0 → [p], [où]. (B.38)
srel0 → [rel]. (B.39)
srel0 → [que]. (B.40)
srel0 → [dont]. (B.41)
srel0 → [où]. (B.42)

Connecteur sep1 Les connecteurs scamb et scs sont également nommés sep1.

sep1 → scamb. (B.43)
sep1 → scs. (B.44)

Connecteur sep2 Le connecteur sep2 regroupe les connecteurs permettant de réaliser la structure de la coordination et il est créé à partir d'une conjonction de coordination cc ou une virgule.

sep2 → [vrg], sep2. (B.45)
sep2 → [cc], [advp]. (B.46)
sep2 → [vrg], [advp]. (B.47)
sep2 → [cc], (cmpinc | subdt). (B.48)
sep2 → [cc]. (B.49)
sep2 → [vrg]. (B.50)

Connecteur sepamb Le connecteur sepamb regroupe les connecteurs capables d'introduire non seulement une proposition mais aussi un syntagme[2].

sepamb → [p], [camb]. (B.51)
sepamb → [p], [ip]. (B.52)
sepamb → [p], [rel]. (B.53)
sepamb → [p], [qui]. (B.54)
sepamb → [p], [où]. (B.55)
sepamb → [camb]. (B.56)
sepamb → [ip]. (B.57)
sepamb → [rel]. (B.58)
sepamb → [qui]. (B.59)
sepamb → [dont]. (B.60)
sepamb → [où]. (B.61)

B.4 Définition des sous-phrases

La définition des sous-phrases se base, en plus des études sur la typologie des connecteurs dont nous venons de parler, sur celles de la typologie des subordonnées vues dans la section 4.6, que nous rappelons ci-dessous.

B.4.1 Typologie des propositions

Nous avons réalisé une classification des propositions selon leur position dans la phrase, et nous avons distingué quatre types de subordonnées selon leur position – donc la fonction qu'elles jouent – dans la phrase :

1. position post-verbale :
   proposition de complément, que nous étiquetons subQ ;
2. positions initiale et finale :
   proposition accessoire (ou périphérique) étiquetée subP ;
3. position post-nominale : proposition secondaire étiquetée subR ;
4. position pré-verbale :
   proposition sujet.

B.4.2 Règles des sous-phrases

Trois types de propositions non autonomes (sous-phrases) sont définis : coordonnée, subordonnée et détachée.

[2]Comme nous l'avons déjà précisé dans la section 9.2.4, on entend ici par « proposition » et « syntagme », des unités purement de surface.

Coordonnées La proposition coordonnée est définie comme une proposition précédée par un connecteur de coordination sep2, éventuellement suivie d'une autre proposition coordonnée.

Elle apparaît à une position définie par la règle B.4.

$$\text{coordonnee} \rightarrow \text{sep2, proposition, coordonnee ?.} \quad \text{(B.62)}$$

Subordonnées Les subordonnées se distinguent elles-mêmes en trois types : subordonnée de complément subQ, subordonnée périphérique (accessoire) subP et subordonnée secondaire subR.

$$\begin{aligned}
\text{subQ} &\rightarrow \text{(sque | sip | scamb), proposition, subQcrd ?.} && \text{(B.63)}\\
\text{subQ} &\rightarrow \text{squi, propss, subQcrd ?.} && \text{(B.64)}\\
\text{subQcrd} &\rightarrow \text{sep2, subQ.} && \text{(B.65)}\\
\text{subP} &\rightarrow \text{(sep1 | sque), proposition, subPcrd ?.} && \text{(B.66)}\\
\text{subPcrd} &\rightarrow \text{sep2, subP.} && \text{(B.67)}\\
\text{subR} &\rightarrow \text{[vrg] ?, squi, propss, subRcrd ?.} && \text{(B.68)}\\
\text{subR} &\rightarrow \text{[vrg] ?, srel, proposition, subRcrd ?.} && \text{(B.69)}\\
\text{subRcrd} &\rightarrow \text{sep2, subR.} && \text{(B.70)}\\
\text{sub} &\rightarrow \text{squi, propss, subRcrd ?.} && \text{(B.71)}\\
\text{sub} &\rightarrow \text{(srel | scs), proposition, sub0crd ?.} && \text{(B.72)}\\
\text{sub0crd} &\rightarrow \text{sep2, subR, subRcrd ?.} && \text{(B.73)}
\end{aligned}$$

Chaque subordonnée est caractérisée par ses connecteurs. La subR est définie comme suivant un syntagme nominal et la subP et la subQ sont définies dans les règles du prédicat que nous présentons un peu plus loin.

Les règles subRcrd, subPcrd et subQcrd sont destinées à traiter les structures de coordination récursives, c'est-à-dire les structures comportant plusieurs sous-phrases coordonnées.

Toutes les subordonnées, sauf celles commençant par un connecteur ip, sont également appelées sous le nom de sub (règles B.71, B.72 et B.73).

Pour toutes ces subordonnées, est définie une règle permettant de les traiter lorsqu'elles sont détachées par un tiret en fin de phrase.

Détachées Ces dernières se divisent elles-mêmes en trois classes : incidente subinc, détachée subdt et relative détachée subdtR.

subinc → [vrg], propinc, [vrg]. (B.74)
subinc → tiret, proposition, tiret. (B.75)
subinc → p_ouv, proposition, p_ferm. (B.76)
subdt → [vrg], (sub | propinc), [vrg]. (B.77)
subdt → tiret, (sub | proposition), tiret. (B.78)
subdt → p_ouv, (sub | proposition), p_ferm. (B.79)
subdtR → [vrg], (squi, propss | srel, proposition), [vrg]. (B.80)
subdtR → tiret, (squi, propss | srel, proposition), tiret. (B.81)
subdtR → p_ouv, (squi, propss | srel, proposition), p_ferm. (B.82)

La proposition détachée est celle entourée de deux symboles de ponctuation de même type, à l'exception des guillemets qui ont vraisemblablement un rôle différent des autres. En effet, les symboles tels que les parenthèses ou les virgules enchâssent et insèrent dans une phrase des éléments plus ou moins périphériques, alors que les guillemets servent à souligner des constituants souvent primaires de la phrase. Aussi, dans notre grammaire, les guillemets apparaissent-ils dans les règles définissant les constituants primaires de la phrase.

Les propositions détachées subdt sont un type général désignant les propositions entourées de deux séparateurs, sauf les relatives subdtR qui ont plus de contraintes quant aux positions où elles apparaissent. Elles concernent les propositions que nous appelons détachées-insérées ainsi qu'une partie de nos subordonnées périphériques, détachées et insérées par deux séparateurs.

Les incidentes subinc correspondent à nos propositions détachées-insérées. Elles sont définies comme apparaissant non seulement aux mêmes positions que celles où apparaissent les subdt, mais aussi après les connecteurs scamb et scs.

B.5 Définition de la proposition

Compte tenu de la définition linguistique adoptée, la proposition est définie comme suit :

proposition → [open-g], proposition. (B.83)
proposition → prop0. (B.84)
proposition → (subP | cmp), [vrg] ?, proposition. (B.85)
proposition → sujet, [vrg], predicat. (B.86)
proposition → sn, prop0. (B.87)
proposition → cmp, (prop0 | propss). (B.88)
proposition → propss. (B.89)
proposition → sn, [vrg], prop0. (B.90)
prop0 → ([clsj] | sujet), predicat, ([vrg], subP) ?. (B.91)
propss → sv, (subQ | cmp | sn) ?, ([vrg], subP) ?. (B.92)
propss → sv, cmp, sn, ([vrg], subP) ?. (B.93)
propinc → v, (cmp | sn) ?. (B.94)

La règle B.84 définit la forme la plus basique de la proposition. La règle B.85 est dédiée au traitement des propositions avec une subordonnée ou un cmp détaché en tête.

La règle B.86 traite les propositions dont le sujet est détaché. Les règles B.87 et B.88 sont dédiées au traitement des propositions précédées par un syntagme, respectivement, nominal et non-nominal. Le syntagme sn ne peut être suivi que d'une proposition canonique prop0, mais cmp peut également être suivi d'une proposition sans sn en tête propss.

La règle B.89 définit que la proposition peut être sans syntagme nomimal en tête propss.

Enfin, la règle B.90 traite les propositions précédées par un syntagme nominal détaché. Le syntagme cmp détaché en tête est traité par la règle B.85. La règle traitant le syntagme sn détaché en tête est définie à part et mise à la fin de l'ensemble des règles afin d'éviter les erreurs que cette règle peut entraîner, à savoir l'interprétation erronée de la phrase constituée d'un sujet détaché et d'un prédicat, comme étant constituée d'un sn détaché et d'une proposition sans sn en tête.

La règle B.91 définit la proposition canonique prop0 constituée d'un sujet ou un clitique sujet et d'un prédicat.

Les règles B.92 et B.93 définissent la proposition sans syntagme nominal préverbal. Ces règles traitent non seulement les phrases à verbe impératif, mais aussi les propositions coordonnées dont le sujet commun avec la racine est omis, ainsi que les propositions dans lesquelles le sujet est post-posé par rapport au verbe.

La règle B.94 définit la proposition apparaissant en tant qu'incidente propinc constituée d'un verbe éventuellement suivi d'un cmp ou d'un sn.

B.6 Définitions du sujet et du prédicat

Le prédicat est repéré par la présence d'un verbe fini et il est défini avec les trois éléments primaires comme suit :

$$\text{sujet} \rightarrow \text{sn}. \tag{B.95}$$
$$\text{predicat} \rightarrow \text{pred0, subP}\,?. \tag{B.96}$$
$$\text{pred0} \rightarrow \text{sv, (subQ} \mid \text{sn} \mid \text{cmp)}\,?. \tag{B.97}$$
$$\text{pred0} \rightarrow \text{sv, cmp, (subQ} \mid \text{sn)}\,?. \tag{B.98}$$

La règle B.95 définit que le sujet est soit un sn soit un syntagme infinitivale svinf.

La règle B.96 définit que le prédicat peut être suivi d'une subordonnée périphérique. La règle B.97 est définie pour les prédicats constitués d'un verbe suivi directement du complément direct et la règle B.98, pour ceux dont le verbe est suivi d'un complément indirect ou accessoire.

À noter qu'on ne définit pas spécifiquement la règle :

$$\text{pred0} \rightarrow \text{sv, sn, cmp}.$$

pourtant très élémentaire. En effet, sn est défini comme susceptible d'être suivi par cmp et nous traitons les structures « SV - COD - COI » par la règle B.97 predicat → sv, sn sans faire de distinction entre les COI et les compléments secondaires suivis d'un sn. Ce choix, difficilement justifiable linguistiquement, a tout de même été retenu car son influence sur la détection des propositions est considérée comme minime, voire nulle.

Les règles B.97 et B.98 sont définies pour les subordonnées complétives ou percontatives suivant directement le verbe principal ou éventuellement un cmp suivi du verbe. Différentes des autres subordonnées (relatives ou incidentes), elles sont rarement suivies par des compléments du verbe principal et aucune règle traitant ce type de structure n'est définie à l'heure actuelle.

B.7 Définition du syntagme verbal

Le syntagme verbal est défini en trois temps : v0, v et sv.

$$\text{v0} \rightarrow \text{[vfin], [clsj]}. \tag{B.99}$$
$$\text{v0} \rightarrow \text{[vfin], [trait], [clsj]}. \tag{B.100}$$
$$\text{v0} \rightarrow \text{[vfin], (cmpinc} \mid \text{subdt)}\,?. \tag{B.101}$$

Les règles B.99 et B.100 traitent le verbe suivi d'un clitique sujet[3]. Cette forme apparaît non seulement dans la proposition sans sn en tête mais aussi dans la pro-

[3]Ces deux règles sont définies séparément, uniquement du fait qu'il arrive parfois que le *tokenizer* utilisé sépare le tiret du clitique et parfois non.

position canonique (c'est-à-dire avec un sujet) telle que « Les espérances de démocratisation des Balkans ont-elles fait long feu ? ».

La règle B.101 traite le verbe fini suivi d'un syntagme ou d'une proposition détachée enchâssée.

$$\text{v} \rightarrow \text{cltq, v.} \tag{B.102}$$
$$\text{v} \rightarrow \text{v0, [advp]?, [vpt].} \tag{B.103}$$
$$\text{v} \rightarrow \text{v0, [advp].} \tag{B.104}$$
$$\text{v} \rightarrow \text{v0.} \tag{B.105}$$

La dernière règle B.105 définit le verbe v le plus basique constitué d'un v0. Les règles B.103 et B.104 définissent qu'il peut être suivi d'un adverbe ou d'un participe passé éventuellement précédé par un adverbe. Enfin, la première règle B.102 définit que le v ainsi constitué peut être précédé par une séquence cltq que nous allons présenter un peu plus bas.

$$\text{sv} \rightarrow \text{[open\_g], sv.} \tag{B.106}$$
$$\text{sv} \rightarrow \text{v, (cmpinc | subdt)?.} \tag{B.107}$$

Le syntagme verbal basique est défini par la règle B.107 : il est constitué d'un v éventuellement suivi d'un syntagme ou d'une proposition détachée enchâssée. La règle B.106 définit que le v peut être précédé par des guillemets ouvrants.

B.8 Définition du clitique

Il s'agit d'un ensemble constitué de clitiques compléments et « ne » de négation apparaissant dans le contexte gauche du verbe.

$$\text{cltq} \rightarrow \text{[ne], cltq0?.} \tag{B.108}$$
$$\text{cltq} \rightarrow \text{cltq0.} \tag{B.109}$$
$$\text{cltq0} \rightarrow \text{[clns], cltq0?.} \tag{B.110}$$

Les règles B.108 et B.109 définissent que le cltq est constitué, soit d'un « ne » de négation éventuellement suivi d'un cltq0, soit d'un cltq0.

Le cltq0 est défini par la règle B.110 comme constitué d'un clitique complément éventuellement suivi d'autres clitiques compléments.

B.9 Définition du syntagme infinitival

Le syntagme infinitival est, comme le sv, défini en trois temps : vinf00, vinf0 et svinf.

vinf00 → [vfin], (cmpinc | subdt) ?. (B.111)

Le vinf00 est défini comme constitué d'un verbe fini éventuellement suivi d'un syntagme ou d'une proposition détachée enchâssée.

vinf0 → ([advp] | cltq) ?, vinf0. (B.112)
vinf0 → vinf00. (B.113)

Le vinf0 est défini comme constitué d'un vinf00 éventuellement précédé par un ou plusieurs clitiques ou adverbes.

svinf → [open_g], svinf. (B.114)
svinf → vinf0, (subq | cmp | sn | svinf), (crdsvinf | subP) ?. (B.115)
svinf → vinf0, cmp, sn, (crdsvinf | subP) ?. (B.116)
svinf → vinf0, (crdsvinf | subP) ?. (B.117)
crdsvinf → sep2, svinf. (B.118)

Le svinf est défini comme constitué d'un vinf0 (règle B.117) et éventuellement d'un ou plusieurs compléments (règles B.115, B.116), éventuellement suivi d'un ou plusieurs autres syntagmes coordonnés ou de subordonnées périphériques. La règle B.114 définit que le svinf peut être précédé par des guillemets ouvrants. Le syntagme coordonné crdsvinf est constitué d'un svinf précédé par un connecteur de coordination.

B.10 Définition du syntagme participial

À la sortie du *chunker*, le *chunk* constitué d'un verbe au participe présent précédé par une préposition « en » est étiqueté vger, et le verbe au participe présent seul est étiqueté vptpr.

Les règles pour le syntagme participial sont destinées à constituer à partir de ces éléments les syntagmes noyaux auxquels seront rattachés les compléments. Il existe trois types de règles définissant chacun vptpr00, vptpr0 et vger.

vptpr00 → [vptpr], (cmpinc | subdt) ?. (B.119)

Le vptpr00 est défini comme constitué d'un verbe au participe présent éventuellement suivi d'un syntagme ou d'une proposition détachée enchâssée.

vptpr0 → ([advp] | cltq) ?, vptpr0. (B.120)
vptpr0 → vptpr00. (B.121)

Le vptpr0 est défini comme constitué d'un vptpr00 éventuellement suivi d'un ou plusieurs clitiques ou adverbes.

vgr → [advp] ?, [p], vptpr0. (B.122)
vgr → vptpr0. (B.123)
vgr → [advp] ?, [vger]. (B.124)

Le vgr est défini comme constitué d'un vgr éventuellement précédé par un adverbe (règle B.124) ou d'un vptpr0 (règle B.123), ou encore d'un vptpr0 précédé par une préposition (règle B.122), éventuellement précédé par un adverbe.

B.11 Définition du syntagme infinitival prépositionnel

Le syntagme infinitival prépositionnel est constitué à partir soit d'un syntagme infinitival précédé par une préposition, soit d'un syntagme participial vgr présenté dans le paragraphe précédent.

Il existe deux types de règles : ppv et svprep.

ppv → [advp] ?, [p], vinf0. (B.125)
ppv → [advp] ?, [ppvinf]. (B.126)

La règle B.125 sert à constituer un syntagme noyau à partir d'un syntagme infinitival et d'une préposition qui le précède, que le *chunker* n'a pas réussi à regrouper.

La règle B.126 définit que le ppv peut être constitué d'un syntagme infinitival prépositionnel composé par le *chunker* éventuellement précédé par un adverbe.

svprep → [open_g], svprep. (B.127)
svprep → svprep0, (subq | cmp | sn), (crdsvinf | crdsvprep | subP) ?. (B.128)
svprep → svprep0, cmp, sn, (crdsvinf | crdsvprep | subP) ?. (B.129)
svprep → svprep0, (crdsvinf | crdsvprep | subP) ?. (B.130)
svprep0 → (ppv | vgr), (cmpinc | subdt) ?. (B.131)
crdsvprep → sep2, svprep. (B.132)

svprep est constitué d'un syntagme noyau svprep0 et d'un ou plusieurs compléments (règles B.129, B.128 et B.130), éventuellement suivi d'un ou plusieurs autres syntagmes coordonnés.

Le svprep0 est défini comme constitué d'un vgr ou ppv (règle B.131) et éventuellement suivi d'un syntagme ou d'une proposition détachée enchâssée. La règle B.127 définit que le svprep peut être précédé par des guillemets ouvrants.

Le syntagme coordonné crdsvprep est constitué d'un svprep précédé par un connecteur de coordination.

B.12 Définition du sn

sn → [d] ?, [open_g], sn. (B.133)
sn → ([d] | [adj]) , sn. (B.134)
sn → sn0, (subR | subdtR | cmp | cmpcrd) ?. (B.135)
sn → sn0, cmp, (subR | subdtR) ?. (B.136)
sn → sn0, (cmpinc | subdt), (subR | subdtR | cmp | cmpcrd) ?. (B.137)
sn → svinf. (B.138)
sn0 → [np], [trait], [np]. (B.139)
sn0 → [np], snprop ?. (B.140)
sn0 → [pro]. (B.141)
sn0 → snprop. (B.142)
snprop → [nprop], snprop. (B.143)

La règle B.133 définit que le sn peut être précédé par des guillemets ouvrants éventuellement précédé eux-mêmes par un déterminant. La règle B.134 définit que le sn peut être précédé par un déterminant ou un adjectif.

La règle B.135 traite le cas où le sn0 est directement suivi d'une relative ou celui où il constitue tout seul un sn. La règle B.136 définit que le sn est constitué d'un syntagme nominal basique sn0 suivi d'un ou plusieurs compléments et éventuellement d'une relative. La règle B.137 définit que le sn est constitué d'un syntagme nominal basique sn0 suivi d'un syntagme ou d'une proposition détachée enchâssée et éventuellement d'une relative ou d'autres compléments.

Le syntagme nominal basique sn0 est défini comme constitué soit de deux np reliés par un trait (règle B.139), soit d'un np éventuellement suivi d'un snprop (règle B.140), soit d'un pro (règle B.141), soit d'un snprop (règle B.142) qui est constitué d'un nom propre suivi d'un ou plusieurs autres noms propres (règle B.143).

B.13 Définition du cmp

Nous distinguons trois types de compléments cmp : complément détaché cmpinc, complément coordonné cmpcrd et complément cmp.

Complément détaché cmpinc

cmpinc → [vrg], [advp] ?, (sn | cmp | cmpcrd), [vrg]. (B.144)
cmpinc → [vrg], [np], [np], [vrg]. (B.145)
cmpinc → [tiret], [advp] ?, (sn | cmp | cmpcrd), [tiret]. (B.146)
cmpinc → [open_p], [advp] ?, (sn | cmp | cmpcrd), [close_p]. (B.147)
cmpinc → [open_p], nonprop, [close_p]. (B.148)

Le complément détaché est un syntagme nominal ou un cmp ou un cmpcrd entouré de virgules ou de parenthèses ou de tirets (règles B.144, B.146 et B.147). La

règle B.145 permet de traiter la séquence de deux np telle que des appositions et la règle B.148 permet d'accepter, quels que soient ses constituants, la séquence entourée d'une paire de parenthèses.

Complément coordonné cmpcrd

cmpcrd → sep2, (sn | cmp | cmpcrd). (B.149)
cmpcrd → [tiret], (sn | cmp | cmpcrd). (B.150)
cmpcrd → [tiret], (sn | cmp), cmpcrd. (B.151)

La règle B.149 définit que le complément coordonné est un sn ou un cmp, ou un cmpcrd, précédé par un connecteur de coordination.

Les règles B.150 et B.151 traitent le complément détaché en fin de phrase[4].

Définition du cmp

cmp → [open_g], cmp. (B.152)
cmp → sepamb, (sn | cmp). (B.153)
cmp → [p], (sn | cmp). (B.154)
cmp → [d], [advp], cmp. (B.155)
cmp → [d], [advp]. (B.156)
cmp → [vrg] ?, svprep. (B.157)
cmp → cmp0, cmpcrd ?. (B.158)
cmp0 → [vrg] ?, atr, cmp ?. (B.159)
cmp0 → [vrg] ?, atr, (sundt | cmpinc), cmp ?. (B.160)
cmp0 → [vrg] ?, [pp], (sundt | cmpinc), (cmp | subR | subdtR) ?. (B.161)
cmp0 → [vrg] ?, [pp], cmp, (subR | subdtR) ?. (B.162)
cmp0 → [vrg] ?, [pp], (subR | subdtR) ?. (B.163)

La règle B.152 définit que le cmp peut être précédé par des guillemets ouvrants.

La règle B.153 définit le cmp introduit par un connecteur ambigu, comme par exemple, « (quatre personnes) dont notre directeur », « (plus intelligente) que belle », « où aller ».

La règle B.154 définit le syntagme prépositionnel qui n'a pas été regroupé par le *chunker*. La règle B.155 traite les séquences telles que « (femme) la [d] plus [advp] considérée [cmp] ». La règle B.156 traite les séquences telles que « (mange) le [d] plus [advp] ».

La règle B.157 définit qu'un svprep, éventuellement précédé par une virgule, peut constituer un cmp. La règle B.158 définit que le cmp peut être constitué d'un cmp0 éventuellement suivi d'un syntagme coordonné cmpcrd.

[4]Nous n'avons pas de moyen de représenter la condition « en fin de phrase » dans ce formalisme. Dans la réalisation, ces règles sont transformées en clauses Prolog, permettant ainsi d'introduire la contrainte qui interdit la structure récursive, c'est-à-dire l'apparition d'un autre complément détaché par un tiret à l'intérieur d'un complément détaché par un tiret.

Les règles cmp0 définissent le syntagme complément basique permettant de regrouper de manière récursive différents éléments pour constituer un cmp. Le cmp0 est constitué d'un syntagme prépositionnel pp ou d'un élément atr. Ce dernier est défini comme l'élément qui n'appartient pas aux catégories suivantes : verbe (sauf participe passé), np, pronom, clitique, déterminant, préposition, ne, pp, séparateur ou connecteur.

Le fait de distinguer la catégorie pp de la classe atr permet d'attacher les relatives uniquement au syntagme prépositionnel, et non pas à d'autres éléments tels qu'un adverbe ou un adjectif (règles B.161, B.162 et B.163).

ANNEXE C

ANNEXE : SIGLÉ

C.1 Règles pour la correction des erreurs d'étiquetage (module postTagging)

Mot « ne »

La figure C.1 (voir page suivante) représente différentes combinaisons du mot « ne » avec son contexte droit. Les éléments sur fond gris reliés avec une ligne continue sont des unités susceptibles d'apparaître à cette position (en l'occurrence dans le contexte droite du mot « ne »). Les autres sont des éléments qui ne peuvent pas apparaître à cette position et s'ils y sont, cela correspond à une erreur d'étiquetage. La ligne discontinue indique la nécessité d'une correction de l'étiquette de l'élément en aval, en la nouvelle catégorie indiquée sur la flèche.

Le mot « ne » n'autorise dans son contexte droit que les clitiques compléments, les verbes ou les adjectifs. Tout autre élément est donc un objet de correction.

* Tout étiquetage est considéré comme fiable : score = 100 ;

1. Si le contexte droit est le mot « en » étiqueté comme préposition, un clitique sujet, un déterminant ou un pronom, alors il est modifié en clitique complément ;
2. Sinon c'est un verbe :
   a) si le mot du contexte droit se termine par « -er » ou par « -ir », on l'étiquette verbe à l'infinitif[1] ;

[1] Il faudrait une analyse beaucoup plus détaillée que nous n'avons pas développée pour ce petit module, que nous considérons comme un outil de premier secours et qui est hors de notre sujet principal. Néanmoins, si nous décidions, dans des travaux futurs, de développer un outil plus per-

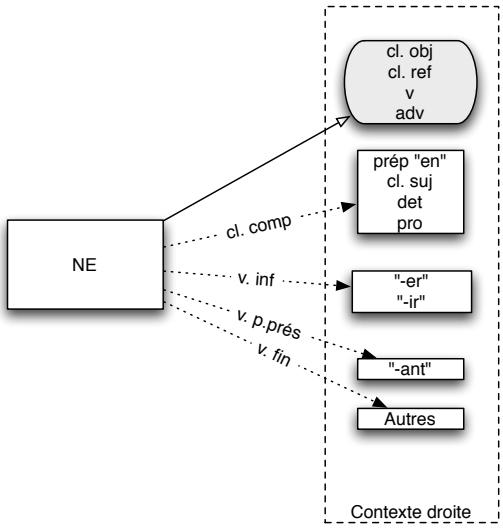


FIG. C.1 – Mot « ne » et son contexte droit

b) si le mot du contexte droit termine par « -ant », on l'étiquette verbe au participe présent ;

c) sinon, on l'étiquette verbe fini.

Par ailleurs, s'il est précédé par une préposition, on lui attribue un score 200, ce qui impose à son contexte droit un verbe à la forme infinitive.

Préposition « en »

Si une préposition « en » est précédée par un verbe sauf à la forme participe présent ou passé, elle est modifiée en clitique objet.
ex. (les marchands...) en [p → cl-obj] ont [v-fin] (commandé par centaines.)
ex. (la population rurale) en [prép → cl-obj] est [v-fin] (restée)

Clitique sujet 1/2 : avec son contexte droit

La figure C.2 page ci-contre représente différentes combinaisons des clitiques sujets avec leur contexte droit.

1. Si le contexte droit est, soit un clitique complément, soit le mot « ne », soit un verbe fini, alors aucune modification n'est nécessaire ;
2. Sinon une modification s'impose :
   a) pour les clitiques sujets non ambigus, c'est-à-dire sauf « je » « tu » « il » « ils » ou celui commençant par un tiret (forme pour l'inversion de sujet),

formant, nous aurions déjà l'essentiel des travaux basiques, linguistiques et algorithmiques, réalisés dans le cadre de précédents travaux (Nakamura-Delloye, 2002).

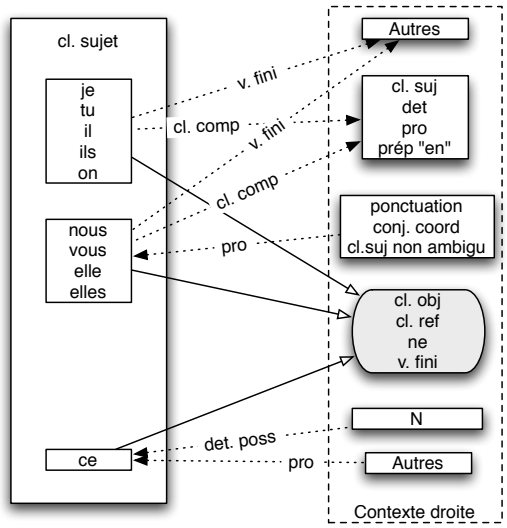


FIG. C.2 – Clitiques sujets et leur contexte droit

* score = 100 ;

i. S'il apparaît suivi d'un clitique sujet, d'un déterminant, d'un pronom ou d'une préposition « en », ce second est un clitique complément
ex. on [cl-suj] nous [cl-suj → cl-comp] (inculque dès l' enfance)

ii. sinon le mot du contexte droit est étiqueté comme verbe fini ;
ex. ils [cl-suj] rognent [v-pt → v-fin] (peu à peu un contrat social ...),
ex. je [cl-suj] change [n → v-fin]
ex. je [cl-suj] restais [n → v-fin] (à écouter ...)

b) pour les clitiques ambigus :

i. S'il apparaît suivi d'un connecteur de subordination tel que « que » ou un point final, il doit être étiqueté comme pronom
ex. (chez) nous [cl-suj → pro] . [ponct-s] ;

ii. S'il apparaît suivi d'un clitique sujet, d'un déterminant, d'un pronom ou d'une préposition « en », ce second est un clitique complément (score = 100) :
ex. elle [cl-suj] le [d-def → cl-comp] (fait avec un masque gris) ;

iii. Sinon, il doit être étiqueté comme un verbe fini
ex. (qu') elle [cl-suj] relègue [v-pt → v-fin] (au second plan ...)

c) pour « ce » :

i. s'il est suivi par un nom, il est modifié en déterminant possessif ;

ii. sinon
ex. ce [cl-suj → pro] que [camb] (la Yougoslavie n' a pas su faire).

Clitique sujet 2/2 : avec son contexte gauche

La figure C.3 représente différentes combinaisons des clitiques sujets avec leur contexte gauche.

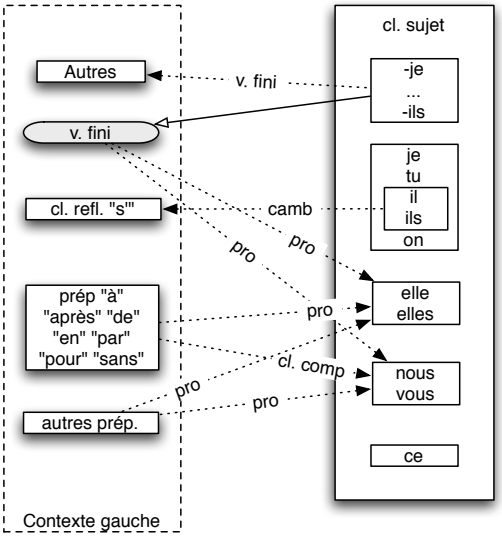


FIG. C.3 – Clitiques sujets et leur contexte gauche

1. pour un clitique sujet commençant par un tiret (forme pour l'inversion du sujet) :
   a) s'il est précédé par un verbe fini, alors aucune modification n'est nécessaire ;
   b) sinon l'élément du contexte gauche doit être étiqueté comme un verbe fini
   ex. déplorait [n → v-fin] -il [cl-suj] ;
2. pour « il » et « ils » :
   a) s'il est précédé par « s' » étiqueté comme clitique complément, ce dernier est modifiée en connecteur ambigu :
   ex. s' [cl-refl → camb] il [cl-suj] (est)
3. pour « elle » et « elles » :
   a) s'il est précédé par un verbe ou par une préposition, le mot étiqueté comme clitique sujet est un pronom :
   ex. (qui) témoigne [v-fin] elle [cl-suj → pro] (aussi)

4. pour « nous » et « vous »
   a) s'il est précédé par un verbe, le mot étiqueté comme clitique sujet est un pronom
      ex. (qui) témoigne [v-fin] elle [cl-suj → pro] (aussi) ;
   b) s'il est précédé par une certaine préposition (à, après, de, en, par, pour, sans), son étiquette est modifiée en clitique complément ;

   c) s'il est précédé par une autre préposition, son étiquette est modifiée en pronom.

Clitique complément

La figure C.4 représente différentes combinaisons des clitiques compléments avec leur contexte droit.

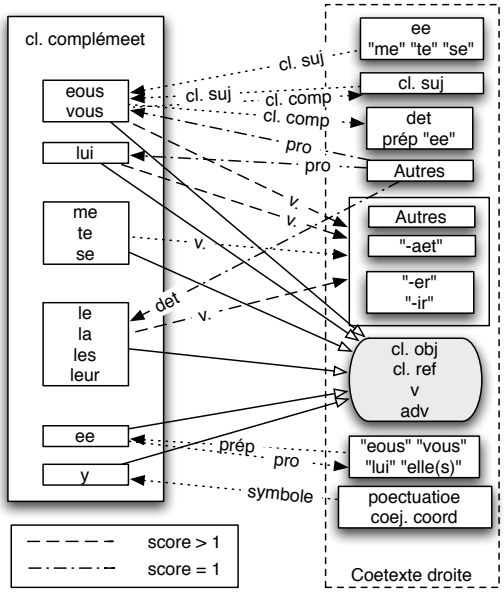


FIG. C.4 – Clitiques compléments et leur contexte droit

1. Si le contexte droit est soit un clitique complément, soit un verbe fini, soit un adverbe, alors aucune modification n'est nécessaire ;
2. Sinon une modification s'impose :
   a) pour « nous » et « vous » :
      i. s'il est suivi d'un « ne » ou d'un clitique « me » « te » « se », il est modifié en clitique sujet (score = 100) :
         ex. nous [cl-obj → cl-suj] ne [ne] (sommes pas),
         ex. nous [cl-obj → cl-suj] n' [ne] (aurions peut-être pas) ;

ii. s'il est suivi d'un déterminant ou d'une préposition « en », ce dernier est modifié en clitique complément ;

iii. s'il est suivi d'un clitique sujet, il est modifié en clitique sujet et l'élément suivant en clitique complément ;

iv. sinon une modification selon son score :
- si son score > 1 (étiquette fiable), alors l'élément suivant est modifié en verbe ;
- sinon (son score = 1, valeur par défaut), il est modifié en pronom :
  ex. (pour) nous [cl-obj → pro] de [préposition] (le dire),
  ex. (vis-à-vis de) nous [cl-obj → pro] comme [camb] (une puissance coloniale),
  ex. (c'est) nous [cl-obj → pro] » [close-g].

b) pour « lui » :
- une modification selon son score :

i. si son score > 1 (étiquette fiable), alors l'élément suivant est modifié en verbe ;

ii. sinon (son score = 1, valeur par défaut), il est modifié en pronom :
  ex. (avant) lui [cl-obj → pro]) [open-p],
  ex. (« chez) lui [cl-obj → pro] » [open-g] ;

c) pour « me », « te », « se », l'élément suivant est modifié en verbe :
ex. (n') en [cl-obj] compte [n → v-fin] (plus que deux),
ex. (des cafés branchés et des publicités commerciales) se [cl-ref] parent [n → v-fin] (également de symboles soviétiques) ;

d) pour « le », « la », « les » ou « leur » :
- une modification selon son score :

i. si son score > 1 (étiquette fiable), alors l'élément suivant est modifié en verbe ;

ii. sinon (son score = 1, valeur par défaut), il est modifié en déterminant :
  ex. les [cl-obj → v-fin] entreprises [n],
  ex. leur [cl-obj → v-fin] projet [n] ;
  Mais, cette règle risque d'entraîner une modification erronée, car le mot étiqueté comme nom, mal analysé, peut en fait être un verbe. Dans ce cas, d'autres analyses morphologiques plus fines – telles que l'examen de la présence ou non d'une terminaison verbal – seraient nécessaires, mais nous n'entrons pas dans ce niveau d'analyse dans le cadre de cette thèse.

e) si un « en » étiqueté comme clitique complément apparaît suivi d'un pronom ambigu (nous, vous, lui, elle, elles), il est modifié en préposition et le mot suivant en pronom :

f) si un « y » étiqueté comme clitique complément apparaît suivi d'un symbole de ponctuation ou d'une conjonction de coordination, son étiquette est modifiée en symbole.

Pronom

La figure C.5 représente différentes combinaisons des pronoms avec leur contexte droit.

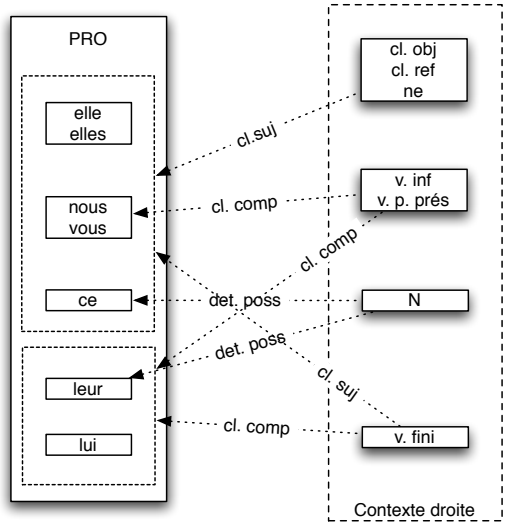


FIG. C.5 – Pronoms et leur contexte droit

1. pour « elle », « elles » :
   – s'il est suivi d'un clitique complément, d'un « ne » ou d'un verbe fini, alors il est modifié en clitique sujet ;
2. pour « nous », « vous » :
   a) s'il est suivi d'un clitique complément, d'un « ne » ou d'un verbe fini, alors il est modifié en clitique sujet ;
   b) s'il est suivi d'un verbe à l'infinitif ou d'un verbe au participe présent, alors il est modifié en clitique complément ;
3. pour « ce » :
   a) s'il est suivi d'un clitique complément, d'un « ne » ou d'un verbe fini, alors il est modifié en clitique sujet ;
   b) s'il est suivi d'un nom, alors il est modifié en déterminant démonstratif :
   ex. ce [pro → det] magma [n]
   Mais, il est également possible, bien que nous n'ayons pas pu trouver

d'exemple dans notre corpus, d'imaginer un pronom suivi d'un nom propre comme :
ex. (pour) cela [pro] Martine [n] (chercha...)

4. pour « leur » (étiqueté comme pronom possessif) :
   a) s'il est suivi d'un verbe à l'infinitif, d'un verbe au participe présent ou d'un verbe fini, alors il est modifié en clitique complément ;
   b) s'il est suivi d'un nom, alors il est modifié en déterminant possessif ;
5. pour « lui » :
   – s'il est suivi d'un verbe à l'infinitif, d'un verbe au participe présent ou d'un verbe fini, alors il est modifié en clitique complément.

Déterminant possessif

Si un déterminant possessif apparaît suivi d'un verbe, quelle que soit sa forme, le mot étiqueté comme verbe est en réalité un nom :
ex. ma [det-poss] destinée [v-pt → n]

Déterminant indéfini

1. Si un déterminant indéfini apparaît suivi d'un verbe, quelle que soit sa forme, le mot étiqueté comme verbe est en réalité un nom :
   ex. une [det-ind] lance [v-fin → n] ;
2. Si un déterminant indéfini apparaît suivi d'un clitique complément, il est modifié en pronom :
   ex. certains [det-ind → pro] se [cl-ref] (sont réfugiés au...)

Déterminant défini

1. avec son contexte droit :
   a) s'il apparaît suivi d'un verbe à la forme participe passé, le mot étiqueté comme verbe est en réalité un nom :
      ex. la [det-def] défunte [v-pt → n] ;
   b) s'il apparaît suivi d'un verbe, sauf à la forme participe passé, il est modifié en clitique complément :
      ex. (pour) le [det-def → cl] faire [v-inf] cuire [v-inf] (dans ...)
      Mais, dans certains cas, l'étiquetage comme déterminant est correct, nécessitant en revanche la correction de l'étiquette verbe en nom (ou pronom) comme :
      ex. sur la [det-def] tienne [v-pt → pro-poss] ;
   c) s'il apparaît suivi d'un clitique complément « lui », « y » ou « en », il est modifié en clitique objet ;
   d) s'il apparaît suivi d'un clitique complément « leur », ce dernier est modifié en pronom.

2. avec son contexte gauche :
   – pour « le », « les » :
     s'il est précédé par « de » ou par « à », alors il est modifié en clitique objet (score = 200).

C.2 Résultats du *chunking*

Le *chunking* des deux phrases brutes suivantes donne les résultats présentés ci-dessous.

Phrases entrées brutes :

1. Almona, qui était fort dévote, fit savoir le jour et l'heure où elle se jetterait dans le feu au son des tambours et des trompettes.
2. il le fit convenir qu'il fallait, si on pouvait, abolir un usage si barbare.

Sortie du *chunking* :

```
<s>
 <NP>
  <N-P-fs>
  Almona
  </N-P-fs>
 </NP>
 <VIRGULE>
 ,
 </VIRGULE>
 <PRO-rel>
 qui
 </PRO-rel>
 <VPfin>
  <V--I3s>
  était
  </V--I3s>
 </VPfin>
 <AdjP>
  <ADV>
  fort
  </ADV>
  <A-qual-fs>
  dévote
  </A-qual-fs>
 </AdjP>
 <VIRGULE>
 ,
 </VIRGULE>
 <VPfin>
  <V--J3s>
  fit
  </V--J3s>
 </VPfin>
 <VPinf>
  <V--W>
   savoir
  </V--W>
 </VPinf>
 <NP>
  <D-def-ms>
  le
  </D-def-ms>
  <N-C-ms>
  jour
   </N-C-ms>
 </NP>
 <CC>
 et
 </CC>
 <D-def-fs>
 l'
 </D-def-fs>
 <NP>
  <N-C-fs>
  heure
  </N-C-fs>
 </NP>
 <PRO-rel>
  <PRO-rel-3ms>
  où
  </PRO-rel-3ms>
 </PRO-rel>
 <CLsuj>
  <CL-suj-3fs>
  elle
  </CL-suj-3fs>
 </CLsuj>
 <CLref>
  <CL-refl-3fs>
  se
  </CL-refl-3fs>
 </CLref>
 <VPfin>
  <V--C3s>
  jetterait
  </V--C3s>
 </VPfin>
 <PP>
  <P>
  dans
  </P>
  <D-def-ms>
  le
  </D-def-ms>
  <N-C-ms>
  feu
  </N-C-ms>
 </PP>
 <PP>
  <P+D-ms>
  au
  </P+D-ms>
  <N-C-ms>
  son
  </N-C-ms>
 </PP>
 <PP>
  <P+D-mp>
  des
  </P+D-mp>
  <N-C-mp>
  tambours
  </N-C-mp>
 </PP>
 <CC>
 et
 </CC>
 <PP>
```

```
  <P+D-mp>
  des
  </P+D-mp>
  <N-C-mp>
  trompettes
  </N-C-mp>
 </PP>
 <PONCT-S>
 .
 </PONCT-S>
</s>

<s>
 <CLsuj>
  <CL-suj-3ms>
  il
  </CL-suj-3ms>
 </CLsuj>
 <CLobj>
  <CL-obj-3ms>
  le
  </CL-obj-3ms>
 </CLobj>
 <VPfin>
  <V--J3s>
  fit
  </V--J3s>
 </VPfin>
 <VPinf>
  <V--W>
  convenir
  </V--W>
 </VPinf>
 <Camb>
 qu'
 </Camb>
 <CLsuj>
  <CL-suj-3ms>
  il
  </CL-suj-3ms>
 </CLsuj>
 <VPfin>
  <V--I3s>
  fallait
  </V--I3s>
 </VPfin>
 <VIRGULE>
 ,
 </VIRGULE>
 <Camb>
 si
 </Camb>
 <CLsuj>
  <CL-suj-3ms>
  on
  </CL-suj-3ms>
 </CLsuj>
 <VPfin>
  <V--I3s>
  pouvait
  </V--I3s>
 </VPfin>
 <VIRGULE>
 ,
 </VIRGULE>
 <VPinf>
  <V--W>
  abolir
  </V--W>
 </VPinf>
 <NP>
  <D-ind-ms>
  un
  </D-ind-ms>
  <N-C-ms>
  usage
  </N-C-ms>
  <ADV>
  si
  </ADV>
  <A-qual-ms>
  barbare
  </A-qual-ms>
 </NP>
 <PONCT-S>
 .
 </PONCT-S>
</s>
```

C.3 Résultats du postChunking et du module chu2pl

Nous présentons ici le résultat fourni par le module postChunking pour les deux phrases d'exemple introduites dans la section précédente et la liste PROLOG produite par le module chu2pl.

Exemple 1 :

> Almona, qui était fort dévote, fit savoir le jour et l'heure où elle se jetterait dans le feu au son des tambours et des trompettes.

Sortie du module postChunking

```
<s>
 <chunk type = 'NP'>
 Almona
 </chunk>
 <chunk type = 'VIRGULE'>
 ,
 </chunk>
 <chunk type = 'C-S'>
 qui
 </chunk>
<chunk type = 'VP_fin'>
était
</chunk>
<chunk type = 'AdjP'>
fort
dévote
</chunk>
<chunk type = 'VIRGULE'>
,
</chunk>
<chunk type = 'VP_fin'>
fit
</chunk>
<chunk type = 'VP_inf'>
```

```
savoir
</chunk>
<chunk type = 'NP'>
le
jour
</chunk>
<chunk type = 'C-C'>
et
</chunk>
<chunk type = 'NP'>
l'
heure
</chunk>
<chunk type = 'C-S'>
où
</chunk>
<chunk type = 'CL_suj'>
elle
</chunk>
<chunk type = 'CL_ref'>
se
</chunk>
<chunk type = 'VP_fin'>
jetterait
</chunk>
<chunk type = 'PP'>
dans
le
feu
</chunk>
<chunk type = 'PP'>
au
son
</chunk>
<chunk type = 'PP'>
des
tambours
</chunk>
<chunk type = 'C-C'>
et
</chunk>
<chunk type = 'PP'>
des
trompettes
</chunk>
</s>
```

Liste PROLOG

[np(1), vrg(2), cs(3), vfin(4), adj(5), vrg(6), vfin(7), vinf(8), np(9), cc(10), np(11), cs(12), clsj(13), clns(14), vfin(15), pp(16), pp(17), pp(18), cc(19), pp(20)].

Exemple 2 :

il le fit convenir qu'il fallait, si on pouvait, abolir un usage si barbare.

Sortie du module postChunking

```
<s>
<chunk type = 'CL_suj'>
il
</chunk>
<chunk type = 'CL_obj'>
le
</chunk>
<chunk type = 'VP_fin'>
fit
</chunk>
<chunk type = 'VP_fin'>
convenir
</chunk>
<chunk type = 'Camb'>
qu'
</chunk>
<chunk type = 'CL_suj'>
il
</chunk>
<chunk type = 'VP_fin'>
fallait
</chunk>
<chunk type = 'VIRGULE'>
,
</chunk>
<chunk type = 'Camb'>
si
</chunk>
<chunk type = 'CL_suj'>
on
</chunk>
<chunk type = 'VP_fin'>
pouvait
</chunk>
<chunk type = 'VIRGULE'>
,
</chunk>
<chunk type = 'VP_fin'>
abolir
</chunk>
<chunk type = 'NP'>
un
```

```
usage
si
barbare
</chunk>
</s>
```

Liste PROLOG

[clsj(1), clns(2), vfin(3), vinf(4), camb(5), clsj(6), vfin(7), vrg(8), camb(9), clsj(10), vfin(11), vrg(12), vinf(13), np(14)].

C.4 Résultats du module principal et du module de post-traitement pl2prop

À la sortie du module principal, nous obtenons comme résultat d'analyse des deux phrases d'exemple précédentes :

Exemple 1 :

[s, [prop, [prop, [sn, np(1), [subinc, [sep, vrg(2)], [sub, [sep, cs(3)], [prop, [prop, [sv, [v, vfin(4)]], [cmp, [cmp, adj(5)]]]]], [sep, vrg(6)]]], [predicat, [sv, [v, vfin(7)]], [svinf, [vinf, vinf(8)], [sn, np(9), [cmpcrd, [sep, cc(10)], [sn, np(11)]]]]], [sub, [sep, cs(12)], [prop, [prop, clsj(13), [predicat, [sv, [v, [cltq, clns(14)], [v, vfin(15)]]], [cmp, [cmp, pp(16), [cmp, [cmp, pp(17), [cmp, [cmp, pp(18)], [cmpcrd, [sep, cc(19)], [cmp, [cmp, pp(20)]]]]]]]]]]]]]]

Exemple 2 :

[s, [prop, [prop, clsj(1), [predicat, [sv, [v, [cltq, clns(2)], [v, vfin(3)]]], [svinf, [vinf, vinf(4)], [sub, [sep, camb(5)], [prop, [prop, clsj(6), [predicat, [sv, [v, vfin(7), [subinc, [sep, vrg(8)], [sub, [sep, camb(9)], [prop, [prop, clsj(10), [predicat, [sv, [v, vfin(11)]]]]], [sep, vrg(12)]]]], [svinf, [vinf, vinf(13)], [sn, np(14)]]]]]]]]]]]

Le module pl2prop transforme ces résultats du module principal en liste au format xml comme suit :

```
<resultat>

<s>
<sid>1</sid>
<phrase id ='1'>
  <prop>
 <chunk type = 'NP'>
 Almona
 </chunk>
  <subinc>
 <chunk type = 'VIRGULE'>
 ,
 </chunk>
<sub>
 <chunk type = 'C-S'>
 qui
 </chunk>
<prop>
```

```
<sv>
<v>
<chunk type = 'VP_fin'>
était
</chunk>
</v>
</sv>
<chunk type = 'AdjP'>
fort
dévote
</chunk>
</prop>
</sub>
<chunk type = 'VIRGULE'>
,
</chunk>
</subinc>
<predicat>
<sv>
<v>
<chunk type = 'VP_fin'>
fit
</chunk>
</v>
</sv>
<svinf>
<vinf>
<chunk type = 'VP_inf'>
savoir
</chunk>
</vinf>
<chunk type = 'NP'>
le
jour
</chunk>
<chunk type = 'C-C'>
et
</chunk>
<chunk type = 'NP'>
l'
heure
</chunk>
</svinf>
</predicat>
<sub>
<chunk type = 'C-S'>
où
</chunk>
<prop>
<chunk type = 'CL_suj'>
elle
</chunk>
<predicat>
<sv>
<v>
<chunk type = 'CL_ref'>
se
</chunk>
<v>
<chunk type = 'VP_fin'>
jetterait
</chunk>
</v>
</v>
</sv>
<chunk type = 'PP'>
dans
le
feu
</chunk>
<chunk type = 'PP'>
au
son
</chunk>
<chunk type = 'PP'>
des
tambours
</chunk>
<chunk type = 'C-C'>
et
</chunk>
<chunk type = 'PP'>
des
trompettes
</chunk>
</predicat>
</prop>
</sub>
</prop>
</phrase>
</s>

<s>
<sid>2</sid>
<phrase id ='2'>
<prop>
<chunk type = 'CL_suj'>
il
</chunk>
<predicat>
<sv>
<v>
<chunk type = 'CL_obj'>
le
</chunk>
<v>
<chunk type = 'VP_fin'>
fit
</chunk>
</v>
</v>
</sv>
<svinf>
<vinf>
<chunk type = 'VP_fin'>
convenir
</chunk>
</vinf>
<sub>
<chunk type = 'Camb'>
qu'
</chunk>
<prop>
<chunk type = 'CL_suj'>
```

```
il
</chunk>
<predicat>
<sv>
<v>
<chunk type = 'VP_fin'>
fallait
</chunk>
<subinc>
<chunk type = 'VIRGULE'>
,
</chunk>
<sub>
<chunk type = 'Camb'>
si
</chunk>
<prop>
<chunk type = 'CL_suj'>
on
</chunk>
<predicat>
<sv>
<v>
<chunk type = 'VP_fin'>
pouvait
</chunk>
</v>
</sv>
</predicat>
</prop>
</sub>
<chunk type = 'VIRGULE'>
,
</chunk>
</subinc>
</v>
</sv>
<svinf>
<vinf>
<chunk type = 'VP_fin'>
abolir
</chunk>
</vinf>
<chunk type = 'NP'>
un
usage
si
barbare
</chunk>
</svinf>
</predicat>
</prop>
</sub>
</svinf>
</predicat>
</prop>
</phrase>
</s>

</resultat>
```

ANNEXE D

ANNEXE : SIGLÉ JP

D.1 Liste des mots agglutinants et des mots variables de support

Mots agglutinants[1] :

分	ブン	(名詞)
類	ルイ	(名詞)
話	ハナシ	(名詞)
点	テン	(名詞)
かど	カド	(名詞)
次第	シダイ	(名詞)
件	ケン	(名詞)
由	ヨシ	(名詞)
趣	オモムキ	(名詞)
儀	ギ	(名詞)
旨	ムネ	(名詞)
節	フシ	(名詞)
段	ダン	(名詞)
場合	バアイ	(名詞)
始末	シマツ	(名詞)
はこび	ハコビ	(名詞)
はめ	ハメ	(名詞)
あんばい	アンバイ	(名詞)
具合	グアイ	(名詞)
様子	ヨウス	(名詞)
調子	チョウシ	(名詞)
有様	アリサマ	(名詞)
ざま	ザマ	(名詞)
とこ	トコ	(名詞)
あたり	アタリ	(名詞)
辺り	アタリ	(名詞)
へん	ヘン	(名詞)
辺	ヘン	(名詞)
際	キワ	(名詞)
ゆえん	ユエン	(名詞)
考え	カンガエ	(名詞)
所存	ショゾン	(名詞)
さなか	サナカ	(名詞)
かわり	カワリ	(名詞)
あいだ	アイダ	(名詞)
間	アイダ	(名詞)
あと	アト	(名詞)

[1] Chaque élément de la liste est constitué de trois informations : forme de l'occurrence, lecture en *katakana* et catégorie lexicale entre parenthèses.

せつな	セツナ	(名詞)
拍子	ヒョウシ	(名詞)
最中	サイチュウ	(名詞)
ま	マ	(名詞)
間	マ	(名詞)
やさき	ヤサキ	(名詞)
矢先	ヤサキ	(名詞)
かたわら	カタワラ	(名詞)
傍ら	カタワラ	(名詞)
そば	ソバ	(名詞)
まえ	マエ	(名詞)
あげく	アゲク	(名詞)
当座	トウザ	(名詞)
節	セツ	(名詞)
時分	ジブン	(名詞)
以前	イゼン	(名詞)
以来	イライ	(名詞)
ほか	ホカ	(名詞)
ものの	モノノ	(助詞-接続助詞)

Mots variables de support :

よる	ヨル	(動詞-自立)
ある	アル	(動詞-自立)
ない	ナイ	(助動詞)
ない	ナイ	(形容詞-自立)
なる	ナル	(動詞-自立)
する	スル	(動詞-自立)
できる	デキル	(動詞-自立)
いる	イル	(動詞-自立)

D.2 Algorithme de transCabo

◇ VARIABLES :
- Fonc : trait fonctionnel – les valeurs possibles sont des entiers, la valeur par défaut étant 0 ;
- MotVar : trait lié au mot variable – les valeurs possibles sont des entiers ;
- Subst : trait lié au substantif – les valeurs possibles sont : Vrai, Faux ;
- Adv : trait lié à l'adverbe – les valeurs possibles sont : Vrai, Faux ;
- conAgg : trait lié au connecteur agglutinant – les valeurs possibles sont : null, PropAgg(`Forme`), MotAgg(`Forme`), `Forme` étant la forme du mot agglutinant constituant le connecteur ;
- Virg : trait lié à la virgule – les valeurs possibles sont : Vrai, Faux ;
- ParF : trait lié à la parenthèse fermante – les valeurs possibles sont : Vrai, Faux ;
- Theme : trait lié au thème – les valeurs possibles sont des entiers ;
- SubVb : trait lié au substantif verbal – les valeurs possibles sont des entiers ;
- PCas : trait lié à la particule de cas – les valeurs possibles sont : null, Cit ou `Forme`, `Forme` étant la forme de la particule concernée ;
- PConj : trait lié à la particule conjonctive – les valeurs possibles sont : null, Cond, Det(`shika`) ou `Forme`, `Forme` étant la forme de la particule concernée ;
- FAuto : trait lié à la forme autonome – les valeurs possibles sont : Vrai, Faux ;
- FCond : trait lié à la forme de condition – les valeurs possibles sont : Vrai, Faux ;

◇ RÉSULTAT : quatre traits pour chaque *chunk*
- trait Theme : les valeurs possibles sont [+ThemeFort], [+ThemeFaible], [-Theme] ;
- trait Prédicat : les valeurs possibles sont [+Pred], [+PredFaible], [+PredSupp], [+PredDekiru], [+PredIu], [?Pred], [-Pred] ;
- trait Fonc : les valeurs possibles sont [PConj(`Forme`)], [Det*][2], [FAuto*], [FNeutre*], [Cond*], [FImp*], [Adv], [SyntAdv*], [PropAdv*] ;
- trait Mot agglutinant : les valeurs possibles sont [+PropAgg(`Forme`)*], [+PredAgg(`Forme`)*], [+MotAgg(`Forme`)*], [-MAgg] ;

D.2.1 Procédure

Pour chaque constituant du *chunk* en cours de traitement :

A. **Modification de certaines étiquettes**

1. modification de l'étiquette des substantifs autonomes appartenant à la liste des mots agglutinants en substantif non autonome[3] ;
2. modification de l'étiquette des verbes *yoru, aru, suru* en verbe-support ;

[2] Le symbole « * » indique ici la présence éventuelle d'une séquence de caractères quelconques.
[3] La liste des substantifs concernés est présentée dans l'Annexe D.1.

3. modification de l'étiquette du qualificatif *nai* en qualificatif-support ;
4. modification de l'étiquette de l'auxiliaire *nai* en auxiliaire-support ;
5. modification de l'étiquette du verbe *dekiru* en verbe-dekiru ;
6. modification de l'étiquette du verbe *iu* en verbe-iu ;

B. **Extraction des informations nécessaires**
constitution de la liste des constituants avec leur forme d'apparition, leur catégorie et éventuellement le nom de leur forme pour les mots variables ;

C. **Examen de la nature des constituants**
pour chaque élément de liste, effectuer toutes les opérations suivantes avant de passer au traitement de l'élément suivant :

a) TRAIT LIÉ AU MOT VARIABLE : VERBE
si l'élément considéré est étiqueté comme verbe autonome, alors :

1. PCas := null ;
2. **si** conAgg = PropAgg, alors conAgg := null : un mot agglutinant suivi d'un verbe fonctionne comme un auxiliaire et non pas un connecteur ;
3. **si** conAgg = MotAgg, alors conAgg := PredAgg : on conserve le trait agglutinant pour pouvoir le fusionner avec une proposition déterminante dans le cas où elle le précède ;
4. **si** c'est le premier élément du *chunk* et qu'il est un verbe de support, alors MotVar := 1000 ;
5. **si** c'est une forme du verbe *dekiru* et que l'élément précédent n'est pas un substantif (Subst = Faux), alors MotVar := 2000 ;
6. **si** c'est une forme du verbe *iu*, alors MotVar := 3000 ;
7. pour les autres verbes, MotVar := 10 ;

b) TRAIT LIÉ AU MOT VARIABLE : QUALIFICATIF ET AUXILIAIRE
si l'élément considéré est étiqueté comme qualificatif autonome ou auxiliaire, alors :

1. **si** c'est le premier élément du *chunk* et qu'il est étiqueté comme un mot variable de support, alors MotVar := 1000 ;
2. **sinon** MotVar += 5 : l'autonomie de ces mots variables est moins sûre que les verbes[4] ;
3. **si** PCas n'est pas null, alors MotVar += 5 et PCas := null : l'autonomie de ces mots variables est plus forte lorsqu'ils sont précédés par une particule ;
4. **si** conAgg = PropAgg, alors MotVar += 5 et conAgg := null : lorsqu'ils sont précédés par un mot agglutinant, ils constituent un auxiliaire fort avec ce dernier et leur autonomie augmente par conséquent ;

[4] Leur autonomie est moins sûre dans le sens où ils peuvent fonctionner comme de simples qualificatifs.

5. **si** conAgg = MotAgg, alors conAgg := PredAgg ;

c) TRAIT LIÉ À LA PONCTUATION

1. **si** l'élément considéré est étiqueté comme virgule, alors :
   i. **s'** il est précédé par le verbe *iu* (MotVar ≥ 3000), alors MotVar -= 3000 : il est fort probable que le verbe *iu* suivi d'une virgule est autonome ;
   ii. **s'** il est précédé par le substantif verbal (SubVb > 0), alors SubVb := 100 : il est possible que le substantif verbal suivi d'une virgule soit un prédicat ;
   iii. Virg := Vrai ;

   sinon Virg := Faux ;
2. **si** l'élément considéré est étiqueté comme point final, alors :
   i. **s'** il est précédé par le verbe *iu* (MotVar ≥ 3000), alors MotVar -= 3000 : il est fort probable que le verbe *iu* suivi d'un point final est autonome ;
3. **si** l'élément considéré est étiqueté comme parenthèse fermante, alors ParF := Vrai ;
   sinon ParF := Faux ;

d) TRAIT LIÉ AU SUBSTANTIF

si l'élément considéré est étiqueté comme substantif :

1. **si** l'élément considéré est étiqueté comme substantif suffixal, alors MotVar := 0 ;
2. **si** l'élément considéré est étiqueté comme substantif non autonome ou particule adverbiale, et que conAgg est null, alors :
   i. **s'**il est précédé par un mot variable (MotVar > 0), alors MotVar := 0, Fonc := 0 et conAgg := PropAgg(Forme)
   ii. **sinon** conAgg := MotAgg(Forme) **si** Fonc = 0 et Subst = Faux : les mots agglutinants précédés directement pas un substantif ou par un syntagme (sans mot variable) déterminant ne sont pas des connecteurs de propositions ;
3. **si** l'élément considéré est étiqueté comme substantif verbal, alors SubVb := 1,
   sinon SubVb := 0 **si** SubVb < 100 ;
4. Subst := Vrai, Fonc := 0 et PCas := null

sinon (l'élément considéré n'est pas un substantif) : Subst := Faux ;

e) TRAIT LIÉ À L'ADVERBE

si l'élément considéré est étiqueté comme adverbe,
alors Adv := Vrai, Fonc := 0 et PCas := null ;
sinon Adv := Faux, **si** Virg = Faux ;

f) TRAIT LIÉ AU THÈME

1. **si** l'élément considéré est la particule *wa*, alors Theme := 10 ;
   sinon Theme := 0 **si** Virg est Faux : si la particule *wa* est suivie d'un élément quelconque (sauf une virgule), elle n'est pas l'élément thématisant le *chunk* ;
2. **si** l'élément considéré est la particule *mo*, alors Theme := 100 ;

g) TRAIT LIÉ À LA PARTICULE CONJONCTIVE

1. **si** l'élément considéré est étiqueté comme particule conjonctive et qu'il est précédé par une forme autonome (FAuto = Vrai), alors PConj := `Forme` ;
2. **si** l'élément considéré est étiqueté comme particule conjonctive et qu'il est précédé par une forme de condition (FCond = Vrai), alors PConj := Cond et Fonc := -50 ;
3. **si** l'élément considéré est une autre particule conjonctive ou une particule de coordination (exceptée *te* ou *de*), alors PConj := `Forme` ;
4. **si** l'élément considéré est étiqueté comme particule *shika* et qu'il est précédé par une forme autonome (FAuto = Vrai), alors PConj := `Det(shika)` ;

sinon (l'élément considéré n'est ni un symbole de ponctuation ni un thème, et il ne correspond à aucun des éléments concernés par les conditions précédentes) PConj := null ;

h) TRAIT LIÉ À LA FORME DU MOT VARIABLE

1. **si** l'élément considéré est étiqueté comme une forme autonome, alors FAuto := Vrai et Fonc := 1 ;
   sinon FAuto := Faux **si** ParF est Faux ;
2. **si** l'élément considéré est étiqueté comme une forme de condition, alors FCond := Vrai et Fonc := -50 ;
   sinon FCond := Faux **si** Virg est Faux ;

i) TRAIT LIÉ À LA PARTICULE DE CAS

1. **si** l'élément considéré est étiqueté comme particule de cas de citation, alors :
   i. PCas := Cit et conAgg := null ;
   ii. MotVar := 10 **si** conAgg est PropAgg ;
2. **si** l'élément considéré est une particule adverbialisante, PCas := `Forme` ;
3. **si** l'élément considéré est étiqueté comme une autre particule de cas, alors PCas := `Forme` : lorsque deux particules de cas succèdent PCas := $\texttt{Forme}_1$+$\texttt{Forme}_2$;
4. **si** l'élément considéré est une particule déterminante, Fonc := 1 ;

j) TRAIT LIÉ À LA FONCTION

1. **si** l'élément considéré est étiqueté comme une forme autonome ou qu'il a un caractère déterminant ou connectif au substantif, alors Fonc := 1 ;
2. **si** l'élément considéré est étiqueté comme une forme *renyô* (adverbiale), ou qu'il a un caractère connectif à la forme en *te*, alors :
   i. **si** l'élément considéré est la particule conjonctive *de*, Fonc := -100 ;
   ii. **sinon** Fonc := -110 ;
3. **si** l'élément considéré est étiqueté comme la particule conjonctive *te* ou l'auxiliaire *zu*, alors Fonc := -10 ;
4. **si** l'élément considéré est une forme impérative, Fonc := -20 ;
5. **si** l'élément considéré est une particule composée appartenant aux particules composées déterminantes, Fonc := 2 ;

D. **Détermination finale des traits**

a) TRAIT THÈME

1. **si** Theme = 10, alors :
   i. **si** PCas n'est pas null, alors le *chunk* en cours de traitement a le trait [+ThemeFaible] ;
   ii. **sinon** le *chunk* en cours de traitement a le trait [+ThemeFort] ;
2. **sinon** le *chunk* en cours de traitement a le trait [-Theme] ;

b) TRAIT PRÉDICAT

1. **si** MotVar > 0, alors :
   i. **si** MotVar < 10, alors le *chunk* en cours de traitement a le trait [+PredFaible] ;
   ii. **si** MotVar ≥ 1000 et MotVar < 2000, alors le *chunk* en cours de traitement a le trait [+PredSupp] ;
   iii. **si** MotVar ≥ 2000 et MotVar < 3000, alors le *chunk* en cours de traitement a le trait [+PredDekiru] ;
   iv. **si** MotVar ≥ 3000, alors le *chunk* en cours de traitement a le trait [+PredIu] ;
2. **sinon** :
   i. **si** SubVb = Vrai et Virg = Vrai, alors le *chunk* en cours de traitement a le trait [?Pred] : les substantifs verbaux peuvent fonctionner comme des prédicats ;
   ii. **sinon** le *chunk* en cours de traitement a le trait [-Pred] ;

c) TRAIT FONCTION

1. **si** Fonc = 1 (fonction déterminante), alors :

   i. **si** PConj n'est pas null, alors le *chunk* en cours de traitement a le trait [PConj(Forme)] ;
   ii. **si** MotVar > 0 et PCas n'est pas null, alors le *chunk* en cours de traitement a le trait [FAuto+PCas], PCas étant la forme de la particule de cas suivant la forme autonome ;
   iii. **si** PCas n'est pas null, alors le *chunk* en cours de traitement a le trait [Det(PCas)] ;
   iv. **si** aucune des conditions précédentes ne s'applique, alors le *chunk* en cours de traitement a le trait [Det] ;
2. **si** Fonc = 2 (fonction déterminante avec particule), alors le *chunk* en cours de traitement a le trait [Det] ;
3. **si** Fonc = -10 ou Fonc = -110 (forme neutre), alors :
   i. **si** PConj n'est pas null, alors le *chunk* en cours de traitement a le trait [FNeutre+PConj], PConj étant la forme de la particule de mise en relief suivant la forme neutre ;
   ii. **si** Fonc = -110 et que PCas n'est ni null, ni Citation, ni ni, alors le *chunk* en cours de traitement a le trait [SyntAdv], et le trait prédicat [-Pred] : cette règle sert à détecter les erreurs d'étiquetage liées aux substantifs analysés comme des verbes ;
   iii. **si** aucune des conditions précédentes ne s'applique, alors le *chunk* en cours de traitement a le trait [FNeutre] :
      – **si** Virg = Vrai, alors le trait est [FNeutreFort] ;
      – **si** PCas n'est pas null, alors le trait est [FNeutre+PCas], PCas étant la forme de la particule de cas suivant la forme neutre ;
      – **si** Theme = 300, alors le trait est [FNeutre+*mo*] ;
4. **si** Fonc = -100 (forme peut-être neutre), alors :
   i. **si** PCas = null et MotVar $<$ 10, alors :
   si Virg = Vrai, alors le trait est [FNeutre ?Fort],
   sinon le trait est [FNeutre ?]
   ii. **sinon** :
   si PCas = null, alors le trait est [FNeutre],
   sinon le trait est [FNeutre+PCas], PCas étant la forme de la particule de cas suivant la forme neutre ;
5. **si** Fonc = -20, alors :
   i. **si** Virg = Vrai, alors le trait est [FImpFort] ;
   ii. **si** PCas n'est pas null, alors le trait est [FImp+PCas] ou [FImpFort+PCas], PCas étant la forme de la particule de cas suivant la forme impérative ;
   iii. **sinon** le trait est [FImp] ;
6. **si** Fonc = -50, alors :

i. le trait est [Cond] ;

ii. **si** PCas n'est pas null, alors le trait est [Cond+PCas], PCas étant la forme de la particule de cas suivant la forme impérative ;

7. **si** Fonc = 0 (fonction adverbiale), alors :

i. **si** Adv = Vrai, alors le trait est [Adv] ;

ii. **si** MotVar > 0, alors le trait est [PropAdv],
et on ajoute à l'étiquette du trait (PConj), **si** PConj n'est pas null ;

iii. **si** aucune des conditions précédentes ne s'applique, alors le trait est [SyntAdv] ;

iv. **si** Virg = Vrai, alors on ajoute à l'étiquette du trait Fort ;

v. **si** PCas n'est pas null, alors on ajoute à l'étiquette du trait +PCas ;

d) TRAIT MOT AGGLUTINANT

1. **si** ConAgg n'est pas null, alors :

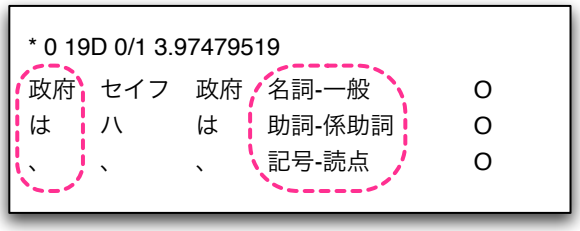
i. le trait est [ConAgg], ConAgg étant la valeur de la variable ConAgg ;

ii. **si** PCas n'est pas null, alors on ajoute à l'étiquette du trait +PCas ;

2. **sinon** le trait est [-MotAgg]

D.2.2 Exemples d'analyse

Exemple 1 : *chunk* thème



A. **Modification d'étiquettes** : aucune

B. **Extraction des informations nécessaires** :
À partir du résultat de CaboCha ci-dessus, en extrayant les informations nécessaires (encadrées dans la figure), nous créons la liste des constituants comme suit :

(a) 政府 (*seifu*, gouvernement) : 名詞-一般 (substantif-commun)

(b) は (*wa*, [thème]) : 助詞-係助詞 (particule-particule *kakari*)

(c) 、 ([virgule]) : 記号-読点 (ponctuation-virgule)

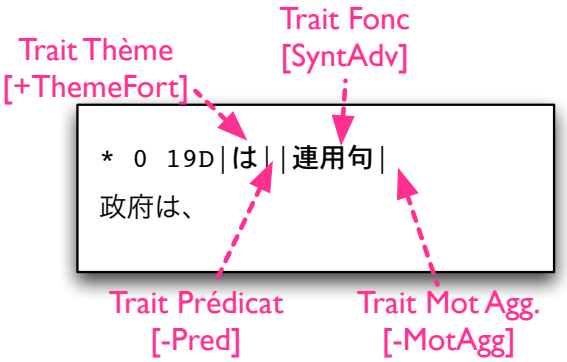
C. **Examen de la nature des constituants** :

1. initialisation des variables :
   Fonc=0, MotVar=0, Subst=Faux, conAgg=null, Virg= Faux, Theme=0, SubVb=0, PCas=null, PConj=null, FAuto=Faux, FCond=Faux
2. examen de l'élément (a) :
   Subst=Vrai, Pcas=null (selon PROCÉDURE C(d)4)
3. examen de l'élément (b) :
   Theme=10 (selon PROCÉDURE C(f)1) Subst=Faux (selon PROCÉDURE C(d)4)
4. examen de l'élément (c) :
   Virg=Vrai (selon PROCÉDURE C(c)1iii)
   Aucune modification de la valeur de la variable Theme car Virg=Vrai (selon PROCÉDURE C(f)1)

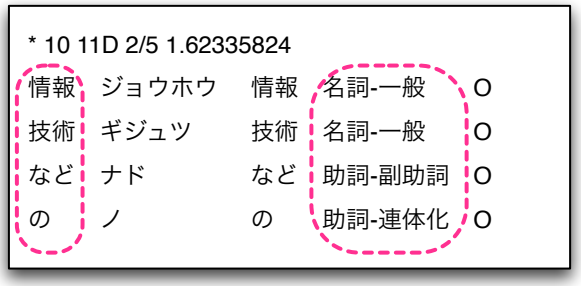
D. **Détermination finale des traits** :

1. Trait Thème : [+ThemeFort]
2. Trait Prédicat : [-Pred]
3. Trait Fonction : [SyntAdv]
4. Trait Mot agglutinant : [-MotAgg]

En sortie du module transCabo, nous obtenons la nouvelle étiquette du *chunk* comportant des informations morpho-syntaxiques ainsi calculée comme suit :



Exemple 2 : *chunk* à fonction déterminante



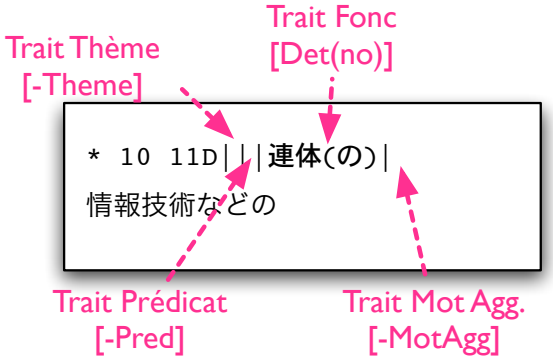
A. **Modification d'étiquettes** : aucune

B. **Extraction des informations nécessaires** :
À partir du résultat de CaboCha ci-dessus, en extrayant les informations nécessaires (encadrées dans la figure), nous créons la liste des constituants comme suit :
   (a) 情報 (*jôhô*, information) : 名詞-一般 (substantif-commun)
   (b) 技術 (*gijutsu*, technique) : 名詞-一般 (substantif-commun)
   (c) など (*nado*, entre autres) : 助詞-副助詞 (particule-particule adverbiale)
   (d) の (*no*, [déterminant]) : 助詞-連体化 (particule-particule déterminante)

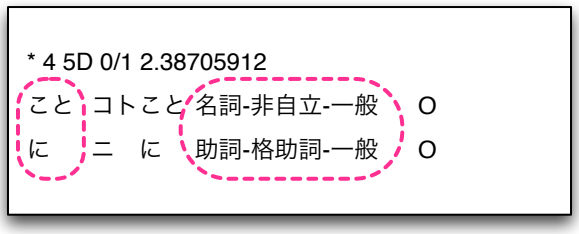
C. **Examen de la nature des constituants** :
   1. initialisation des variables :
      Fonc=0, MotVar=0, Subst=Faux, conAgg=null, Virg= Faux, Theme=0, SubVb=0, PCas=null, PConj=null, FAuto=Faux, FCond=Faux
   2. examen de l'élément (a) :
      Subst=Vrai, Pcas=null (selon PROCÉDURE C(d)4)
   3. examen de l'élément (b) :
      Subst=Vrai, Pcas=null (selon PROCÉDURE C(d)4)
   4. examen de l'élément (c) :
      Aucune affectation à la variable ConAgg car Subst=Vrai (selon PROCÉDURE C(d)2)
      Subst=Faux (selon PROCÉDURE C(d)4)
   5. examen de l'élément (d) :
      Fonc=1, PCas=no (selon PROCÉDURE C(i)4)

D. **Détermination finale des traits** :
   1. Trait Thème : [-Theme]
   2. Trait Prédicat : [-Pred]
   3. Trait Fonction : [Det(no)]
   4. Trait Mot agglutinant : [-MotAgg]

En sortie du module transCabo, nous obtenons la nouvelle étiquette du *chunk* comportant des informations morpho-syntaxiques ainsi calculée comme suit :



Exemple 3 : *chunk* à connecteur agglutinant



A. **Modification d'étiquettes** : aucune

B. **Extraction des informations nécessaires** :
À partir du résultat de CaboCha ci-dessus, en extrayant les informations nécessaires (encadrées dans la figure), nous créons la liste des constituants comme suit :

(a) こ と (*koto*, chose) : 名 詞-非 自 立-一 般 (substantif-non autonome-commun)

(b) に (*ni*, [*ni*]) : 助詞-格助詞-一般 (particule-particule de cas-général)

C. **Examen de la nature des constituants** :

1. initialisation des variables :
Fonc=0, MotVar=0, Subst=Faux, conAgg=null, Virg= Faux, Theme=0, SubVb=0, PCas=null, PConj=null, FAuto=Faux, FCond=Faux

2. examen de l'élément (a) :
conAgg=MotAgg(koto) (selon PROCÉDURE C(d)2)
Subst=Vrai, Pcas=null (selon PROCÉDURE C(d)4)

3. examen de l'élément (b) :
PCas=ni (selon PROCÉDURE C(i)3) Subst=Faux (selon PROCÉDURE C(d)4)

D. **Détermination finale des traits** :

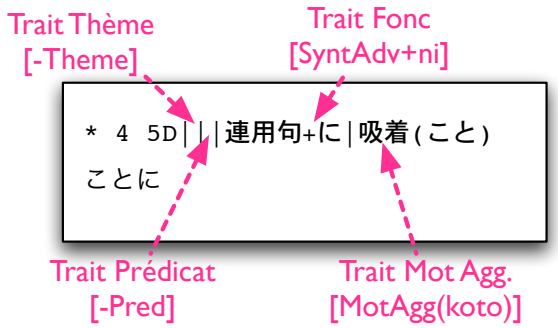
1. Trait Thème : [-Theme]

2. Trait Prédicat : [-Pred]

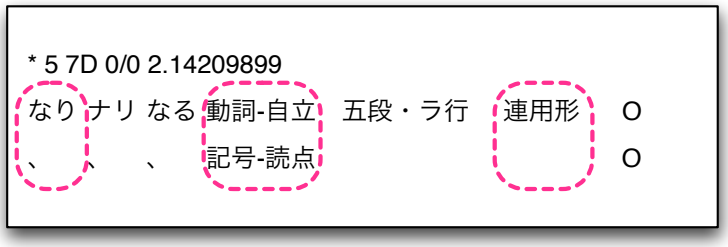
3. Trait Fonction : [SyntAdv+ni]

4. Trait Mot agglutinant : [MotAgg(koto)+ni]

En sortie du module transCabo, nous obtenons la nouvelle étiquette du *chunk* comportant des informations morpho-syntaxiques ainsi calculée comme suit :



Exemple 4 : *chunk* verbal



A. **Modification d'étiquettes** :
L'étiquette du verbe なる (*naru*), 動詞-自立 (verbe-autonome), est modifiée en 動詞-自立-支持 (verbe-autonome-support) ;

B. **Extraction des informations nécessaires** :
En extrayant les informations nécessaires (encadrées dans la figure), nous créons la liste des constituants comme suit :

(a) なり (*nari*, devenir) : 動詞-自立-支持 (verbe-autonome-support), 連用形 (forme *renyô*)

(b) 、 ([virgule]) : 記号-読点 (ponctuation-virgule)

C. **Examen de la nature des constituants** :

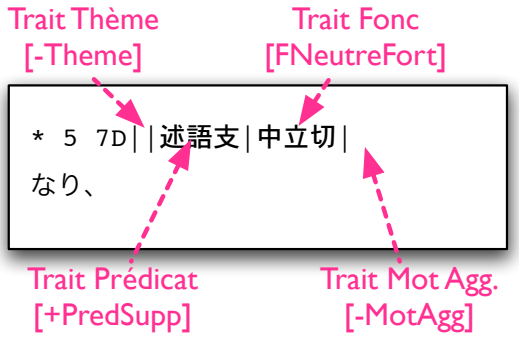
1. initialisation des variables :
Fonc=0, MotVar=0, Subst=Faux, conAgg=null, Virg= Faux, Theme=0, SubVb=0, PCas=null, PConj=null, FAuto=Faux, FCond=Faux
2. examen de l'élément (a) :
Pcas=null, conAgg=null, MotVar=1000 (selon PROCÉDURE Ca)
Fonc=-10 (selon PROCÉDURE C(j)2)
3. examen de l'élément (b) :
Virg=Vrai (selon PROCÉDURE C(c)1iii)

D. **Détermination finale des traits** :

1. Trait Thème : [-Theme]
2. Trait Prédicat : [+PredSupp]

3. Trait Fonction : [FneutreFort]
4. Trait Mot agglutinant : [-MotAgg]

En sortie du module transCabo, nous obtenons la nouvelle étiquette du *chunk* comportant des informations morpho-syntaxiques ainsi calculée comme suit :



D.3 Algorithme de regroupement des *chunks*

◇ VARIABLES

Pour le segment courant :

- LMembres : liste des membres du segment – les valeurs possibles sont des entiers ;
- Pere : indice du *chunk* dont dépend le segment – les valeurs possibles sont des entiers et F pour le dernier *élément* ;
- Theme : trait Thème du segment courant – les valeurs possibles sont des entiers ;
- Pred : trait Prédicat du segment courant – les valeurs possibles sont des entiers ;
- Fonc : trait Fonction du segment courant – les valeurs possibles sont des entiers ;
- MAgg : trait Mot agglutinant du segment courant – les valeurs possibles sont des entiers ;
- nbMembres : nombre d'éléments dans LMembres – les valeurs possibles sont des entiers ;

Pour le segment précédent :

- LMembresP : liste des membres du segment précédent – les valeurs possibles sont des entiers ;
- PereP : indice du *chunk* dont dépend le segment précédent – les valeurs possibles sont des entiers, la valeur par défaut étant 100 ;
- ThemeP : trait Thème du segment précédent – les valeurs possibles sont des entiers ;
- PredP : trait Prédicat du segment précédent – les valeurs possibles sont des entiers ;

- FoncP : trait Fonction du segment précédent – les valeurs possibles sont des entiers ;
- MAggP : trait Mot agglutinant du segment précédent – les valeurs possibles sont des entiers ;
- FusionP : trait du segment précédent lié à la possibilité de fusion avec le segment qui le suit – les valeurs possibles sont des entiers, la valeur par défaut étant 100 ;

Autres variables :

- LPhrase : liste Phrase contenant les segments constituant la phrase – à l'état initial, elle contient tous les *chunks* constituant la phrase ;
- LRes : liste des segments constituants obtenus après la fusion ;

D.3.1 Procédure

Pour chaque *chunk* constituant la phrase en cours de traitement, répéter les opérations suivantes jusqu'à épuisement des possibilités de fusion :

A. **Lecture du segment courant dans LPhrase** :
Affectation des valeurs correspondantes aux variables pour le segment courant : SegC, LMembres, Theme, Pred, Fonc et MAgg ;

B. **Examen de la possibilité de fusion du segment courant avec le segment précédent**

a) CAS CLASSIQUE DE FUSION :
si le segment précédent dépend d'un des *chunks* constituant le segment courant (FusionP ∈ LMembres), alors fusionner le segment courant avec le segment précédent :

1. LMembres := LMembresP +LMembres
2. **si** PereP < 100, alors MAgg := [-MotAgg] : les mots agglutinants déterminés par un syntagme non verbal ne fonctionnent pas comme des connecteurs ;
sinon MAgg := [+PropAgg(Forme)] **si** MAgg = [+MotAgg(Forme)] ;
3. **si** le syntagme précédent est un syntagme en particule de cas (PereP = 150), alors :
   i. **si** le syntagme courant se termine par le substantif verbal (Pred = [?Pred]), on le considère comme un prédicat à la forme neutre sans terminaison, Pred := [FNeutre] ;
   ii. **si** le syntagme courant se termine par *de* (Pred = [FNeutre ?]), son statut de verbe est confirmé, Pred := [FNeutre] ;
4. **si** le segment courant n'est pas le dernier élément de la liste (Pere n'est pas F), alors passer à l'actualisation des traits du segment courant (→ C) ;
sinon créer la nouvelle étiquette avec les variables du segment courant et la stocker dans la liste LRes et passer à la préparation du nouveau tour (→ D) ;

b) CAS DE FUSION PRÉDICAT (NON FORT) + CONNECTEUR AGGLUTINANT :
si le segment précédent contient un prédicat non fort (FusionP = 200) et que le segment courant contient un mot aggultinant (MAgg = [+MotAgg]) et que le segment précédent dépend d'un des *chunks* constituant le segment courant (PereP ∈ LMembres), alors fusionner le segment courant avec le segment précédent :

1. LMembres := LMembresP +LMembres
2. MAgg := [PropAgg] ;
3. affecter les valeurs des variables du segment courant aux variables correspondantes du segment précédent ;
4. FusionP := 200 ;
5. **si** le segment courant n'est pas le dernier élément de la liste (Pere n'est pas F), alors passer à la lecture de l'élément suivant LPhrase (→ A) ;
   sinon créer la nouvelle étiquette avec les variables du segment courant et la stocker dans la liste LRes et passer à la préparation du nouveau tour (→ D) ;

c) CAS DE FUSION TO (CITATION) + PRÉDICAT IU :
si le segment précédent est terminé par la particule de citation *to* (FusionP = 700) et que le segment courant contient le prédicat *iu* non suivi de ponctuation (Pred = [+PredIu]) et que le segment précédent dépend d'un des *chunks* constituant le segment courant (PereP ∈ LMembres), alors fusionner le segment courant avec le segment précédent :

1. LMembres := LMembresP +LMembres
2. Pred := [+Pred] ;
3. affecter les valeurs des variables du segment courant aux variables correspondantes du segment précédent ;
4. FusionP := 200 ;
5. **si** le segment courant n'est pas le dernier élément de la liste (Pere n'est pas F), alors passer à la lecture de l'élément suivant LPhrase (→ A) ;
   sinon créer la nouvelle étiquette avec les variables du segment courant et la stocker dans la liste LRes et passer à la préparation du nouveau tour (→ D) ;

d) CAS DE FUSION SHIKA + PRÉDICAT SUPPORT :
si le segment précédent contient la particule *shika* (FusionP = 400) et que le segment courant contient le prédicat support (Pred = [+PredSupp]) et que le segment précédent dépend d'un des *chunks* constituant le segment courant (PereP ∈ LMembres), alors fusionner le segment courant avec le segment précédent :

1. LMembres := LMembresP +LMembres

2. Pred := [+Pred] ;
3. affecter les valeurs des variables du segment courant aux variables correspondantes du segment précédent ;
4. FusionP := 200 ;
5. **si** le segment courant n'est pas le dernier élément de la liste (Pere n'est pas F), alors passer à la lecture de l'élément suivant LPhrase (→ A) ;
   sinon créer la nouvelle étiquette avec les variables du segment courant et la stocker dans la liste LRes et passer à la préparation du nouveau tour (→ D) ;

e) CAS DE FUSION WA + PRÉDICAT SUPPORT :
si le segment précédent est étiqueté comme thème (FusionP = 250) et que le segment courant contient le prédicat support (Pred = [+PredSupp]) et que le segment précédent dépend d'un des *chunks* constituant le segment courant (PereP ∈ LMembres), alors fusionner le segment courant avec le segment précédent :

1. LMembres := LMembresP +LMembres
2. Pred := [+Pred] ;
3. affecter les valeurs des variables du segment courant aux variables correspondantes du segment précédent ;
4. FusionP := 200 ;
5. **si** le segment courant n'est pas le dernier élément de la liste (Pere n'est pas F), alors passer à la lecture de l'élément suivant LPhrase (→ A) ;
   sinon créer la nouvelle étiquette avec les variables du segment courant et la stocker dans la liste LRes et passer à la préparation du nouveau tour (→ D) ;

f) MODIFICATION DE L'ÉTIQUETTE : WA + SYNTAGME ADVERBIAL OU ADVERBE :
si le segment précédent est étiqueté comme thème (FusionP = 250) et que le segment courant est un syntagme adverbial ou un adverbe (Fonc = [SyntAdv] ou [Adv]) sans mot variable, et que le segment précédent dépend d'un des *chunks* constituant le segment courant (PereP ∈ LMembres), alors modifier l'étiquette du segment courant :

1. **si** le segment courant est un syntagme adverbial se terminant par la particule *de*, il doit être considéré comme un verbe à la forme neutre, Pred := [+Pred] et Fonc := [FNeutre] ;
2. **si** le segment courant est étiqueté adverbe (Fonc = [Adv]) ou syntagme adverbial suivi de la particule *ni* (Fonc = [SyntAdv+ni]), on considère le segment en *wa* précédent comme un segment en *wa* faible, ThemeP := [+ThemeFaible] ;
3. **si** le segment courant est un syntagme adverbial fort (Fonc = [SyntAdvFort]), il doit être considéré comme un verbe à la forme neutre, Pred := [+Pred] et Fonc := [FNeutre] ;

g) MODIFICATION DE L'ÉTIQUETTE : PROPOSITION À CONNECTEUR AGGLUTINANT + PRÉDICAT EN *de* OU SUBSTANTIF VERBAL :
si le segment précédent est une proposition à connecteur agglutinant (FusionP = 500 ou 600) et que le segment courant est un syntagme terminé par *de* (Fonc = [FNeutre ?]) ou par un substantif verbal (Pred = [?Pred]), et que le segment précédent dépend d'un des *chunks* constituant le segment courant (PereP ∈ LMembres),
alors modifier l'étiquette du segment courant, Pred := [+Pred] et Fonc := [FNeutre] ;

h) CAS DE FUSION PROPOSITION AGGLUTINANTE + PRÉDICAT SUPPORT :
si le segment précédent est étiqueté comme proposition à connecteur agglutinant (FusionP = 500), que le segment courant contient le prédicat support (Pred = [+PredSupp]) et que le segment précédent dépend d'un des *chunks* constituant le segment courant (PereP ∈ LMembres), alors fusionner le segment courant avec le segment précédent :
   1. LMembres := LMembresP +LMembres
   2. Pred := [+Pred] ;
   3. affecter les valeurs des variables du segment courant aux variables correspondantes du segment précédent ;
   4. FusionP := 200 ;
   5. **si** le segment courant n'est pas le dernier élément de la liste (Pere n'est pas F), alors passer à la lecture de l'élément suivant LPhrase (→ A) ;
      sinon créer la nouvelle étiquette avec les variables du segment courant et la stocker dans la liste LRes et passer à la préparation du nouveau tour (→ D) ;

i) CAS DE FUSION PROPOSITION EN KOTO + PRÉDICAT SUPPORT / DEKIRU :
si le segment précédent est étiqueté comme proposition à connecteur agglutinant *koto* (FusionP = 600), que le segment courant contient le prédicat support ou le prédicat *dekiru* (Pred = [+PredSupp] ou [+PredDekiru]) et que le segment précédent dépend d'un des *chunks* constituant le segment courant (PereP ∈ LMembres), alors fusionner le segment courant avec le segment précédent :
   1. LMembres := LMembresP +LMembres
   2. Pred := [+Pred] ;
   3. affecter les valeurs des variables du segment courant aux variables correspondantes du segment précédent ;
   4. FusionP := 200 ;
   5. **si** le segment courant n'est pas le dernier élément de la liste (Pere n'est pas F), alors passer à la lecture de l'élément suivant LPhrase (→ A) ;
      sinon créer la nouvelle étiquette avec les variables du segment courant et la stocker dans la liste LRes et passer à la préparation du nouveau tour (→ D) ;

j) CAS DE FUSION PROPOSITION NEUTRE + PRÉDICAT ISOLÉ :
si le segment précédent est étiqueté comme proposition à la forme neutre (FusionP = 150), que le segment courant contient le prédicat (Pred = [+Pred*] et qu'il est constitué d'un seul *chunk* (nbMembres = 1), et que le segment précédent dépend d'un des *chunks* constituant le segment courant (PereP ∈ LMembres),
alors fusionner le segment courant avec le segment précédent :

1. LMembres := LMembresP +LMembres
2. Pred := [+Pred] ;
3. affecter les valeurs des variables du segment courant aux variables correspondantes du segment précédent ;
4. **si** le segment courant n'est pas le dernier élément de la liste (Pere n'est pas F), alors passer à la lecture de l'élément suivant LPhrase (→ A) ;
   sinon créer la nouvelle étiquette avec les variables du segment courant et la stocker dans la liste LRes et passer à l'actualisation des traits (→ C) ;

k) MODIFICATION DE L'ÉTIQUETTE : SUBSTANTIF VERBAL :
si le segment courant est un substantif verbal (Pred = [?Pred], alors :

1. **si** le segment précédent se termine par la particule de citation (FusionP n'est pas 700), alors Pred := [+Pred] ;
2. **sinon** il est peu probable que le substantif verbal courant fonctionne comme un prédicat, Pred := [-Pred] ;

l) CAS D'AUCUNE FUSION :
si aucune des conditions ci-dessus ne s'applique, alors pas de fusion :

1. **si** le segment courant n'est pas le premier élément de la liste (FusionP n'est pas 100), alors créer une étiquette avec les variables du segment précédent et la stocker dans la liste LRes ;

C. **Actualisation des traits du segment courant liés à la possibilité de fusion avec le segment suivant**

a) SEGMENT CITATION :
si le segment courant est étiqueté comme citation (Fonc = [*+Cit]) et qu'il contient un mot variable (Pred = [+Pred*]), alors FusionP = 700 et PereP = Pere : possibilité de fusion avec le prédicat support qui le suit ;

b) SEGMENT CONNECTEUR AGGLUTINANT :
si le segment courant est étiqueté comme proposition à connecteur agglutinant (MAgg = [+PropAgg*]), alors :

i. **si** le segment courant se termine par *de* (Fonc = [FNeutre ?]), alors son statut de verbe est confirmé (Fonc := [FNeutre]), FusionP = 300 et PereP = Pere : aucune fusion possible ;

ii. **si** le mot agglutinant constituant le connecteur est *koto* (MAgg = [+Prop(koto)*]), alors FusionP = 600 et PereP = Pere : possibilité de fusion avec le prédicat support qui le suit ;

iii. **sinon** FusionP = 500 et PereP = Pere : possibilité de fusion avec le prédicat support qui le suit ;

c) SEGMENT THÈME :
si le segment courant est étiqueté comme thème (Theme = [+Theme*]), alors FusionP = 250 et PereP = Pere : peu de possibilité de fusion avec le segment suivant ;

d) SEGMENT PRÉDICAT FAIBLE À UNE FORME DÉTERMINANTE :
si le segment courant est étiqueté comme déterminant (Fonc = [Det]) et qu'il contient un mot variable (Pred = [+Pred*]), alors :

i. **si** ce mot variable est étiqueté comme prédicat faible (Pred = [+Pred-Faible*]) et qu'il est constitué d'un seul *chunk* (nbMembres = 1), alors FusionP = Pere et PereP = 100 : fusion avec le segment suivant :

ii. **sinon** FusionP = 200 et PereP = Pere : possibilité de fusion avec le connecteur agglutinant qui le suit ;

e) SEGMENT DÉTERMINANT SHIKA :
si le segment courant est étiqueté comme déterminant en *shika* (Fonc = [Det(shika)]), alors FusionP = 400 et PereP = Pere : possibilité de fusion avec le prédicat support qui le suit ;

f) SEGMENT PRÉDICAT À UNE FORME NEUTRE :
si le segment courant comporte le prédicat à une forme neutre (Pred = [+PredNeutre*]), alors :

i. **s'**il est constitué d'un seul *chunk* (nbMembres = 1) et qu'il est précédé par un segment qui ne dépend pas de lui (PereP non ∈ LMembres), alors FusionP = Pere et PereP = Pere : fusion avec le segment suivant ;

ii. **sinon** FusionP = 150 et PereP = Pere : possibilité de fusion avec le segment suivant sous certaines conditions ;

g) SEGMENT PRÉDICAT FORT :
si le segment courant comporte le prédicat fort (Pred = [+PredFort*]), alors FusionP = 300 et PereP = Pere : aucune fusion possible :

h) AUTRES SEGMENTS À MOT VARIABLE :
si le segment courant comporte le prédicat (Pred = [+Pred]), alors FusionP = 200 et PereP = Pere : possibilité de fusion avec le connecteur agglutinant qui le suit ;

i) AUTRES SEGMENTS SANS MOT VARIABLE :

i. **si** le segment courant comporte la particule de cas (Fonc = [*+*]) non déterminante, alors FusionP = Pere et PereP = 150 : possibilité de fusion avec le connecteur agglutinant qui le suit ;

ii. **sinon** FusionP = Pere et PereP = Pere : fusion avec le segment suivant ;

j) MISE À JOUR DES VARIABLES DU SEGMENT PRÉCÉDENT :
pour tout type de segment, affecter les valeurs des variables du segment courant aux variables correspondantes du segment précédent (sauf PereP et FusionP), avant de passer à la lecture de l'élément suivant LPhrase (→ A) ;

D. **Préparation d'un nouveau tour**

1. mise à jour de la liste des segments : vider la liste LPhrase et y affecter tous les éléments de LRes ;
2. initialiser les autres variables avant de commencer la lecture du premier segment de la liste pour le nouveau tour (→ A).

D.4 Règles de détermination du type de proposition

D.4.1 Quatre traits de proposition

Les propositions ainsi reconstituées sont toutes caractérisées pour les quatre traits par une valeur :

Trait	Valeurs possibles
Thème	[+Thème fort], [+Thème faible], [-Thème]
Prédicat	[+Prédicat], [-Prédicat]
Fonction	[Neutre+PART. ?], [Autonome+PART. ?], [Adverbial+PARTCAS ?], [Déterminant], [Citation+PARTCAS ?], [Conjonction(FORME)], [Interrogation+PARTCAS ?]
Prop. Agg.	[PropAgg(FORME)+PARTCAS ?], [-PropAgg]

TAB. D.1 – Traits des propositions détectées

D.4.2 Règles

Les règles de détermination du type de proposition à partir de ces quatre traits sont les suivantes :

1. Trait Thème = [-Thème] ou [+Thème faible]
   a) Trait Proposition à connecteur agglutinant = [-PropAgg]
      i. Trait Prédicat = [-Prédicat]
         A. si le segment considéré précède le thème fort, alors c'est un élément externe et son type est Externe ;
         B. sinon c'est le complément d'un prédicat (l'insertion dans la proposition constituée par le prédicat régissant) ;

ii. Trait Prédicat = [+Prédicat]

A. si le segment considéré précède le thème fort, alors c'est une proposition externe et son type est Externe(TRAITFONCTION) ;

B. sinon c'est une proposition subordonnée interne à une autre proposition et son type est TRAITFONCTION ;

b) Trait Proposition à connecteur agglutinant = [PropAgg(FORME)+PARTCAS ?]

i. si le segment considéré précède le thème fort, alors c'est une proposition externe et son type est Externe(PropAgg(FORME)+PARTCAS ?) ;

ii. sinon c'est une proposition subordonnée interne à une autre proposition et son type est PropAgg(FORME+PARTCAS ?) ;

2. Trait Thème = [+Thème fort]

a) Trait Proposition à connecteur agglutinant = [-PropAgg]

i. Trait Fonction = [Citation] ou [Interrogation], alors le type est ThèmeFaible(TraitFonction) ;

ii. sinon le type est ThèmeFort ;

b) Trait Proposition à connecteur agglutinant = [PropAgg(FORME)+PARTCAS ?], alors son type est Thème(PropAgg(FORME)+PARTCAS ?).

D.4.3 Exemple

Considérons maintenant un exemple de détermination du type de proposition réalisée par ce module pour mieux illustrer l'explication.

La figure D.2 montre le résultat fourni par le troisième module (cadre du bas) à partir du résultat du premier regroupement réalisé par le deuxième module (cadre du haut) pour l'analyse de la phrase :

治験は、新薬について、製薬会社が
厚労省に　承認申請する　際に
必要な　安全性、有効性の　データを　集める　ために　実施される。
(*chiken wa - shinyaku nitsuite - seiyaku gaisha ga*
- kôrôshô ni - shônin shinsei suru - sai ni
- hitsuyôna - anzensei - yûkôsei no - dêta wo - atsumeru - tameni - jisshi sareru)
(essai clinique [wa] - pour les nouveaux médicaments - les sociétés pharmaceutiques [ga]
- le Ministère de la santé et du travail [ni] - demander l'autorisation - à l'occasion de
- nécessaire - sécurité - efficacité [no] - données [wo] - collecter - afin de - réaliser)
« Des essais cliniques de médicaments sont réalisés pour les nouveaux produits, afin d'obtenir des données sur leur sécurité et leur efficacité, nécessaires lorsqu'une société pharmaceutique fait une demande d'autorisation auprès du Ministère de la santé et du travail. »

1. Le segment 1 résultant du premier regroupement ayant le trait [+ThèmeFort] constitue à lui-seul l'unité 1 (marqué id='1') étiquetée « thème » ;

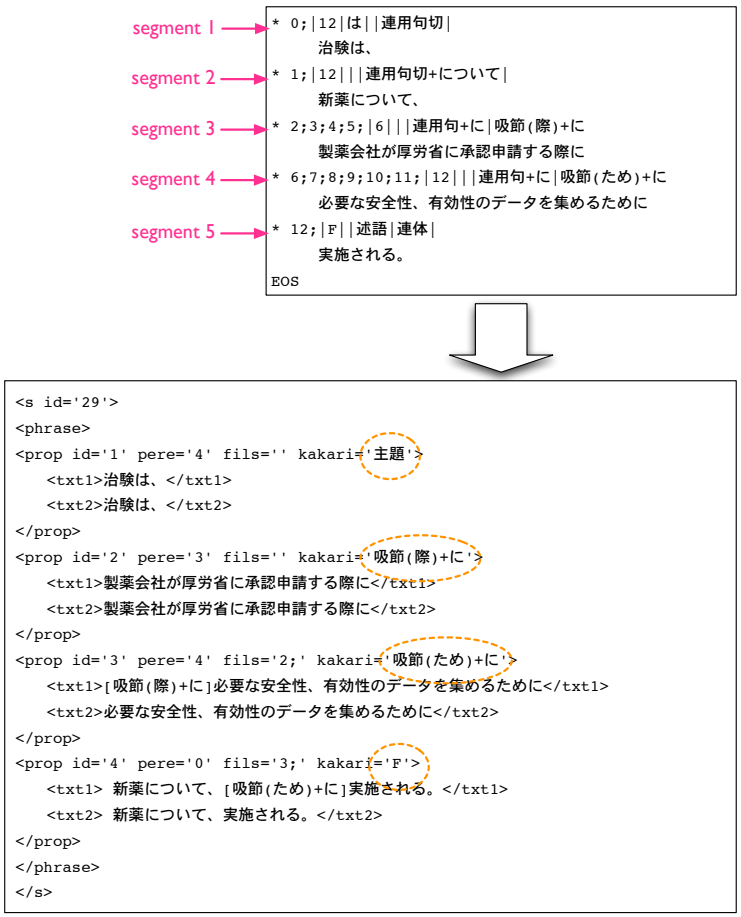


FIG. D.2 – Exemple du résultat de la détermination du type de proposition

2. Le segment 2 qui ne comporte pas de mot variable et est situé postérieurement au thème – donc interne à la proposition – est inséré dans la proposition 4 (marquée id='4') que constitue le prédicat régissant ;
3. Le segment 3 ayant le trait [PropAgg(SAI)+ni] constitue la proposition 2 (marquée id='2') ayant l'étiquette héritée du trait Proposition à connecteur agglutinant ;
4. Le segment 4 ayant le trait [PropAgg(TAME)+ni] constitue la proposition 3 (marquée id='3') ayant l'étiquette héritée du trait Proposition à connecteur agglutinant. Il comporte également l'indication de la présence d'une subordonnée (entre crochets dans la ligne entourée des tags <txt1>), la proposition 2 marquée dans l'étiquette comme fils='2' ;
5. Le segment 5, prédicat principal, constitue la proposition 4 (marquée id='4'), proposition racine étiquetée F. Il comporte un complément discon-

tinu, le segment 2, et l'indication de la présence d'une subordonnée (entre crochets dans la ligne entourée des tags <txt1>), la proposition 3 marquée dans l'étiquette comme fils='3'.

LISTE DES CORPUS UTILISÉS

Corpus utilisés lors de l'évaluation des systèmes

(dans l'ordre alphabétique)

1. **Asahi** : texte composé de six articles du journal Asahi version électronique, disponibles sur `http://www.asahi.com/`
   1) 村上世彰被告「無罪と確信」検察側と全面対決の姿勢 (*murakami yoshiaki hikoku "muzai to kakushin" kensatsu gawa to zenmen taiketsu no shisei*, « "Certain d'être acquitté", le suspect MURAKAMI Yoshiaki manifeste son intention d'une confrontation totale avec le procureur »), 30 novembre 2006 ;
   2) 防衛庁の「省」昇格　衆院を通過 (*bôeichô no "shô" shôkaku, shûin wo tsûka*, « Promotion de l'Agence de la Défense au statut de "ministère", votée par le parlement »), 30 novembre 2006 ;
   3) 目黒区議会が大混乱、公明党区議団は総辞職 (*meguro ku gikai ga daikonran, kômeitô kugidan wa sôjishoku*, « Chaos dans le conseil d'arrondissement de Meguro, démission de tous les membres du parti Kômei »), 30 novembre 2006 ;
   4) 「科協」、北朝鮮に技術情報パイプ (*"kakyô" kitachôsen ni gijutsu jôhô paipu*, « "Kakyo", Tuyau pour les informations techniques vers la Corée du nord »), 30 novembre 2006 ;
   5) 住宅ローン、減税目減り分救済へ (*jûtaku rôn, genzei meberi bun kyûsai e*, « Prêt immobilier – aide à la perte de réduction des impôts »), 30 novembre 2006 ;
   6) 教育基本法改正案を単独採決　衆院特別委 (*kyôiku kihonhô kaisei an wo tandoku saiketsu, shûin tokubetsu i*, « Vote en solo de la proposition de modification de la loi fondamentale sur l'éducation »), 15 novembre 2006.
2. **Balth** : constitué de la traduction anglaise du texte intégral de *Balthasar* de Anatole France, repris du site internet : `http://www.nogginworks.org/index.html`
3. **BalthJP** : traduction japonaise de *Balthasar* par Ryunosuke AKUTAGAWA, réalisé à partir de la version électronique distribuée sur le site Aozora-

bunko :
`http://www.aozora.gr.jp`

4. **Bio** : article en français « L'homme à l'épreuve de la révolution biologique » *Label France*, no. 49, janvier/mars 2003, Ministère des Affaires Étrangères.
5. **BioJP** : traduction japonaise de l'article ci-sessus « 生物学上の大変革に直面する人類 ». *Label France* (version japonaise), no. 49, janvier/mars 2003, Ministère des Affaires Étrangères.
6. **Brevet1** : 電動機用変速機の直流母線上の電圧管理システムと方法 (*dendôki yô hensokuki no chokuryûbosen jô no den'atsu kanri shisutemu to hôhô*), original en français « Procédé et système de gestion de la tension sur le bus continu d'un variateur de vitesse pour moteur électrique ».
7. **BRVF** : composés de deux brevets techniques :
   – Système de transport et de distribution d'énergie électrique ;
   – Dispositif de commande d'un transistor de puissance.
8. **BRVFJP** : composés de deux traductions des brevets techniques français ci-dessus :
   – 送電・配電システム ;
   – パワートランジスター指令装置.
9. **BRVJ** : composés d'une traduction du brevet technique japonais ci-dessous :
   – Structure de connexion à la terre de modules.
10. **BRVJJP** : composés d'un brevet technique :
    – モジュールの接地構造.
11. **EU** : texte en anglais de l'Union européen « EU priority proposals for regulatory reform in Japan », du 16 octobre 2003.
12. **EUJP** : traduction japonaise du texte ci-dessus « 日本の規制改革に関するEU優先提案（仮訳） », par Delegation of the European Commission in Japan.
13. **FdT** : MURAKAMI, Haruki. *La fin des temps*, Éditions du Seuil, 1992.
14. **FdTJP** : MURAKAMI, Haruki. 世界の終りとハードボイルド・ワンダーランド (*sekai no owari to hâdoboirudo wandâ rando*), Shinchosha 1985.
15. **FIV** : article en français « La fabrication de l'être humain » *Label France*, no. 49, janvier/mars 2003, Ministère des Affaires Étrangères.
16. **FIVJP** : traduction japonaise de l'article ci-sessus : « 人間の製造、不妊治療から優生学まで ». *Label France* (version japonaise), no. 49, janvier/mars 2003, Ministère des Affaires Étrangères.
17. **G8** : texte en français du sommet G8 2003 « Lutte contre la famine », disponible sur le site :
    `http://www.g8.fr/evian/francais`
18. **G8JP** : version japonaise du texte ci-dessus « 特にアフリカにおける飢餓に対する行動（仮訳） », traduction publiée sur le site de Ministère des

Affaires Étrangères du Japon :
`http://www.mofa.go.jp/mofaj/gaiko/summit/evian_paris03/af_kigah_z.html`

19. **LMD** : constitué de 15 articles de Le Monde Diplomatique tirés des numéros de janvier, février et mars 2004 (édition informatique disponible sur `http://www.diplo.jp/`) :
    1) *Sous la pression des Églises*, par Christian Terras, janvier 2004.
    2) *Comment se meurent les géants : Seraing repense son avenir sans les hauts-fourneaux*, par Sergio Carrozzo, janvier 2004.
    3) *Une obsession nommée Bombay*, par Mila Khalon, janvier 2004.
    4) *L'histoire populaire des États-Unis : « Un pouvoir que nul ne peut réprimer »*, par Howard Zinn, janvier 2004.
    5) *Création d'un poisson d'aquarium génétiquement modifié : Le sacre des mutants*, par Franck Mazoyer, janvier 2004.
    6) *Exception française*, par Dominique Vidal, février 2004.
    7) *Discours « populiste » et défense des privilèges : Cette Amérique qui vote George W. Bush*, par Tom Frank, février 2004.
    8) *Cogestion de la régression sociale : La défaite programmée des syndicats allemands*, par Udo Rehfeldt, février 2004.
    9) *Vingt-cinq ans après la révolution islamique : Le réveil de l'Iran*, par Bernard Hourcade, février 2004.
    10) *Au Sri Lanka, crise politique et paix en suspens : Un État « de facto » pour les Tigres tamouls*, par Cédric Gouverneur, février 2004.
    11) *Difficile transition pour une Algérie meurtrie*, par Lyes Si Zoubir, mars 2004.
    12) *Les États-Unis pris au piège : Quelle autonomie pour les Kurdes d'Irak ?*, par Michel Verrier, mars 2004.
    13) *UN CASSE-TÊTE POUR L'OCCIDENT : Poussée ultranationaliste en ex-Yougoslavie*, par Jean-Arnault Dérens, mars 2004.
    14) *Le devenir du passé : En Russie, nostalgie soviétique et nouveau patriotisme d'Etat*, par Jean-Marie Chauvier, mars 2004.
    15) *Des dizaines de milliers de réfugiés tenus à distance : L'Europe enterre le droit d'asile*, par Alain Morice, mars 2004.
20. **LMDJP** : constitué des traductions des 15 articles de Le Monde Diplomatique du corpus LMD ci-dessus, disponible sur `http://www.diplo.jp/` (même numérotation que les articles originaux) :
    1) 欧州カトリック勢力のロビー工作
    2) 高炉の火が消えるベルギーの町
    3) ボンベイ、それは

4) 民衆のアメリカ史を記す
5) フランケンフィッシュの物語
6) イスラムのスカーフに対するヨーロッパ諸国の姿勢
7) アメリカのポピュリズム
8) 危地に立つドイツの労働組合
9) イランの覚醒
10) スリランカ和平への長い道のり
11) アルジェリア社会の深い傷跡
12) イラクのクルド地域を歩く
13) 旧ユーゴ諸国に躍る民族主義の影
14) ロシアに広まるソ連ノスタルジー
15) 難民庇護の否定へと繋がるEUの動き

21. **LMDJP2** : article du Monde Diplomatique
   – フランス郊外団地で火を噴いたものは何か (*furansu kougai danchi de hi wo fuita mono wa nani ka*), article d'origine « Révolte des banlieues : Les raisons d'une colère », par Laurent Bonelli, decembre 2005.
22. **Unicode** : version française de la page internet « How to Unicode »
`http://www.freenix.fr/unix/linux/HOWTO/Unicode-HOWTO.html`
23. **UnicodeJP** : version japonaise de la page internet « How to Unicode »
`http://www.linux.or.jp/JF/JFdocs/Unicode-HOWTO.html`
24. **Zadig** : constitué du texte intégral de *Zadig* de Voltaire, réalisé à partir de la version électronique distribuée par Olivier Tableau (disponible sur Internet)
25. **ZadigJP** : traduction japonaise de *Zadig* par Takenori NOUMI, réalisé à partir de la version électronique distribuée sur le site Aozora-bunko :
`http://www.aozora.gr.jp`

Corpus utilisés lors des études linguistiques

26. **Fujiwara** : FUJIWARA, Masahiko. 若き数学者のアメリカ (*wakaki sûgakusha no amerika*), 1953.
27. **LMD0704** : article du Monde Diplomatique et sa traduction japonaise
   – *Développement ne rime pas forcément avec croissance : Vers une société économe et solidaire*, par Jean-Marie Harribey, juillet 2004.
   – 必ずしも発展に成長は必要ない (*kanarazushimo hatten ni seichô wa hitsuyônai*).
28. **Murakami-kaze** : MURAKAMI, Haruki. 風の歌を聴け (*kaze no uta wo kike*). Kodansha, 1982.
29. **Murakami-kokkyo** : MURAKAMI, Haruki. 国境の南、太陽の西 (*kokkyô no minami, taiyô no nishi*). Kodansha, 1995.

30. **Shincho** : 11 romans représentant 66 899 phrases, extraits du CD-ROM « 新潮文庫の100冊 [*Shinchô-bunko no 100 satsu*), 1995, Shinchosha » :
    1) TANIZAKI, Junichiro. 痴人の愛 (*chijin no ai*), 1926.
    2) KAWABATA, Yasunari. 雪国 (*yuki guni*), 1935.
    3) MISHIMA, Yukio. 金閣寺 (*kinkakuji*), 1956.
    4) KAIKO, Ken. パニック・裸の王様 (*panikku, hadaka no ôsama*), 1957.
    5) YOSHIYUKI, Junnosuke. 砂の上の植物群 (*suna no ue no shokubutsu-gun*), 1963.
    6) INOUE, Hisashi. ブンとフン (*bun to fun*), 1970.
    7) WATANABE, Junichi. 花埋み (*hanauzumi*), 1970.
    8) SAWAKI, Kotaro. 一瞬の夏 (*isshun no natsu*), 1981.
    9) corpus Fujiwara.
    10) corpus Tsutsui (cf . ci-dessous).
    11) corpus FdT.
31. **Tsutsui** : TSUTSUI, Yasutaka. エディプスの恋人 (*edipusu no koibito*), Shinchosha, 1977.
32. **Yomiuri** : 241 articles du journal *Yomiuri* représentant 3 237 phrases, version électronique disponible sur `https://db.yomiuri.co.jp/bunshokan/`
    1) 「飛鳥会」横領　駐車場収益を決算書に記載せず　大阪府、近く立ち入り検査, 09/05/06 (éd. Osaka-soir) ;
    2) 大学公開講座に「履修証明」　再就職や転職後押し　政府、来年度にも, 08/05/06 (éd. Tokyo-matin) ;
    3) 年々長期化、８９か月　新薬の治験短縮へ専門家育成を支援／厚労省, 08/05/06 (éd. Tokyo-soir) ;
    4) 大阪市長が人事評価　区長、局長を直接面談　９月から導入, 08/05/06 (éd. Osaka-matin) ;
    5) 大阪市「同和対策事業」で着服　財団理事長を逮捕　２年間で１億５０００万円, 08/05/06 (éd. Osaka-soir) ;
    6) 雇用保険で少子化対策　来年度財源に積立金１０００億円活用／政府検討, 07/05/06 (éd. Tokyo-matin) ;
    7) 韓国の名刹、極彩色壁画　日韓台で合同修復　元興寺研中心、来月にも始動, 07/05/06 (éd. Osaka-matin) ;
    8) 飛鳥会「評議員会」設置せず　公益法人運用指針に違反　大阪府の指導もなし, 14/05/06 (éd. Osaka-matin) ;
    9) 村上ファンド「外資」に　国内顧問業廃業、シンガポールを本拠地に　, 13/05/06 (éd. Tokyo-matin) ;
    10) 外国人に「在留カード」　不法滞在を判別　国が許可・登録を一元管理へ, 13/05/06 (éd. Tokyo-soir) ;
    11) 飛鳥会駐車場　契約内容は大阪市が決定　公社元役員「判押すだけ」, 13/05/06 (éd. Osaka-matin) ;
    12) モノづくりの技を国内へ海外へ　シニアを生かす　職業紹介ＮＰＯを設立／大阪, 13/05/06 (éd. Osaka-soir) ;

13) 別納郵便２７億円不正値引き　業者優遇　長岡郵便局３０人以上処分へ／郵政公社, 12/05/06 (éd. Tokyo-matin) ;
14) 「高松塚」損傷　文化庁に隠ぺい？メモ　報道対応に「自然劣化と説明する」, 12/05/06 (éd. Tokyo-soir) ;
15) 飛鳥会横領事件　小西容疑者着服認める　「個人資金回収」と弁明／大阪府警, 12/05/06 (éd. Osaka-matin) ;
16) 大証に最高規制責任者　上場審査、売買監視を強化, 12/05/06 (éd. Osaka-soir) ;
17) 中央青山、法定監査２か月停止処分　７月から、対象２３００社／金融庁, 11/05/06 (éd. Tokyo-matin) ;
18) 普天間移設　沖縄県が受け入れ　政府案基本に　額賀防衛長官と稲嶺知事合意, 11/05/06 (éd. Tokyo-soir) ;
19) 「飛鳥会」横領事件　公社が収入報告ねつ造　１９年間過少に、押印も大阪市職員, 11/05/06 (éd. Osaka-matin) ;
20) キトラ古墳壁画「白虎」　村民と初対面　はぎ取り修復完了／奈良・飛鳥資料館, 11/05/06 (éd. Osaka-soir) ;
21) 中学校教科書６５冊に誤記２０８か所　校正強化求める／文科省緊急調査, 10/05/06 (éd. Tokyo-matin) ;
22) ライブドア事件　堀江被告側、全面否認「違法性認識なし」　公判前整理手続きで, 10/05/06 (éd. Tokyo-soir) ;
23) 大阪市　不適切「同和事業」、飛鳥会以外に５件　４年で発注８億円, 10/05/06 (éd. Osaka-matin) ;
24) 「飛鳥会」横領　駐車場純利益、年１億円　大阪国税局が税務調査へ, 10/05/06 (éd. Osaka-soir) ;
25) 露サミット　日本式省エネを各国に提案へ　「トップランナー方式」採用促す, 09/05/06 (éd. Tokyo-matin) ;
26) 容疑者取り調べを録画・録音　「裁判員制度」向けに試行　検察当局が方針, 09/05/06 (éd. Tokyo-soir) ;
27) 「飛鳥会」横領　大阪市が駐車場過少申告を“放置”　９１年実態調査、算定, 09/05/06 (éd. Osaka-matin) ;
28) 不明の小１男児、遺体発見　首絞められた跡　自宅１０キロの川沿いで／秋田, 19/05/06 (éd. Tokyo-matin) ;
29) 米産牛肉の輸入７月にも再開　来月決定　日本側３５施設を事前査察, 19/05/06 (éd. Tokyo-soir) ;
30) 京都のパチンコ業者、数十億円所得隠し　容疑で捜索／大阪地検など, 19/05/06 (éd. Osaka-matin) ;
31) 神戸市議汚職　村岡被告長男の市議を逮捕　あっせん収賄容疑／地検, 19/05/06 (éd. Osaka-soir) ;
32) ヒューザー小嶋社長、詐欺容疑で逮捕　耐震偽装隠し販売　木村元社長も再逮捕, 18/05/06 (éd. Tokyo-matin) ;
33) 耐震偽装公表３か月前　木村建設、危険性認識か　子会社に福岡の建築士が指摘, 18/05/06 (éd. Tokyo-soir) ;
34) ［遠近おちこち］備讃瀬戸（３）女木島（連載）, 18/05/06 (éd. Osaka-matin) ;
35) 関大新キャンパス「予定地、無償提供を」　突然申し出に大阪・高槻市「想定外」, 18/05/06 (éd. Osaka-soir) ;
36) 耐震強度計算の新方式　自治体４５％、対応不能／２７６自治体・読売新聞社調査, 17/05/06 (éd. Tokyo-matin) ;
37) ヒューザー小嶋社長逮捕へ　耐震偽装、詐欺容疑　木村元社長も再逮捕へ, 17/05/06 (éd. Tokyo-soir) ;

38) 飛鳥会理事長への融資５０億円、ほぼ不良債権化　資産４０億円分散？つかめず, 17/05/06 (éd. Osaka-matin) ;

39) サッカー独Ｗ杯　日本代表２３人が決定　経験重視、充実の布陣＝号外も発行, 16/05/06 (éd. Tokyo-matin) ;

40) 韓国民団と朝鮮総連、和解へ　あすにも初トップ会談　南北融和ムード反映, 16/05/06 (éd. Tokyo-soir) ;

41) 大阪府、所管の全公益法人検査へ　飛鳥会事件の無届け収益事業発覚で　, 16/05/06 (éd. Osaka-matin) ;

42) 高松塚壁画損傷　文化庁と東文研対立が背景に？　引き継ぎ中に事故　, 16/05/06 (éd. Osaka-soir) ;

43) 領事館「職務重圧で自殺」　上海警察に説明、署名　脅迫否定の中国側根拠に, 15/05/06 (éd. Tokyo-matin) ;

44) 米軍再編特措法案の概要判明　新交付金で地域振興　閣僚会議も新設, 15/05/06 (éd. Tokyo-soir) ;

45) 信楽鉄道事故、風化させないで　１５年追悼法要　「尼崎事故」遺族が初参列, 15/05/06 (éd. Osaka-matin) ;

46) 飛鳥会事件　大阪市職員、理事長親族を１０年以上“専従介護”　上司は黙認, 15/05/06 (éd. Osaka-soir) ;

47) 厚生年金加入「週２０時間労働で」　パートの範囲と義務拡大へ／政府方針, 14/05/06 (éd. Tokyo-matin) ;

48) センター試験　受験番号の記入ミス、２３年間救済　今春は７０００件, 25/05/06 (éd. Tokyo-matin) ;

49) ライブドア粉飾事件　宮内被告、堀江被告の発言供述「その金額、予算に乗せて」, 25/05/06 (éd. Tokyo-soir) ;

50) ［私たちは何を学んだか］（１）付属池田小副校長・津田一司さん４６歳（連載）, 25/05/06 (éd. Osaka-soir) ;

51) 大手銀６グループの０６年３月期決算　最高益、３兆円超す　バブル期の１．７倍, 24/05/06 (éd. Tokyo-matin) ;

52) 野菜じわり高騰　前線が列島に停滞　日照不足は記録的, 24/05/06 (éd. Tokyo-soir) ;

53) 年金保険料無断免除　大阪社保事務局長を更迭　厚労相指示「監督責任重い」, 24/05/06 (éd. Osaka-matin) ;

54) 飛鳥会事件　関連の社福法人に大阪府有地を格安で貸与　２６年間“放置”, 24/05/06 (éd. Osaka-soir) ;

55) 社会保障負担、７３％増　財源議論必要に／２０２５年度厚労省推計, 23/05/06 (éd. Tokyo-matin) ;

56) 汚泥談合　公取委が１１社告発　幹事社部長級ら、一斉取り調べ／大阪地検, 23/05/06 (éd. Tokyo-soir) ;

57) 神鋼ばい煙データ改ざん　５年で１５０時間以上　基準超える排出, 23/05/06 (éd. Osaka-matin) ;

58) トヨタ、１０工場を海外に新設　２０１０年までに　年１０００万台体制へ, 22/05/06 (éd. Tokyo-matin) ;

59) 国家公務員の再就職に初の支援ビジネス　パソナ、今秋から　「削減」論議も背景, 22/05/06 (éd. Tokyo-soir) ;

60) 脱線の教訓刻む資料館　福知山線事故の車両や遺品展示／ＪＲ西日本, 22/05/06 (éd. Osaka-matin) ;

61) 市有地、無契約状態に　大阪市、５４か所で　「無償」再提案へ　, 22/05/06 (éd. Osaka-soir) ;

62) イラク本格政府発足　治安閣僚決まらず　フセイン政権崩壊３年、民主主義体制に, 21/05/06 (éd. Tokyo-matin) ;

63) 神戸市議汚職　市の内部調査　不適切対応認める　産廃施設設置巡り　, 21/05/06 (éd. Osaka-matin) ;

64) 地震保険料、平均７．７％下げ　予測データを４０年ぶり見直し, 20/05/06 (éd. Tokyo-matin) ;

65) イラク陸自、撤収開始来月にも　治安権限など、地元移譲にメドで, 20/05/06 (éd. Tokyo-soir) ;

66) 村上ファンド、阪神総会に「検査役」請求　大阪地裁が受理, 20/05/06 (éd. Osaka-matin) ;

67) 昨秋の国勢調査　大阪・中央区で未回収３３％　, 20/05/06 (éd. Osaka-soir) ;

68) サッカーW杯　中国ルートのチケット届かず　１２００人、ツアー中止, 31/05/06 (éd. Tokyo-soir) ;

69) 三洋電機、井植元会長の退職金見送り　在任４５年でも株主の理解得られぬ, 31/05/06 (éd. Osaka-matin) ;

70) 飛鳥会横領　小西被告、独断で４７億円融資枠　福祉法人から暴力団に流れる？, 31/05/06 (éd. Osaka-soir) ;

71) 年金不適正手続き　２６都府県１１万件超す　２７０万件、再確認へ／社保庁調査, 30/05/06 (éd. Tokyo-matin) ;

72) 在日米軍再編　基本方針を閣議決定　「普天間」移設、地名記載せず, 30/05/06 (éd. Tokyo-soir) ;

73) 「飛鳥会」横領事件　小西被告特養用地　大阪市が９４年に購入・無償貸与, 30/05/06 (éd. Osaka-matin) ;

74) 関空滑走路、全面改修　アスファルト上塗り　２４時間運用は維持、来秋にも着工, 30/05/06 (éd. Osaka-soir) ;

75) インドネシア・ジャワ島地震　死者４６００人超す　救助活動を本格化, 29/05/06 (éd. Tokyo-matin) ;

76) 阪急が阪神株をＴＯＢ　村上氏、応諾検討　決議で１株９３０円、１９日期限, 29/05/06 (éd. Tokyo-soir) ;

77) 阪神株買い取り　価格差「数十円に」　阪急と村上氏側、交渉なお継続, 29/05/06 (éd. Osaka-matin) ;

78) ジャワ島地震、死者４９８３人　非常事態宣言／インドネシア政府, 29/05/06 (éd. Osaka-soir) ;

79) インドネシア・ジャワ島地震　死者３０００人　負傷数千人、２０万人が家失う, 28/05/06 (éd. Tokyo-matin) ;

80) ライブドア初公判　粉飾、堀江被告が指示　宮内被告が不正認める／東京地裁, 27/05/06 (éd. Tokyo-matin) ;

81) ［土曜ナビ］老いも若きも「脳」トレ　知育教育、携帯ゲーム、高齢者講座…, 27/05/06 (éd. Tokyo-soir) ;

82) 大阪・和泉の市営住宅で火事　留守番３兄妹死亡　母の携帯に助け求め, 27/05/06 (éd. Osaka-matin) ;

83) ニート６割、部活未経験　希薄な社会性が未就労の原因／読売新聞ネット調査, 26/05/06 (éd. Tokyo-matin) ;

84) ４月消費者物価指数　６か月連続で前年比プラス　「脱デフレ」鮮明に＝訂正あり, 26/05/06 (éd. Tokyo-soir) ;

85) 村上氏・阪急ＨＤ交渉　阪神の大幅増配案浮上　１０倍程度　株売却価格は下げ, 26/05/06 (éd. Osaka-matin) ;

86) 石綿被害、国を集団提訴　大阪・泉南の住民ら８人、２億４０００万円賠償請求, 26/05/06 (éd. Osaka-soir) ;

87) 村上代表を逮捕　ニッポン放送株巡るインサイダー取引容疑＝号外も発行, 06/06/06 (éd. Tokyo-matin) ;

88) 秋田の小１殺害　畠山容疑者「首絞め殺害」供述　死体遺棄も１人で, 06/06/06 (éd. Tokyo-soir) ;

89) 阪急百貨店、阪神百貨店統合を検討　来月以降、価格交渉へ, 06/06/06 (éd. Osaka-matin) ;

90) 「明日の神話」よみがえった　岡本太郎さん作の巨大壁画、修復終了／愛知・東温, 06/06/06 (éd. Osaka-soir) ;

91) 秋田小１殺害　水死女児の母親逮捕　死体遺棄容疑認める　「殺人はやってない」, 05/06/06 (éd. Tokyo-matin) ;

92) インサイダー認める　村上氏ら午後逮捕　ライブドア意向認識後、株買い増す, 05/06/06 (éd. Tokyo-soir) ;

93) 小１男児殺害　団地内の女性宅捜索へ　不明直後に車トランク閉める／秋田県警, 04/06/06 (éd. Tokyo-matin) ;

94) 火に願う、山の安全／鳥取・大山, 04/06/06 (éd. Osaka-matin) ;

95) 村上ファンド立件へ　ニッポン放送株売買で証取法違反容疑, 03/06/06 (éd. Tokyo-matin) ;

96) 村上氏を任意聴取　「インサイダー」本格捜査へ／東京地検, 03/06/06 (éd. Tokyo-soir) ;

97) 大阪市の同和行政見直し　施設の職員加配を是正　関市長が表明, 03/06/06 (éd. Osaka-matin) ;

98) 村上ファンド　阪神株ＴＯＢに応募か　「１週間後、態度表明」, 03/06/06 (éd. Osaka-soir) ;

99) 村上ファンド幹部聴取　ニッポン放送株の購入巡り　東京地検、任意で, 02/06/06 (éd. Tokyo-matin) ;

100) インサイダー疑惑　村上氏を近く聴取へ　発端はライブドア情報？／東京地検, 02/06/06 (éd. Tokyo-soir) ;

101) インサイダー疑惑　「村上ファンド」関連銘柄軒並み下落, 02/06/06 (éd. Osaka-soir) ;

102) 少子化対策　０～２歳の児童手当増額　企業の育児支援公開／政府・与党案概要, 01/06/06 (éd. Tokyo-matin) ;

103) 駐車違反の民間監視スタート　即摘発、ピリピリ　繁華街で効果あり, 01/06/06 (éd. Tokyo-soir) ;

104) プロ野球・阪神タイガースの新オーナーに宮崎恒彰氏, 01/06/06 (éd. Osaka-matin) ;

105) ［眼］島で産みたい　産婦人科医不在の隠岐の島　妊婦７０人、本土出産, 01/06/06 (éd. Osaka-soir) ;

106) 「教育基本法」先送り　会期大幅延長せず　改正は小泉後継で／政府・与党, 31/05/06 (éd. Tokyo-matin) ;

107) 平日２２時まで対応　小児救急医療センター設置／東京・荒川区, 06/06/06 (éd. Tokyo-matin) ;

108) 総領事館員自殺問題　日本側、中国に改めて抗議／日中協議, 06/06/06 (éd. Tokyo-matin) ;

109) 対中円借款の凍結解除、きょう決定　２００５年度分「１００億円減」, 06/06/06 (éd. Tokyo-matin) ;

110) 対中円借款、７４０億円決定　中川農相は異論, 06/06/06 (éd. Tokyo-soir) ;

111) 日中外相会談　７月、マレーシアで／麻生外相, 05/06/06 (éd. Tokyo-matin) ;
112) 日中外相会談　７月末開催で調整, 04/06/06 (éd. Tokyo-matin) ;
113) 中韓との関係、次期首相で改善／森前首相, 03/06/06 (éd. Tokyo-soir) ;
114) 安倍官房長官　靖国参拝、明言せず　総裁選公約で触れない方針, 03/06/06 (éd. Tokyo-soir) ;
115) 島大、中国・寧夏大に共同研究所長ら２人派遣　森林復活事業など指導＝島根, 03/06/06 (éd. Osaka-matin) ;
116) ミャンマー情勢で国連安保理決議　米要求、日中露は反対, 02/06/06 (éd. Tokyo-matin) ;
117) ［国家戦略を考える］第５部（８）「管理」進める世界に背（連載）, 01/06/06 (éd. Tokyo-matin) ;
118) ［夕景時評］古代に導く「小石丸」　編集委員・井上茂男, 01/06/06 (éd. Tokyo-soir) ;
119) ［ペット最新事情］（２）散歩必携ペットボトル　おしっこを洗い流す（連載）, 31/05/06 (éd. Tokyo-matin) ;
120) ［国家戦略を考える］第５部（７）対中配慮、資源開発遅れ（連載）, 31/05/06 (éd. Tokyo-matin) ;
121) 中国から省エネ視察　東ソーなど３企業を訪問＝山口, 31/05/06 (éd. Ouest-matin) ;
122) 福岡の五輪理念を分かってもらえた　山崎市長、北京から帰国＝福岡, 31/05/06 (éd. Ouest-matin) ;
123) トラフズク、子育て中　大仙の民家のモミの木で＝秋田, 30/05/06 (éd. Tokyo-matin) ;
124) 公開シンポジウム「東アジアの平和と繁栄に向けて」＝特集, 30/05/06 (éd. Tokyo-matin) ;
125) 安倍官房長官「父親超えるよう努力したい」, 30/05/06 (éd. Tokyo-matin) ;
126) ワールドインテック、技術者派遣など中国企業と提携, 30/05/06 (éd. Ouest-matin) ;
127) 耐震偽装　姉歯被告を偽証立件へ　国会証人喚問「木村建設から圧力」証言で, 06/06/06 (éd. Tokyo-soir) ;
128) 耐震偽装発覚後　６か月ぶり営業再開　岡崎のサンホテル, 06/06/06 (éd. Centre-matin) ;
129) 耐震偽装　市川市、改修費助成しない方針　強度７３％の「ＧＳ下総中山」＝千葉, 05/06/06 (éd. Tokyo-matin) ;
130) ［会う聞く話す］県建築物安全推進室長・生田博隆さん＝熊本, 04/06/06 (éd. Ouest-matin) ;
131) 〈解〉日本建築構造技術者協会, 02/06/06 (éd. Tokyo-matin) ;
132) 耐震偽装　マンション管理組合連合会、「使用禁止」の住民に義援金, 01/06/06 (éd. Tokyo-soir) ;
133) 耐震偽装木村関与　八代のホテル「安全」　県内検証終了、強度不足は４棟＝熊本, 01/06/06 (éd. Ouest-matin) ;
134) 耐震偽装のサンホテル大和郡山　県が改修計画承認＝奈良, 31/05/06 (éd. Osaka-matin) ;
135) ［マンション快適ライフ］耐震強度（９）業者と説明会　住民で委員会, 30/05/06 (éd. Tokyo-matin) ;
136) 耐震偽装　ＧＳ稲城で検証開始　退去から４か月、新局面＝多摩, 30/05/06 (éd. Tokyo-matin) ;
137) 耐震偽装　イーホームズの「指定」を取り消し／国交省, 30/05/06 (éd. Tokyo-matin) ;

138) 浅沼物件の強度不足　滝川のマンションでも　千歳の物件は数値改ざん＝北海道, 30/05/06 (éd. Tokyo-soir) ;

139) 姉歯被告、国会証人喚問での「証言ウソ」　「ＧＳ池上」前に偽装, 29/05/06 (éd. Tokyo-soir) ;

140) 欠陥住宅被害の全国大会、静岡で始まる　法整備求める＝静岡, 28/05/06 (éd. Tokyo-matin) ;

141) 耐震偽装　信頼回復へ建築塾＝大分, 28/05/06 (éd. Ouest-matin) ;

142) 賃貸マンションの構造計算に不正　札幌の不動産業者、建て直し＝北海道, 27/05/06 (éd. Tokyo-matin) ;

143) 県情報公開制度　請求件数、過去最多の２万件超え　背景に耐震偽装か＝神奈川, 27/05/06 (éd. Tokyo-matin) ;

144) 耐震偽装　前橋のホテル、賠償額に風評被害も＝群馬, 27/05/06 (éd. Tokyo-matin) ;

145) 東京２３区の特別区長会、国に支援策を要請　イーホームズ廃業で, 26/05/06 (éd. Tokyo-matin) ;

146) サマワで自衛隊攻撃事件　主犯格容疑認める／イラク, 06/06/06 (éd. Tokyo-matin) ;

147) バグダッドでバス乗客ら５０人を拉致／イラク, 06/06/06 (éd. Tokyo-matin) ;

148) 緊迫の街バスラ　英軍パトロールに同行／イラク, 06/06/06 (éd. Tokyo-matin) ;

149) イラクのロシア大使館員殺害　国連が非難, 05/06/06 (éd. Tokyo-matin) ;

150) イラクでバス襲撃、乗客２１人射殺　シーア派とクルド人標的, 05/06/06 (éd. Tokyo-matin) ;

151) イラク・サマワで５００人デモ　政府施設を襲撃, 05/06/06 (éd. Tokyo-matin) ;

152) イラク陸自撤収時期調整へ　空自輸送範囲拡大も／額賀長官意向, 05/06/06 (éd. Tokyo-matin) ;

153) イラク駐留２８か国、情勢を注視, 05/06/06 (éd. Tokyo-matin) ;

154) イラク・バスラ　警察腐敗、治安最悪　英軍撤退出口見えず, 05/06/06 (éd. Tokyo-matin) ;

155) 陸自のイラク支援　第９次群が全員帰国, 04/06/06 (éd. Tokyo-matin) ;

156) バスラの市場で爆発、１５人死亡／イラク, 04/06/06 (éd. Tokyo-matin) ;

157) イラク米軍虐殺疑惑　ブッシュ政権、対応に忙殺, 04/06/06 (éd. Tokyo-matin) ;

158) 米軍虐殺疑惑　イラクが独自捜査へ　「解明、米に任せるな」　国民の怒り増幅, 04/06/06 (éd. Tokyo-matin) ;

159) ロシア人外交官、殺害される／バグダッド, 04/06/06 (éd. Tokyo-matin) ;

160) 非常事態宣言下、バスラ警戒強化／イラク, 04/06/06 (éd. Tokyo-matin) ;

161) イラク駐留米兵に特別訓練, 03/06/06 (éd. Tokyo-matin) ;

162) イラク・サマワに異例の夜間外出禁止令, 03/06/06 (éd. Tokyo-matin) ;

163) イラク・サマワで陸自への不満募る　「２年たっても生活よくならない」, 03/06/06 (éd. Tokyo-matin) ;

164) 治安回復へ“切り札”　イラク陸軍に新型装甲車, 02/06/06 (éd. Tokyo-matin) ;

165) 陸自攻撃の３容疑者逮捕　サドル派と接点も／イラク当局, 02/06/06 (éd. Tokyo-matin) ;

166) 衆院比例ブロック大会　総裁選告示前の開催を決定／自民, 06/06/06 (éd. Tokyo-matin) ;

167) 自民総裁選　福田氏「あんまり引っ張られては困る」, 06/06/06 (éd. Tokyo-matin) ;

168) 国会運営で小泉首相批判相次ぐ　「重要案件が山積」／自民・古賀氏, 06/06/06 (éd. Tokyo-matin) ;

169) 自民党総裁選　候補一本化に否定的／安倍氏, 05/06/06 (éd. Tokyo-matin) ;

170) 小泉改革「流れ止める勇気も」／自民・古賀氏, 04/06/06 (éd. Tokyo-matin) ;

171) 自民総裁選　額賀氏、出馬に「ノーコメント」, 04/06/06 (éd. Tokyo-matin) ;

172) 「国会延長せず」安倍氏のためだった？　小泉首相の“翻意”片山氏が明かす, 04/06/06 (éd. Tokyo-matin) ;

173) 自民総裁選　安倍氏が実質“始動”　相次ぎＴＶ出演　「靖国」深入りせぬ戦略, 04/06/06 (éd. Tokyo-matin) ;

174) 再チャレンジ議連、安倍応援団でない／森前首相, 04/06/06 (éd. Tokyo-matin) ;

175) 自民総裁選　森氏、深まる苦悩　小泉首相に不快感も　, 04/06/06 (éd. Tokyo-matin) ;

176) 自民・再チャレンジ議連に９４人　総裁選で安倍氏支持の中核組織？, 03/06/06 (éd. Tokyo-matin) ;

177) 自民総裁選　候補者一本化論が相次ぐ／森派幹部会, 02/06/06 (éd. Tokyo-matin) ;

178) 小泉首相、会期延長を強く否定　自民、法案先送りに不満　口出し「おかしい」, 02/06/06 (éd. Tokyo-matin) ;

179) 自民総裁選　「世代間抗争」の気配　年配議員の会は福田氏支持？, 01/06/06 (éd. Tokyo-matin) ;

180) 自民総裁選　森派あいまい政策提言　候補２人で“苦悩”　見解並べただけ, 31/05/06 (éd. Tokyo-matin) ;

181) 自民総裁選　福田氏と安倍氏、消費税上げで対立鮮明　靖国問題でも, 30/05/06 (éd. Tokyo-matin) ;

182) 自民総裁選　与謝野経財相、出馬検討に含み, 29/05/06 (éd. Tokyo-matin) ;

183) 自民総裁選　安倍氏、派内で優勢　森前首相、「福田氏一本化」に期待感も, 29/05/06 (éd. Tokyo-matin) ;

184) ライブドア初公判　「堀江主犯」の構図、検察の立証に注目（解説）, 27/05/06 (éd. Tokyo-matin) ;

185) ライブドア初公判　粉飾「何とかなるでしょ」　堀江被告、部下に要求強める, 27/05/06 (éd. Tokyo-matin) ;

186) ライブドア事件　民事で追及続々と　個人株主訴訟、５０００人規模か, 27/05/06 (éd. Tokyo-matin) ;

187) ライブドア粉飾事件　宮内被告、堀江被告の発言供述「その金額、予算に乗せて」, 25/05/06 (éd. Tokyo-soir) ;

188) ライブドア事件　堀江被告の保身に嫌気、宮内被告が“決別”決意, 23/05/06 (éd. Tokyo-matin) ;

189) ライブドア事件・第１回公判前整理手続き　迅速裁判モデルケースに（解説）, 10/05/06 (éd. Tokyo-soir) ;

190) ライブドア事件・第１回公判前整理手続き　“前哨戦”早くも激突, 10/05/06 (éd. Tokyo-soir) ;

191) ライブドア事件　堀江被告側、全面否認「違法性認識なし」　公判前整理手続きで, 10/05/06 (éd. Tokyo-soir) ;

192) ライブドア・堀江被告保釈　「公判前整理」で９４日の早期実現（解説）, 28/04/06 (éd. Tokyo-matin) ;

193) ライブドア・堀江被告保釈　拘置中、一心に公判対策　８キロ減「ベスト体重」, 28/04/06 (éd. Tokyo-matin) ;

194) ライブドア証取法違反　堀江被告保釈認める　保釈金３億、検察準抗告／東京地裁, 27/04/06 (éd. Tokyo-matin) ;

195) ライブドア・堀江被告　５月１０日に公判前手続き　, 26/04/06 (éd. Tokyo-soir) ;

196) 使命忘れた監査法人　後絶たぬなれ合い　中小も体制強化を（解説）, 06/04/06 (éd. Tokyo-matin) ;

197) ライブドア粉飾事件　熊谷被告を保釈, 06/04/06 (éd. Tokyo-matin) ;

198) ライブドア事件　会計士２人を告発／証券取引等監視委, 31/03/06 (éd. Tokyo-matin) ;

199) ライブドア粉飾決算　２会計士、０３年も黙認　発覚恐れ翌年「適正」、在宅起訴, 31/03/06 (éd. Tokyo-soir) ;

200) ライブドア粉飾決算　一両日中に会計士を告発／証券取引等監視委, 30/03/06 (éd. Tokyo-matin) ;

201) ライブドア粉飾決算　２会計士、適正意見拒む　監査責任者ら午後告発, 30/03/06 (éd. Tokyo-soir) ;

202) 港陽監査法人が解散発表　虚偽許されぬ退場劇（解説）, 28/03/06 (éd. Tokyo-matin) ;

203) 「港陽監査法人」６月解散方針　ライブドア事件の会計士所属, 26/03/06 (éd. Tokyo-matin) ;

204) 村上代表逮捕　与党、投資ルール厳格化を強調　野党「小泉政治が問題」と批判, 06/06/06 (éd. Tokyo-matin) ;

205) 村上代表逮捕、こう見る　吉見俊彦氏、郷原信郎氏, 06/06/06 (éd. Tokyo-matin) ;

206) ［崩壊・マネーゲーム］（１）村上代表逮捕　モノ言う株主、誤算（連載）, 06/06/06 (éd. Tokyo-matin) ;

207) 村上代表逮捕　強気一転「罪は私に」　側近は「旧友」　大手証券と警察官僚出身, 06/06/06 (éd. Tokyo-matin) ;

208) 村上ファンドの証取法違反　「行き過ぎ」「遺憾」各界から相次ぎ批判, 06/06/06 (éd. Tokyo-matin) ;

209) 村上ファンド事件　「自覚ない行動　反省されるべき」／与謝野金融相, 06/06/06 (éd. Tokyo-matin) ;

210) 村上代表逮捕　日枝・フジテレビ会長「私は守旧派、２人は英雄視されたが…」, 06/06/06 (éd. Tokyo-matin) ;

211) 村上ファンドのインサイダー取引事件　シンガポール、調査に乗り出す方針, 06/06/06 (éd. Tokyo-matin) ;

212) 「村上ファンド」の証取法１６７条違反　Ｍ＆Ａ巡る情報交換に警鐘, 06/06/06 (éd. Tokyo-matin) ;

213) 村上代表逮捕　ファンドの投資先企業、株放出に警戒感, 06/06/06 (éd. Tokyo-matin) ;

214) 村上代表逮捕　４０００億円ファンド崩壊危機　資金引き揚げ加速へ, 06/06/06 (éd. Tokyo-matin) ;

215) ［社説］村上代表逮捕　“プロ”を狂わせた市場原理主義　, 06/06/06 (éd. Tokyo-matin) ;

216) 村上代表逮捕　宮内供述で疑惑浮上　ニッポン放送株取得「村上代表が持ちかけ」, 06/06/06 (éd. Tokyo-matin) ;

217) 村上代表逮捕　海外でも速報, 06/06/06 (éd. Tokyo-matin) ;

218) 村上代表逮捕　ファンドの暴走、歯止めを　経済部長・杉山美邦, 06/06/06 (éd. Tokyo-matin) ;

219) ［転落・村上ファンド］（上）市場監視強める検察（連載）, 06/06/06 (éd. Tokyo-matin) ;

220) 村上代表を逮捕　ニッポン放送株巡るインサイダー取引容疑＝号外も発行, 06/06/06 (éd. Tokyo-matin) ;

221) 村上ファンド、開示遅らせる株売買　報告ルールの甘さ利用, 06/06/06 (éd. Tokyo-soir) ;

222) ［顔］水中写真家として５０年　舘石昭さん, 06/06/06 (éd. Tokyo-matin) ;

223) ［顔］全国身体障害者野球で１０連覇を果たした監督　岩崎廣司さん５６, 05/06/06 (éd. Tokyo-matin) ;

224) ［顔］第５２回江戸川乱歩賞に決まった　鏑木蓮さん, 03/06/06 (éd. Tokyo-matin) ;

225) ［顔］出版文化産業振興財団理事長に就任した　肥田美代子さん, 02/06/06 (éd. Tokyo-matin) ;

226) ［顔］ＷＨＯ西太平洋地域事務局で活躍する　葛西健さん, 01/06/06 (éd. Tokyo-matin) ;

227) ［顔］院内コンサート活動を続ける医師　上杉春雄さん, 31/05/06 (éd. Tokyo-matin) ;

228) ［ひと］現役教諭、優しい味わいの新作　作家・瀬尾まいこさん, 30/05/06 (éd. Tokyo-matin) ;

229) ［ひと］俳誌「波」３０周年　全国に広げた裾野　俳人・倉橋羊村さん, 30/05/06 (éd. Tokyo-matin) ;

230) ［顔］画家バルテュスの夫人　節子・クロソフスカ・ド・ローラさん, 30/05/06 (éd. Tokyo-matin) ;

231) ［顔］認知症の早期診断装置を開発した　武者利光さん, 27/05/06 (éd. Tokyo-matin) ;

232) ［顔］弁護士から最高裁判事に２５日、就任した　那須弘平さん　, 26/05/06 (éd. Tokyo-matin) ;

233) ［顔］世界大会に出場する柏レイソルＵ―１２監督　酒井直樹さん, 25/05/06 (éd. Tokyo-matin) ;

234) ［顔］パリの人気ジャズクラブで日本語の歌を披露した　小野寺あやのさん, 24/05/06 (éd. Tokyo-matin) ;

235) ［顔］映画「典子は、今」に主演、講演・執筆活動を始めた　白井のり子さん, 23/05/06 (éd. Tokyo-matin) ;

236) ［顔］初優勝の白鵬を育てた前師匠　熊ヶ谷親方, 22/05/06 (éd. Tokyo-matin) ;

237) ［顔］「地域創造ネット」専務理事になる　田中尚輝さん, 20/05/06 (éd. Tokyo-matin) ;

238) ［顔］第１９回三島由紀夫賞に決まった　古川日出男さん, 19/05/06 (éd. Tokyo-matin) ;

239) ［ひと］本に命を吹き込む書評　国際日本文化研究センター助教授・池内恵さん, 17/05/06 (éd. Tokyo-matin) ;

240) ［顔］６月に本格再開する「小千谷闘牛」の若手リーダー　平沢隆一さん, 17/05/06 (éd. Tokyo-matin) ;

241) ［顔］日韓大会に続き独Ｗ杯で主審を務める　上川徹さん, 16/05/06 (éd. Tokyo-matin).

BIBLIOGRAPHIE

ABEILLÉ, A., & CLEMENT, L. (2003). Annotation morpho-syntaxique : Les mots simples - les mots composés, corpus *Le Monde*. `www.llf.cnrs.fr/fr/Abeille/guide-morpho-synt.02.pdf`.

ABEILLÉ, A., CLEMENT, L., & REYES, R. (1998). TALANA annotated corpus : first results. In *Proceedings of the First Conference on Linguistic Resources*, (pp. 992–999).

ABNEY, S. (1990). Rapid incremental parsing with repair. In *Proceedings of the 6th New OED Conference*. University of Waterloo, Waterloo, Ontario.
URL `citeseer.ist.psu.edu/abney90rapid.html`

ABNEY, S. (1997). Part-of-speech tagging and partial parsing. In K. Church, S. Young, & G. Bloothooft (Eds.) *Corpus-Based Methods in Language and Speech*. Kluwer Academic Publishers.

AKIHIRO, H. (2004). *Contribution à l'étude de la valence verbale en français contemporain, la non réalisation du complément d'objet direct*. Thèse de doctorat, École Pratique des Hautes Études.

ALFONSO, A. (1966). *Japanese Language Patterns*, vol. 1. Tokyo : Sophia University LL Center of Applied Linguistics.

BALLY, C. (1965). *Linguistique générale et linguistique française*. Berne : Francke, quatrième ed.

BEARTH, T. (2003). *Glossaire français-anglais de terminologie linguistique*. `http://www.sil.org/linguistics/glossary_fe/index.aspl`. Site internet, SIL International.

BISKRI, I., & DESCLÉS, J.-P. (2005). Analyse de la coordination et de la subordination au moyen de la grammaire catégorielle combinatoire applicative. In *Colloque Subordination-Coordination*. `http://www.cavi.univ-paris3.fr/ilpga/colloque-coord-subord-2005/pre-textes/index.html`.

BLANC, O., CONSTANT, M., & WATRIN, P. (2007). Segmentation en super-chunks. In *Actes de la TALN*, (pp. 33 – 42).

BLANCHE-BENVENISTE, C. (2000). *Approche de la langue parlée en français*. Essentiel français. Gap, Paris : Ophrys.

BLANCHE-BENVENISTE, C., BILGER, M., ROUGET, C., & VAN DEN EYNDE, K. (1990). *Le français parlé : Études grammaticales*. Sciences du langage. Paris : Éditions du centre national de la recherche scientifique.

BLOCH, B. (1946). Studies in colloquial japanese, part i, inflection. *Journals of the American Oriental Society*, (66). Repris dans ブロック日本語論考 [burokku nihongo ronkô] publié en 1975 par Kenkyûsha.

BLOOMFIELD, L. (1970). *le Langage*. Paris : Payot.

BOITET, C. (2000). Traduction assistée par ordinateur. In J.-M. Pierrel (Ed.) *Ingénierie des langues*, chap. 12, (pp. 271 – 291). Paris : HERMES.

BONNOT, C. (1999). Pour une définition de la notion de thème (russe moderne). In *La thématisation dans les langues*. Rennes. Actes du colloque de Caen (9-11 octobre 1997).

BORIN, L. (1998). Linguistics isn't always the answer : Word comparaison in computational linguistics. In *The 11th Nordic Conference on Computational Linguistics (NODALIDA)*, (pp. 140 –151).

BOURIGAULT, D. (1992). Surface grammatical analysis for the extraction of terminological noun phrases. In *Proceedings of the 14th International Conference on Computational Linguistics (COLING'92)*, (pp. 977– 981).

BOURIGAULT, D. (1996). Lexter, a natural language processing tool for terminology extraction. In *Proceedings of the 7th EURALEX International Congress*.

BOURIGAULT, D. (2002). UPERY : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus. In *Actes de la 9ème conffrence annuelle sur le Traitement Automatiques des Langues (TALN 2002)*, (pp. 75–84).

BOURIGAULT, D., FABRE, C., FRÉROT, C., JACQUES, M.-P., & OZDOWSKA, S. (2005). Syntex, analyseur syntaxique de corpus. In *TALN 2005*.

BOUTSIS, S., & PIPERIDIS, S. (1998). Aligning clauses in parallel texts. In *Proceedings of the 3rd Conference on Empirical Methods in Natural Language Processing*, (pp. 17 – 26).

BROWN, P. F., COCKE, J., DELLA PIETRA, S. A., DELLA PIETRA, V. J., JELINEK, F., LAFFERTY, J. D., MERCER, R. L., & ROOSSIN, P. S. (1990). A statical approach to machine translation. *Computational Linguistics*, *16*(2), 79–81.

BROWN, P. F., DELLA PIETRA, S. A., DELLA PIETRA, V. J., & MERCER, R. L. (1993). The mathematics of statistical machine translation : Parameter estimation. *Computational Linguistics*, *19*(2), 263–311.

BROWN, P. F., LAI, J. C., & MERCER, R. L. (1991). Aligning sentences in parallel corpora. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, (pp. 169 – 176).

CARPENTIER, F.-G. (2005). *Introduction aux analyses multidimensionnelles*. `http://geai.univ-brest.fr/~carpenti/`.

CARRERAS, X., & LLUÍS, M. (2001). Boosting trees for clause splitting. In *Proceedings of CoNLL-2001*.

CHAFE, W. (1970). *Meaning and the structure of language*. Chicago : University of Chicago Press.

CHAFE, W. (1976). Givenness, contrastiveness, definiteness, subjects, topics, and point of view. In *Subject and topic*, (pp. 25 – 55). New York : Academic Press Inc.

CHAROLLES, M. (1997). L'encadrement du discours : univers, champs, domaines et espaces. *Cahiers de recherche linguistique, 6*, 1 – 73.

CHAROLLES, M. (2003). De la topicalité des adverbiaux détachés en tête de phrase. *Travaux de Linguistique, 47*, 11–51.

CHEN, S. F. (1993). Aligning sentences in bilingual corpora using lexical information. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, (pp. 9–16).

CHENON, C. (2005). *Une meilleure utilisabilité des mémoire de traduction fondée sur un alignement sous-phrastique*. Thèse de doctorat en informatique, Université de Joseph Fourier Grenoble I.

CHEVALIER, J.-C., BLANCHE-BENVENISTE, C., ARRIVÉ, M., & PEYTARD, J. (1964). *Grammaire du français contemporain*. Paris : Larousse.

CHIANG, D. (2005). A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, (pp. 263–270).

CHURCH, K., DAGAN, I., GALE, W., FUNG, P., HELFMAN, J., & SATISH, B. (1993). Aligning parallel texts : do methodes developed for english-french generalize to asian languages ? In *Proceedings of the Pascific Asia Conference on Formal and Computational Linguistics*.

CHURCH, K. W. (1988). A stochastic parts program and noun phrase parser for unrestricted text. In *Proceedings of the Second ACL Conference on Applied Natural Language Processing*.

CHURCH, K. W. (1993). Char_align : a program for aligning parallel texts at the character level. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, (pp. 1 – 8).

COLLIER, N., HIRAKAWA, H., & KUMANO, A. (1998). Machine translation vs. dictionary term translation – a comparison for english-japanese news article alignment. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL 98)*, (pp. 263 – 267).

COLLIER, N., & TAKAHASHI, K. (1995). Sentence alignment in parallel corpora : The asahi corpus of newspaper editorials. Tech. Rep. 95/11, Centre for Computational Linguistics (CCL/UMIST).

COMBETTES, B. (1998). *Les constructions détachées en français*. Gap, Paris : Ophrys.

COVINGTON, M. A. (1996). An algorithm to align words for historical comparison. *Computational Linguistics*, *22*(4), 481–496.

CULIOLI, A. (1999). *Pour une linguistique de l'énonciation*, vol. 2. Paris/Gap : Ophrys.

CULIOLI, A., & DESCLÉS, J.-P. (1982a). Traitement formel des langues naturelles. 1ère partie : mise en place des concepts à partir d'exemples. *Mathématiques et sciences humaines*, *77*, 93–125.

CULIOLI, A., & DESCLÉS, J.-P. (1982b). Traitement formel des langues naturelles. 2ème partie : dérivations d'exemples. *Mathématiques et sciences humaines*, *78*, 5–31.

DAGAN, I., & CHURCH, K. (1994). Termight : Identifying and translating technical terminology. In *Proceedings of the 4th Conference on Applied Natural Language Processing (ANLP'94)*, (pp. 34–40).

DAGAN, I., CHURCH, K. W., & GALE, W. A. (1993). Robust bilingual word alignment for machine aided translation. In *Proceedings of the Workshop on Very Large Corpora : Academic and Industrial Perspectives*, (pp. 1–8).

DAILLE, B. (1994). *Approche mixte pour l'extraction de terminologie : statistique lexicale et filtres linguistiques*. Thèse en informatique fondamentale, Université de Paris VII.

DAILLE, B. (1999). Identification des adjectifs relationnels en corpus. In *Actes de la conférence de Traitement Automatiques du Langage Naturel (TALN '99)*.

DAMOURETTE, J., & PICHON, E. (1971). *Essai de grammaire de la langue française 7 (1911-1940)*, vol. 7 of *Des mots à la pensée*. Paris : D'Artrey.

DANLOS, L. (2005). ILIMP : Outil pour repérer les occurrences du pronom impersonnel *il*. In *TALN 2005*.

DAVID, S., & PLANTE, P. (1990). De la nécessité d'une approche morpho-syntaxique dans l'analyse de textes. *Intelligence Artificielle et Sciences Cognitives au Québec*, *3*(3), 140 – 154.

DEBILI, F., & SAMMOUDA, E. (1992). Appariement des phrases de textes bilingues. In *Proceedings of the 14th International Conference on Computational Linguistics (COLING 92)*, (pp. 517 – 538).

DELAVEAU, A. (2001). *Syntaxe : La phrase et la subordination*. Paris : Armand Colin.

DIK, S. (1989). *The theory of functional grammar*. Paris : Foris Publications.

DUBOIS, J., GIACOMO, M., GUESPIN, L., MARCELLESI, C., MARCELLESI, J.-B., & MÉVEL, J.-P. (1994). *Dictionnaire de linguistique et des sciences du langage*. Paris : Larousse, 1999 ed.

EHARA, T., FUKUSHIMA, T., WADA, Y., & SHIRAI, K. (2000). Automatic sentence partitioning of tv news sentences for closed caption service to hearing impaired people. *IPSJ SIG Notes, NL 138*(3). En japonais.

EJERHED, E. (1988). Finding clauses in unrestricted text by finitary and stochastic methods. In *ACL Proceedings, Second Conference on Applied Natural Language Processing*, (pp. 219–227).
URL `citeseer.ist.psu.edu/eva88finding.html`

ENGUEHARD, C., & PANTERA, L. (1995). Automatic natural acquisition of a terminology. *Journal of Quantitative Linguistics*, *2*(1), 27 – 32.

FLUHR, C., BISSON, F., & FAÏZA, E. (2000). Parallel text alignment using crosslingual information retrieval techniques. In J. Véronis (Ed.) *Parallel text processing*, chap. 9, (pp. 187 – 200). Dordrecht : Kluwer Academic Publishers.

FRAIKIN, C., NESTEROV, Y., & VAN DOOREN, P. (2006). A gradient-type algorithm optimizing the coupling between matrices and application to graph matching. In *Proceedings of the 13-th ILAS conference in Amsterdam*.

FUCHS, C. (1992). Les subordonnées introduites par *encore que* en français. *Subordination (Travaux Linguistiques du CERLICO), 5*.

FUCHS, C. (1996). *Les ambiguïtés du français*. Paris : Ophrys.

FUCHS, C., & VICTORRI, B. (1993a). Sémantique. In *Linguistique et traitement automatiques des langues*, chap. 5. Paris : Hachette.

FUCHS, C., & VICTORRI, B. (1993b). Syntaxe. In *Linguistique et traitement automatiques des langues*, chap. 4. Paris : Hachette.

FUKUI, M., HIGUCHI, S., FUJII, A., & ISHIKAWA, T. (2001). Bilingual lexicon extraction using japan-us patent family corpora. *IPSJ SIG Notes, NL 145*(4), 23–28. En japonais.

FUNG, P., & CHURCH, K. W. (1994). K-vec : A new approach for aligning parallel texts. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING 94)*, vol. 2, (pp. 1096 – 1102).

FUNG, P., & MCKEOWN, K. (1994). Aligning noisy parallel corpora across language groups : Word pair feature by dynamic time warping. In *Proceedings of AMTA-94*, (pp. 517 – 538).

GALE, W. A., & CHURCH, K. W. (1991). Identifying word correspondences in parallel texts. In *Proceedings of the 4th DARPA Speach and Natural Language Workshop*, (pp. 152 – 157).

GALE, W. A., & CHURCH, K. W. (1993). A program for aligning sentences in bilingual corpora. *Computational Linguistics, 19*(3), 75–102.

GARDES-TAMINE, J. (1998). *La Grammaire 2, Syntaxe*. Paris : Colin.

GARDES-TAMINE, J. (2003). Phrase, proposition, énoncé etc. : Pour une nouvelle terminologie. *L'information grammaticale, 98*, 23 – 27.

GARNIER, C. (1982). *La phrase japonaise : structures complexes en japonais moderne*. Paris : Publications orientalistes de France.

GAUSSIER, E. (1998). Flow network models for word alignment and terminology extraction from bilingual corpora. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL'98)*, (pp. 444–450).

GIGUET, E. (1998). *Méthode pour l'analyse automatique de structures formelles sur documents multilingues*. Thèse de doctorat, Université de Caen.

GINESTE, M.-D. (2003). De la phrase à la proposition sémantique : Un point de vue de la psychologie cognitive du langage. *L'information grammaticale, 98*, 48 – 51.

GOCHET, P., & GRIBOMONT, P. (1990). *Logique*, vol. 1 : Méthodes pour l'informatique fondamentale. Paris : HERMES.

GREIBACH, S. A. (1965). A new normal-form theorem for context-free phrase structure grammars. *Journal of the association for computing machinery, 12*(1), 42 – 52.

GREVISSE, M. (1969). *Cours d'analyse grammaticale*. Paris : De Boeck Duculot, sixième ed.

GREVISSE, M. (1990). *Précis de grammaire française.* Paris : Duculot, vingt neuvième ed.

GREVISSE, M. (1993). *Le bon usage : grammaire française.* Paris : Duculot, treizième ed. Édition par André GOOSSE.

GUIMIER, C. (1993). L'établissement d'un corpus de circonstants. In C. Guimier (Ed.) *1001 circonstants*, (pp. 11 – 45). Caen : Presses Universitaires de Caen.

HAGA, Y. (1954). "陳述"とは何もの？ [chinjutsu towa nani mono ?]. 国語国文 *[kokugo kokubun]*, *23*(4).

HALLIDAY, M. A. K. (1962). Linguistique générale et linguistique appliqée à l'enseignement des langues. *Études de linguistique appliquée*, *1*, 5 – 42.

HARRIS, B. (1988a). Are you bi-textual ? *Language Technology*, *7*, 41.

HARRIS, B. (1988b). Bi-text : A new concept in translation theory. *Language Monthly*, *54*, 8–10.

HARRIS, Z. S. (1976). *Notes du cours de syntaxe.* Paris : Seuil.

HARUNO, M., IKEHARA, S., & YAMAZAKI, T. (1996). Learning bilingual collocations by word-level sorting. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING'96)*, (pp. 525– 530).

HARUNO, M., & YAMAZAKI, T. (1996). Bilingual text alignment using statistical and dictionary information. *IPSJ SIG Notes, NL 112*(4), 23–30. En japonais.

HASHIMOTO, S. (1934). 新文典別記 *[shin butten bekki]*. Tokyo : Fuzambo.

HASHIMOTO, S. (1938). 改制新文典別記　口語篇 *[kaisei shin butten bekki - kôgo hen]*. Tokyo : Fuzambo.

HASHIMOTO, S. (1948). 国語法研究 *[kokugo hô kenkyû]*, vol. 2 of 橋本進吉博士著作集 *[Hashimoto Shinkichi hakushi chosakushû]*. Tokyo : Iwanami.

HASHIMOTO, S. (1969). 助詞・助動詞の研究 *[joshi-jodôshi no kenyû]*, vol. 8 of 橋本進吉博士著作集 *[hashimoto shinkichi hakase chosaku-shu]*. Tokyo : Iwanami.

HATIER (Ed.) (1990). *La grammaire pour tous.* Bescherelle 3. Paris : Hatier.

HAYAMI, A. (1923). 論理学 *[Ronri gaku]*. Iwanami, nouvelle ed.

HAYASHI, O., KINDAICHI, H., & SHIBATA, T. (Eds.) (1988). *An Encyclopaedia of the Japanese Language.* Tokyo : Taishukan Shoten.

HAYASHI, S. (1960). 基本文型の研究 *[kihon bunkei no kenkyû]*. Meiji tosho.

HUOT, H. (2001). *Morphologie : Forme et sens des mots du français*. Paris : Armand Colin.

HWANG, D., & NAGAO, M. (1994). Aligning of japanese and korean texts by analogy. *IPSJ SIG Notes, NL 99*(12), 87–94. En japonais.

HWANG, D., NAGAO, M., & SATO, S. (1993). Construction of a thesaurus for korean. *IPSJ SIG Notes, NL 94*(12), 79–84. En japonais.

ICHIKAWA, T. (1976). 副用語 [fukuyôgo]. In 文法 *I [bunpô I]*, vol. 6 of 岩波講座日本語 *[Iwanami kôza nihongo]*, chap. 6. Tokyo : Iwanami Shoten.

IKEHARA, S., SHIRAI, S., & UCHINO, H. (1996). A statical method for extracting uninterrupted and interrupted collocations from very large corpora. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING'96)*, (pp. 574– 615).

IMAMURA, K. (2000). A hierarchical phrase alignment from english and japanese bilingual text. In *Proceedings of CICLing 2001*.

ISABELLE, P. (1992). La bi-textualité : vers une nouvelle génération d'aides à la traduction et la terminologie. *Meta, XXXVII*(4), 721–737.

ISAHARA, H., & HARUNO, M. (2000). Japanese-english aligned bilingual corpora. In J. Véronis (Ed.) *Parallel text processing*, (pp. 313 – 334). Kluwer Academic Publishers.

ISHIMOTO, H., UTSURO, T., MATSUMOTO, Y., & NAGAO, M. (1993). Structural matching between parallel sentences of english and japanese. *IPSJ SIG Notes, NL 95*(11), 81–88. En japonais.

JACQUEMIN, C. (1997). *Variation terminologique : Reconnaissance et acquisition automatiques de termes et de leurs variantes en corpus*. Mémoire d'habilitation à diriger des recherches en informatique fondamentale, Université de Nantes.

JACQUEMIN, C. (1999). Syntagmatic ans paradigmatic representations of term variation. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, (pp. 341–348).

JACQUES, M.-P. (2005). Que : la valise des étiquettes. In *TALN 2005*.

KAJI, H., KIDA, Y., & MORIMOTO, Y. (1992). Learning translation templates from bilingual text. In *Proceedings of the 14th International Conference on Computational Linguistics (COLING'92)*, (pp. 672–678).

KASHIOKA, H., MARUYAMA, T., & TANAKA, H. (2003). Building a parallel corpus for monologue with clause alignment. In *Proceedings of the 9th Machine Translation Summit*, (pp. 216 – 223).

KAWAMOTO, S. (1958). フランス語のthèmeとsujet、日本語の「は」と「が」 [furansugo no thème to sujet, nihongo no "wa" to "ga"]. フランス語研究 *[furansugo kenkyû]*, (17). Repris dans 言語の構造 [gengo no kôzô] publié en 1985 par Hakusuisha.

KAY, M., & RÖSCHEISEN, M. (1988). Text-translation alignment. Tech. rep., Xerox Palo Alto Research Center.

KAY, M., & RÖSCHEISEN, M. (1993). Text-translation alignment. *Computational Linguistics*, *19*(1), 121–142.

KIKUCHI, Y. (1995). は構文の概観 [wa-kôbun no gaikan]. In 日本語の主題と取り立て *[Nihongo no shudai to toritate]*. Tokyo : Kuroshio.

KIM, Y.-B., & EHARA, T. (1993). An automatic sentence breaking method for japanese-to-english machine translation. *IPSJ SIG Notes*, *NL 93*(3). En japonais.

KIMURA, M., NOMURA, K., & HIRAKAWA, H. (1993). Sentence division at preediting for japanese to english machine translation. *IPSJ SIG Notes, NL 93*(3). En japonais.

KINDAICHI, H. (1953). 不変化助動詞の本質（その一）（その二） [fuhenka jodôshi no honshitsu (sono ichi) (sono ni)]. 国語国文 *[kokugo kokubun]*, *22*(2, 3).

KINDAICHI, H. (1988). 日本語 *[nihongo]*, vol. 2. Iwanami.

KITAHARA, Y. (1976). 文の構造 [bun no kôzô]. In 文法 *I [bunpô I]*, vol. 6 of 岩波講座日本語 *[Iwanami kôza nihongo]*, chap. 2. Tokyo : Iwanami Shoten.

KITAHARA, Y. (1988). 文の構造と文の成分 [bun no kôzô to bun no seibun]. In *An Encyclopaedia of the Japanese Language*, chap. IV-4. Tokyo : Taishukan Shoten.

KITAMURA, M., & MATSUMOTO, Y. (1997). Automatic extraction of translation patterns in parallel corpora. *IPSJ Journal*, *38*(4), 727– 736. En japonais.

KLINGLER, D. (2003). Spécificité du dispositif créé par le marqueur *wa* en japonais comparaison avec le français. *Travaux de linguistique*, *2*(47).

KNUTH, D. E. (1997). *The art of computer programming*, vol. 3 : Sorting and searching. Addison-Wesley, third ed.

KOKURITSU KOKUGO KENKYÛ JO (1963). 話し言葉の文型（*2*）独話資料による研究 *[hanashi kotoba no bunkei (2) dokuwa shirô ni yoru kenkyû]*, vol. 23 of 国語研報告 *[kokugoken hôkoku]*. Shûei shuppan.

KOSINOV, S., & CAELLI, T. (2002). Inexact multisubgraph matching using graph eigenspace and clustering models. In *Proceedings of SSPR/SPR*, vol. 2396, (pp. 133–142).

KOSINOV, S., & CAELLI, T. (2004). An eigenspace projection clustering method for inexact graph matching. *IEEE transactions on pattern analysis and machine intelligence, 26*(4), 515–519.

KRAIF, O. (1999). Identification des cognats et alignement bi-textuel : une étude empirique. In *Actes de la 6ème conférence annuelle sur le Traitement Automatique des Langues Naturelles (TALN 99)*, (pp. 205–214).

KRAIF, O. (2001). Exploitation des cognats dans les systèmes d'alignement bi-textuel : architecture et évaluation. *TAL, 42*(3).

KRAIF, O. (2002). Méthodes de filtrage pour l'extraction d'un lexique bilingue à partir d'un corpus aligné. In J. Véronis (Ed.) *Lexicometrica*, vol. Numéro spécial, (pp. 1–22). ILPGA - Paris 3.

KÜBLER, N., & FRÉROT, C. (2003). Verbs in specialised corpora : From manual corpus-based description to automatic extraction in an english-french parallel corpus. In *Proceedings of the Corpus Linguistics Conference*. Format PPT disponible sur `http://wall.jussieu.fr/~nkubler/`.

KUDO, H. (1989). 現代日本語の文の叙法性　序章 [gendai nihongo no bun no johôsei - joshô]. 東京外国語大学論集 *[tokyô gaikokugo daigaku ronshû]*, (39).

KUDO, H. (1997). 評価成分をめぐって [hyôka seibun wo megutte]. In 日本語文法　体系と方法 *[nihongo bunpô - taikei to hôhô]*, vol. 14 of ひつじ研究叢書 *[hitsuji kenkyû sôsho]*. Tokyo : Hitsuji shobo.

KUDO, H. (2000). 副詞と文の陳述的なタイプ [fukushi to bun no chinjutsu tekina taipu]. In モダリティー *[modality]*, vol. 3 of 日本語の文法 *[nihongo no bunpô]*, chap. 3. Tokyo : Iwanami shoten.

KUDO, T., & MATSUMOTO, Y. (2000). Japanese dependency analysis based on support vector machines. In *Proceedings of the EMNLP/VLC 2000*.
URL `http://chasen.org/~taku/index.html.en`

KUDO, T., & MATSUMOTO, Y. (2002). Japanese dependency analyisis using cascaded chunking. In *Proceedings of the CONLL 2002*.
URL `http://chasen.org/~taku/index.html.en`

KUHN, M. G. (1999). *UTF-8 and Unicode FAQ for Unix/Linux*. http ://www.cl.cam.ac.uk/ mgk25/unicode.html, 31 janvier 2003 ed.

KUNO, S. (1973). 日本文法研究 *[nihon bunpô kenkyû]*. Tokyo : Taishukan.

KUNO, S. (1978). 談話の文法 *[danwa no bunpô]*. Tokyo : Taishukan.

KURODA, S. Y. (1965). *Generative Grammatical Studies in the Japanese Language*. Ph. d. dissertation, Massachusetts Institute of Technology. Publié par Garland Publishing, 1979.

KURODA, S. Y. (1973). Le jugement catégorique et le jugement thétique ; exemples tirés de la syntaxe japonaise. *Langages*, *30*, 81 – 110.

KUROHASHI, S. (2000). 結構やるな、KNP. *IPSJ Magazine*, *41*(11), 1215 – 1220.

KUROHASHI, S., & NAGAO, M. (1991). A method for analyzing conjunctive structures in japanese. *IPSJ SIG Notes, NL - 86*(2), 1 – 8.

KUROHASHI, S., & NAGAO, M. (1992). A method for analyzing conjunctive structures in japanese. *IPSJ Journal*, *33*(8), 1022 – 1031.

LANGÉ, J.-M., & GAUSSIER, É. (1995). Alignement de corpus multilingues au niveau des phrases. *TAL*, *36*(1–2).

LANGLAIS, P. (1997). Alignement de corpus bilingues : intérêt, algorithmes et évaluation. In *Bulletin de Linguistique Appliquée et Générale, numéro Hors Série*, (pp. 245–254). Université de Franche-Comté.

LANGLAIS, P., & EL-BÈZE, M. (1997). Alignement de corpus bilingues : algorithmes et évaluation. In *Proceedings of 1ères Journées Scientifiques et Techniques du Réseau Francophone de l'Ingénierie de la langue de l'AUPELF-UREF (JST)*, (pp. 191–197).

LE GOFFIC, P. (1992). *Que* en français : Essai de vue d'ensemble. *Subordination (Travaux Linguistiques du CERLICO)*, *5*.

LE GOFFIC, P. (1993a). *Grammaire de la phrase française*. Paris : Hachette.

LE GOFFIC, P. (1993b). Les subordonnées circonstancielles et le classement formel des subordonnées. In C. Guimier (Ed.) *1001 circonstants*, (pp. 69 – 102). Caen : Presses Universitaires de Caen.

LE GOFFIC, P. (2002). Marqueurs d'interrogation / indéfinition / subordination : Essai de vue d'ensemble. *Verbum*, *XXIV*(4), 315 – 340.

LE NY, J.-F. (1979). *La sémantique psychologique*. Le psychologue. Paris : Presses Universitaires de France.

LE NY, J.-F. (1987). Sémantique psychologique. In J. R. et Jean-Pierre THIBAUT (Ed.) *Problèmes de psycholinguistique*. Mardaga.

LEBART, L., PIRON, M., & MORINEAU, A. (2006). *Statistique exploratoire multidimensionnelle : visualisation et inférence en fouilles de données*. Dunod, 4ème ed.

LEFFA, V. J. (1998). Clause processing in complex sentences. In *Proceedings of LREC'98*.
URL `citeseer.ist.psu.edu/eva88finding.html`

LÉON, J. (2003). Proposition, phrase, énoncé dans la grammaire : Parcours historique. *L'information grammaticale, 98*, 5 – 16.

LERALLUT, R. (2006). *Modélisation et interprétation d'images à l'aide de graphes*. Thèse de doctorat, École des Mines de Paris.

LI, C. N., & THOMPSON, S. A. (1976). Subject and topic : a new typology of language. In *Subject and topic*, (pp. 457 – 491). New York : Academic Press Inc.

MAAREK, Y. S., BERRY, D. M., & KAISER, G. E. (1991). An informationretrieval approach for automatically constructing software libraries. *IEEE Translations on Software Engineering, 17*(8), 800–813.

MAEGAARD, B., & SPANG-HANSSEN, E. (1973). Segmentation of French sentences. In *Proceedings of COLING '73*.

MARTIN, W. J. R., AL, B. P. F., & VAN STERKENBURG, P. J. G. (1983). On the processing of a text corpus : From textual data to lexicographic information. In *Lexicography : Principes and Practice*. London : Hartmann.

MARUYAMA, T., KASHIOKA, H., KUMANO, T., & TANAKA, H. (2004). Development and evaluation of japanese clause boundaries annotation program. *Shizen gengo shori, 11*(3). En japonais.

MASUOKA, T. (1991). モダリティーの文法 *[modaritî no bunpô]*. Tokyo : Kurishio Shuppan.

MASUOKA, T. (1997). 複文 *[fukubun]*, vol. 2 of 新日本語文法選書 *[shin nihongo bunpô sensho]*. Tokyo : Kurishio Shuppan.

MASUOKA, T. (2002). 複文各論 [fukubin kakuron]. In 複文と談話 *[fukubun to danwa]*, vol. 4 of 日本語の文法 *[nihongo no bunpô]*, chap. 2. Tokyo : Iwanami Shoten.

MASUOKA, T., & TAKUBO, Y. (1992). 基礎日本語文法 *[kiso nihongo bunpô]*. Tokyo : Kurishio Shuppan.

MATSUMOTO, Y., ISHIMOTO, H., & UTSURO, T. (1993). Structural matching of parallel texts. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, (pp. 23– 30).

MATSUMOTO, Y., KITAUCHI, A., YAMASHITA, T., HIRANO, Y., MATSUDA, H., TAKAOKA, K., & ASAHARA, M. (2002). *Morphological Analysis System Chasen 2.2.9*. Nara Institute of Science and Technology.

MATSUMOTO, Y., & SUGIMURA, R. (1986). 論理型言語に基づく構文解析システムsax. コンピュータソフトウェア, *3*(4), 308 – 315.

MATSUMOTO, Y., TANAKA, H., HIRAKAWA, H., MIYOSHI, H., & YASUKAWA, H. (1983). Bup : A bottom-up parser embedded in prolog. *New Generation Computing, 1*, 145 – 158.

MATSUMOTO, Y., & UTSURO, T. (2000). Lexical knowledge acquisition. In R. DALE, H. MOISL, & H. SOMERS (Eds.) *Handbook of Natural Language Processing, Part II*, (pp. 563 – 610). Marcel Dekker.

MATSUSHITA, D. (1928). 改撰標準日本文法 *[kaisen hyôjun nihon bunpô]*. Tokyo : Kigensha.

MEILLET, A. (1903). *Introduction à l'étude comparative des langues indo-européenne*. Paris : Hachette.

MELAMED, I. D. (1996). A geometric approach to mapping bitext correspondence. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

MIKAMI, A. (1953). 現代語法序説 *[gendaigohô josetsu]*. Tokyo : Toko shoin. Nouvelle édition publiée en 1972 par Kuroshio Shuppan.

MIKAMI, A. (1955). 現代語法新説 *[gendaigohô shinsetsu]*. Tokyo : Toko shoin. Nouvelle édition publiée en 1972 par Kuroshio Shuppan.

MIKAMI, A. (1959). 新訂版　現代語法序説 *[sintei ban - gendaigohô josetsu]*. Tokyo : Toko shoin. Nouvelle édition publiée en 1972 par Kuroshio Shuppan, sous le nouveau titre : 続・現代語法序説 [zoku - gendaigohô josetsu].

MIKAMI, A. (1960). 象は鼻が長い *[zô wa hana ga nagai]*. Tokyo : Kuroshio Shuppan.

MIKAMI, A. (1963a). 日本語の構文 *[nihongo no kôbun]*. Tokyo : Kuroshio Shuppan.

MIKAMI, A. (1963b). 日本語の論理 *[nihongo no ronri]*. Tokyo : Kuroshio Shuppan.

MIKAMI, A. (1970). 文法小論集 *[bunpô shôron shû]*. Tokyo : Kuroshio Shuppan.

MILLER, P., & TORRIS, T. (1990). *Formalismes syntaxiques pour le traitement automatique du langage naturel*. Paris : HERMES.

MINAMI, F. (1974). 現代日本語の構造 *[gendai nihongo no kôzô]*. Tokyo : Taishukan.

MINAMI, F. (1993). 現代日本語文法の輪郭 *[gendai nihongo bunpô no rinkaku]*. Tokyo : Taishukan.

MIO, I. (1942). 話言葉の文法－言葉遣篇 *[hanashi kotoba no bunpô - kotoba zukai hen]*. Tokyo : Teikoku kyoiku kai.

MUNTEANU, D. S., & MARCU, D. (2002). Processing comparable corpora with bilingual suffix trees. In *Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*.

MURAKI, M. (1974). *Presupposition and Thematization*. Kaitakusha.

MURAO, H. (1991). *Studies on bilingual text alignment*. Bachelor thesis, Kyoto University. En japonais.

NAGAO, M., & MORI, S. (1994). A new method of n-gram statistics for large number of n and automatic extraction of words and phrases from large text data of japanese. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING'94)*, (pp. 611– 615).

NAKAMURA-DELLOYE, Y. (2002). *Étiqueteur de verbes et constructeur de listes de mots - implémentation en C++ et en PROLOG*. Mémoire de maîtrise, Institut National des Langues et Civilisations Orientales.

NAKAMURA-DELLOYE, Y. (2003a). *Analyse syntaxique du japonais*. Mémoire de D.E.A., Institut National des Langues et Civilisations Orientales.

NAKAMURA-DELLOYE, Y. (2003b). Complément à la fiche de lecture sur l'article de Kuroda. Dissertation du séminaire Méthodes d'Analyse Linguistique de l'INALCO, `http://www.lattice.cnrs.fr`.

NAKAMURA-DELLOYE, Y. (2003c). Rapport entre les systèmes d'interrogatifs et d'intégratifs en japonais. Dissertation du séminaire Phrase Complexe de Paris III, `http://www.lattice.cnrs.fr`.

NODA, H. (1998). 文の構造と機能からみた日本語の主題 *[Bun no kôzo to kinô kara mita nihongo no shudai]*. Thèse de doctorat en linguistique, Université de Tsukuba (Japon).

NODA, H. (2002). 単文・複文とテキスト [tanbun, fukubin to tekisuto]. In 複文と談話 *[fukubun to danwa]*, vol. 4 of 日本語の文法 *[nihongo no bunpô]*, chap. 1. Tokyo : Iwanami Shoten.

NUMATA, Y. (1986). とりたて詞 [toritateshi]. In いわゆる日本語助詞の研究 *[Iwayuru nihongo joshi no kenkyû]*, chap. 2. Tokyo : Bonjinsha.

NUMATA, Y. (2000). とりたて [toritate]. In 時・否定と取り立て *[toki, hitei to toritate]*, chap. 3. Tokyo : Iwanami.

NUMATA, Y., & JO, K. (1995). とりたて詞「も」のフォーカスとスコープ [toritateshi "mo" no fôkasu to sukôpu]. In 日本語の主題と取り立て *[Nihongo no shudai to toritate]*. Tokyo : Kuroshio.

NUMAZAKI, H., TAMURA, N., & TANAKA, H. (1989). An implementation of generalized lr-parsing algorithm based on parallel logic programming language. *IPSJ SIG Notes, NL - 74*(5), 33 – 40.

OKUTSU, K. (1974). 生成日本文法論 *[seisei nihon bunpô ron]*. Tokyo : Taishukan.

OKUTSU, K. (1978). 「ボクハウナギダ」の文法－ダとノー *["boku wa unagida" no bunp^- da to no –]*. Tokyo : Kuroshio.

OKUTSU, K. (1986). 形式副詞 [keishiki fukushi]. In いわゆる日本語助詞の研究 *[Iwayuru nihongo joshi no kenkyû]*, chap. 1. Tokyo : Bonjinsha.

OKUTSU, K., NUMATA, Y., & SUGIMOTO, T. (1986). いわゆる日本語助詞の研究 *[Iwayuru nihongo joshi no kenkyû]*. Tokyo : Bonjinsha.

OMORI, K., TSUTSUMI, J., & NAKANISHI, M. (1996). 統計情報を用いた対訳辞書の作成 [constitution du dictionnaire de lexiques bilingues utilisant des informations statistiques]. In *Proceedings of the 2nd Annual Meeting of the ISPJ*, (pp. 49– 52).

PAPAGEORGIOU, H. (1997). Clause recognition in the framework of alignment. In R. Mitkov, & N. Nicolov (Eds.) *Recent advances in natural language processing*, (pp. 417 – 425). John Benjamins.

PEREIRA, F. C. N., & SHIEBER, S. M. (1987). *Prolog and Natural-Language Analysis*. Stanford : CSLI.

PIPERIDIS, S., PAPAGEORGIOU, H., & BOUTSIS, S. (2000). From sentences to words and clauses. In J. Véronis (Ed.) *Parallel text processing*, (pp. 117 – 138). Kluwer Academic Publishers.

PRÉVOST, S. (2003). Les compléments spatiaux : Du topique au focus en passant par les cadres. *Travaux de Linguistique, 47*, 51–78.

RAPP, R. (1995). Identifying word translations in non-parallel texts. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL '95)*, (pp. 320– 322).

RAYON, N. (2003). *Segmentation et analyse morphologique automatiques du japonais en univers ouvert*. Thèse de doctorat en traitement automatique des langues, Institut National des Langues et Civilisations Orientales.

RIEGEL, M., PELLAT, J.-C., & RIOUL, R. (1994). *Grammaire méthodique du français*. Paris : Presses Universitaires de France.

SAEKI, T. (1998). 要説　日本文の語順 *[yôsetsu nihon-bun no gohun]*. Tokyo : Kuroshio Shuppan.

SAITO, H., NOYORI, M., MORI, K., & SHIMOMURA, N. (1978). 特許請求範囲文の段落分割 [segmentation des paragraphes de la section revendications du brevet]. *IPSJ SIG Notes, NL 1977*(046). En japonais.

SAKAKURA, A. (1979). 日本語の構造上の特色 *[nihongo no kôzôjô no tokushoku]*, vol. 10 of「ことば」シリーズ *[kotoba shirîzu]*. Tokyo : Bunkachô.

SAKUMA, K. (1940a). 現代日本語の表現と語法 *[gendai nihongo no hyôgen to gohô]*. Tokyo : Koseikaku. Édition complétée publiée en 1966 par Koseisha Koseikaku.

SAKUMA, K. (1940b). 現代日本語法の研究 *[gendai nihongohô no kenkyû]*. Tokyo : Koseikaku. Nouvelle édition publiée en 1952 par Koseisha Koseikaku.

SEMMAR, N., & FLUHR, C. (2007). Utilisation d'une approche basée sur la recherche cross-lingue d'information pour l'alignement de phrases à partir de textes bilingues arabe-français. In *TALN 2007*, (pp. 411 – 420).

SHIBATANI, M. (1978). 日本語の分析 *[nihongo no bunseki]*. Tokyo : Taishukan.

SHIBATANI, M. (1985). 主語プロトタイプ論 [shugo puroto taipu ron]. 日本語学 *[nihongo gaku]*, *4*(10).

SIMARD, M. (1998). The BAF : A corpus of english-french bitext. In *Proceedings of First International Conference on Language Resources and Evaluation (LREC)*, (pp. 489–496).

SIMARD, M. (2003). *Mémoire de traduction sous-phrastique*. Thèse de doctorat en informatique, Université de Montréal.

SIMARD, M., CANCEDDA, N., CAVESTRO, B., DYMETMAN, M., GAUSSIER, E., GOUTTE, C., LANGLAIS, P., MAUSER, A., & YAMADA, K. (2005). Traduction automatique statistique avec des segments discontinus. In *TALN 2005*, (pp. 233–242).

SIMARD, M., FOSTER, G., & ISABELLE, P. (1992). Using cognates to align sentences in bilingual corpora. In *Proceedings of the Fourth International Conference on Thoretical and Methodological Issues in Machine Translation (TMI)*, (pp. 67 –81).

SIMARD, M., & PLAMONDON, P. (1998). Bilingual sentence alignment : Balancing robustness and accuracy. *Machine Translation*, *13*(1), 59–80.

SMADJA, F. (1993). Retieving collocations from text : Xtract. *Computational Linguistics*, *19*(1), 143 – 177.

SMADJA, F., MCKEOWN, K., & HATZIVASSILOGLOU, V. (1996). Translating collocations for bilingual lexicons : A statistical approach. *Computational Linguistics*, *22*(1), 1 – 38.

SUZUKI, S. (1972). 日本語文法・形態論 *[nihongo bunpô - keitai ron]*. Tokyo : Mugi shobo.

SUZUKI, S. (1978). 文法Ｉ *[bunpô I]*, vol. 3 of 教師用日本語教育ハンドブック *[kyôshiyô nihongo kyôiku hando bukku]*. Tokyo : Bonjinsha.

TAKAGAKI, Y. (2001). Les stratégies discursives du français et du japonais et le problème de l'organisation textuelle. フランス語フランス文学研究 *[furansu-go furansu-bungaku kenkyû], 79*. En japonais.

TAKAMI, K. (1995). 機能的構文論による日英語比較 *[kinôteki kôbunron niyoru nichi-eigo hikaku]*. Tokyo : Kuroshio.

TAKEISHI, E., & HAYASHI, Y. (1992). 接続構造解析に基づく日本語複文の分割 [segmentation des phrases complexes japonaises basée sur l'analyse des structures de connexion]. *IPSJ Journal, 33*(5). En japonais.

TAMBA, I., & TERADA, A. (1991). La phrase japonaise et son double dispositif d'intégration des noms : Les particules dites relationnelles et casuelles. *Langages*, (104).

TANAKA, H. (1989). 自然言語解析の基礎 *[Base de l'analyse linguistique du japonais]*. Tôkyô : Sangyô Tosho.

TANAKA, K., & IWASAKI, H. (1996). Extraction of lexical translations from non-aligned corpora. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING'96)*, (pp. 580– 585).

TERAMURA, H. (1978). 連体修飾のシンタクスとその意味－その４－ [rentai shushoku no shintakusu to sono imi - sono 4]. 日本語・日本文化 *[nihongo - nihon bunka]*, (7). Reprise dans 寺村秀夫論文集Ｉ日本語文法編, Kuroshio Shuppan, 1993.

TERAMURA, H. (1982a). 日本語における単文、複文認定の問題 [nihongo ni okeru tanbun, fukubun nintei no mondai]. In 講座日本語学 *[kôza nihongo-gaku]*, vol. 11. Tokyo : Meiji shoin. Reprise dans 寺村秀夫論文集Ｉ日本語文法編, Kuroshio Shuppan, 1993.

TERAMURA, H. (1982b). 日本語のシンタクスと意味 *[Nihongo no shintakusu to imi]*, vol. 1. Tokyo : Kuroshio Shuppan.

TERAMURA, H. (1984). 日本語のシンタクスと意味 *[Nihongo no shintakusu to imi]*, vol. 2. Tokyo : Kuroshio Shuppan.

TERAMURA, H. (1991). 日本語のシンタクスと意味 *[Nihongo no shintakusu to imi]*, vol. 3. Tokyo : Kuroshio Shuppan.

TERAMURA, H. (1993). 寺村秀夫論文集Ｉ日本語文法編 *[teramura hideo ronbun shû - nihongo bunpô hen]*, vol. 1. Tokyo : Kuroshio Shuppan.

TESNIÈRE, L. (1988). *Éléments de Syntaxe Structurale*. Paris : KLINCKSIECK, deuxième ed. Cinquième tirage.

TIEDEMANN, J. (1991). *Automatical lexicon extraction from aligned bilingual corpora*. Diploma thesis in computer science, Otto-von-Guericke-Universität Magdeburg.

TJONG, E. F., SANG, K., & DÉJEAN, H. (2001). Introduction to the conll-2001 shared task : Clause identification. In *Proceedings of CoNLL-2001*.

TOKIEDA, M. (1950). 日本文法　口語篇 *[nihon bunpô kôgo hen]*. Tokyo : Iwanami.

TSUJI, K. (2002). Automatic extraction of translational japanese-katakana and english word pairs from bilingual-corpora. In *Proceedings of the Third International Conference on Language Ressources and Evaluation (LREC 2002)*.

TSUJI, K., DAILLE, B., & KAGEURA, K. (2002). Extracting french-japanese word pairs from bilingual corpora based on transliteration rules. In *Proceedings of the Third International Conference on Language Ressources and Evaluation (LREC 2002)*.

TSUJI, K., YOSHIKANE, F., & KAGEURA, K. (2000). Low-frequency words in bilingual corpora : A step towards automatic extraction of bilingual word pairs. *IPSJ SIG Notes, NL 138*(7), 47–54. En japonais.

UCHIYAMA, M., & ISAHARA, H. (2003). Reliable measures for alignment japanese-english news articles and sentences. In *Proceedings of the 41st Annual meeting of the ACL*.

UTSURO, T., IKEDA, H., YAMANE, M., MATSUMOTO, Y., & NAGAO, M. (1994). Bilingual text, matching using bilingual dictionary and statistics. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING 94)*, (pp. 1076 – 1082).

VAN DER EIJIK, P. (1993). Automationg the acquisition of bilingual terminologiy. In *Proceedings of the 6th Conference of the European Chapter of the Association for Computational Linguistics (EACL'93)*, (pp. 113 – 119).

VAN RIJSBERGEN, C. J. (1979). *Information retrieval*. Addison-Wesley, butterworths ed.

VERGNE, J. (1989). *Analyse morpho-syntaxique automatique sans dictionnaire*. Thèse de doctorat, Université Paris 6.

VERGNE, J. (1990). A parser without a dictionary as a tool for research into French syntax. In *Proceedings of COLING '90*.

VERGNE, J., GIGUET, E., LUCAS, N., & COUSIN, G. (1999). Analyseur syntaxique du Greyc. `http://users.info.unicaen.fr/~jvergne`.

VÉRONIS, J. (2000a). Alignement de corpus multilingues. In J.-M. Pierrel (Ed.) *Ingénierie des langues*, chap. 6, (pp. 313 – 334). Paris : HERMES.

VÉRONIS, J. (2000b). From the rosetta stone to the information society : A survey of parallel text processing. In J. Véronis (Ed.) *Parallel text processing*, chap. 1, (pp. 1 – 24). Dordrecht : Kluwer Academic Publishers.

VÉRONIS, J. (2000c). *Parallel text processing : Alignment and use of translation corpora*. Dordrecht : Kluwer Academic Publishers.

WAGNER, R. L., & PINCHON, J. (1991). *Grammaire du français*. Paris : Hachette supérieur.

WAKAO, T., EHARA, T., & SHIRAI, K. (1998). Partitioning long sentences : how useful it is for sentence extraction. *IPSJ SIG Notes, NL 126*(9). En japonais.

WANG, X., & REN, F. (2005). Chinese-japanese clause alignment. *Lecture Notes in Computer Science, 3406*, 400–412.

WATANABE, H., KUROHASHI, S., & ARAMAKI, E. (2000). Finding structural correspondences from bilingual parsel corpus for corpus-based traslation. In *Proceedings of COLING 2000*, (pp. 906–912).

WATANABE, M. (1971). 国語構文論 *[Kokugo kôbun ron]*. Tokyo : Haniwa Shobo.

WILMET, M. (1997). *Grammaire critique du français*. Paris : Hachette supérieur, Duculot.

WU, D. (1994). Aligning a parallel english-chinese corpus statistically with lexical criteria. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, (pp. 80 –87).

YAMADA, Y. (1908). 日本文法論 *[Nihon bunpô ron]*. Tokyo : Hobunkan.

YAMADA, Y. (1936). 日本文法学概論 *[Nihon bunpô gaku gairon]*. Tokyo : Hobunkan.

YAMAMOTO, K., & MATSUMOTO, Y. (2001). Translation pattern acquisition using dependency structures. *IPSJ Journal, 42*(9), 2239– 2247. En japonais.

YAZAWA, M. (1995). 辞書の記述と利用－言語系－「一字漢字」と「連語」をめぐって [jisho no kijutsu to riyô - gengo kei - "ichiji kanji" to "rengo" wo megutte]. 日本語学 *[nihon go gaku], 14*(4).

YAZAWA, M. (2000). 副詞的修飾の諸相 [fukushi teki shûshoku no shosô]. In 文の骨格 *[bun no kokkaku]*, vol. 1 of 日本語の文法 *[nihongo no bunpô]*, chap. 4. Tokyo : Iwanami shoten.

YOSHIKAWA, T. (1989). 日本語文法入門 *[nihongo bunpô nyûmon]*. Tokyo : ALC.

ZHANG, B.-T., & KIM, Y.-T. (1990). Morphological analysis and synthesis by automated discovery and acquisition of linguistic rules. In *Proceedings of the 13th International Conference of the Association for Computational Linguistics*, vol. 2, (pp. 431–436).