Outils et ressources linguistiques pour l'alignement de textes multilingues français-vietnamiens

Thi Minh Huyen Nguyen

Département de formation doctorale en informatique **École doctorale IAEM**
UFR STMIA

Thèse

présentée pour l'obtention du titre de

Docteur de l'Université Henri Poincaré, Nancy 1

en Informatique

par

NGUYỄN Thị Minh Huyền

Outils et ressources linguistiques pour l'alignement de textes multilingues français-vietnamiens

Soutenue en publique le 10 octobre 2006

Membres du jury :

Président du jury :	Jean-Marie PIERREL	Professeur, Université Henri Poincaré - Nancy I
Rapporteurs externes :	LƯƠNG Chi Mai	Directeur de Recherche, Académie des Sciences et Technologies du Vietnam, Hanoi, Vietnam
	Jean CAELEN	Directeur de Recherche CNRS, CLIPS, Grenoble
Référent interne :	Hazel EVERETT	Professeur, Université Nancy II
Directeur de thèse :	Laurent ROMARY	Directeur de Recherche INRIA, LORIA, Nancy

Laboratoire Lorrain de Recherche en Informatique et ses Applications - UMR 7503

Remerciements

Je tiens à remercier :

Patrice Bonhomme, pour avoir été à l'initiative de mon projet de thèse ;

Laurent Romary, pour son encadrement, sa direction, son support et sa confiance pendant mes années de thèse ;

Mme Lương Chi Mai, pour son rôle d'intermédiaire de toutes les collaborations avec les linguistes vietnamiens dans le cadre de ma thèse, et aussi pour avoir accepté d'être rapporteur de ma thèse ;

Mme Hazel Everett et M. Jean Caelen, qui ont accepté d'être mes rapporteurs. M. Thierry Declerck et M. Jean-Marie Pierrel pour leur participation au jury de thèse ;

les membres de l'équipe Langue et Dialogue, qui ont toujours été prêts à m'aider, de mon stage de maîtrise à aujourd'hui. Remerciements en particulier à Hélène, Suzanne, Ashwani, Erica, Bôn, Phương, Jean-Luc, Eric, Yannick, Sébastien, Azim, Mathieu et Bertrand pour leur amitié et leur support durant ma thèse, ainsi qu'à Isabelle, la meilleure assistante d'équipe que je connaisse ;

l'Insitut National de Technologie d'Information du Vietnam, pour leur support de mon projet de thèse ;

M. Vũ Xuân Lương et ses collègues du Centre Vietnamien de Lexicographie : Hoàng Thị Tuyền Linh, Đặng Thanh Hoà, Đào Minh Thu et Phạm Thị Thuỷ, pour leur collaboration et leur encouragement tout au long de ma thèse ;

le projet national vietnamien KC01-03 « Recherche et Développement en Reconnaissance et Traitement de la Langue Vietnamienne » pour le financement du travail linguistique dans le cadre de cette thèse ;

le comité technique de l'ISO TC 37/SC 4, pour les expériences acquises durant ces années ;

les professeurs du département de Linguistique de la faculté des Sciences Humaines et Sociales de l'Université Nationale de Hanoi, et les linguistes de l'Institut National de Linguistique du Vietnam, pour m'avoir donné des conseils précieux durant ma thèse ;

M. Benjamin Dumontet (Maison de Droits Vietnamo-Français), M. Alain Fontanel (ADETEF-Vietnam : Association pour le Développement des Échanges en Technologies Économiques et Financières) pour avoir offert des textes bilingues (français et vietnamien) dans les domaines du droit et de l'économie ;

le Département de Mathématiques, de Mécanique et d'Informatique de la faculté des Sciences, Université Nationale de Hanoi, pour m'avoir permis de suspendre mon travail durant les périodes passées en France, et pour m'avoir encouragée et soutenue pour aboutir à la fin de cette thèse ;

mes parents et ma grande famille, et les amis proches pour leur encouragement durant ces longues années. Remerciements en particulier à Minh pour sa grande amitié.

Enfin, merci à Mathias de sa compagnie merveilleuse, et d'être un lecteur et correcteur attentif de mon manuscrit.

TABLE DES MATIÈRES

Liste des tableaux

Liste des figures

Mots clés

alignement multilingue
analyse syntaxique
annotation linguistique
corpus annotés
étiquetage lexical / morphosyntaxique
grammaire d'arbres adjoints
lexique
normalisation
partie du discours
ressources linguistiques
segmentation
traitement automatique des langues
vietnamien

Sigles et Abréviations

AP	Adjectival Phrase
ARCADE	Action de Recherche Concertée sur l'Alignement de Documents et son Évaluation
CES	Corpus Encoding Standard
CKIP	Chinese Knowedge Information Processing
CLIF	Corpus et Lexiques Informatisés du Français
DCR	Data Category Registry
DCS	Data Category Selection/Specification
DEC	Dictionnaire explicatif et combinatoire
DI	Dominance Immédiate
DTD	Document Type Definition
EAGLES	Expert Advisory Group on Language Engineering Standards
ELR	Electronic Lexical Resources
ELRA	European Language Resources Association
FSR	Feature Structure Representation
GENELEX	*GENeric* LEXicon
GPSG	Generalized Phrase Structure Grammar
GRACE	Grammaires et Ressources pour les Analyseurs de Corpus et leur Évaluation
HPSG	Head-driven Phrase Structure Grammar
ISLE	International Standards for Language Engineering
ISO	International Standardization Organization
LAF	Linguistic Annotation Framework
LFG	Lexical Functional Grammar
LMF	Lexical Markup Framework
LORIA	Laboratoire LOrrain de la Recherche en Informatique et ses Applications
RL	Ressources Linguistiques
LTAG	Lexicalized Tree Adjoining Grammar
MARTIF	MAchine-Readable Terminology Interchange Format
MILE	Multilingual Isle Lexical Entry

MULTEXT	Multilingual Text Tools and Corpora
PAROLE	Preparatory Action for Linguistic Resources Organisation for Language Engineering
SC	Sub-Committee
SAV	Structure Attributs – Valeurs
SGML	Standard Generalized Markup Language
SIMPLE	Semantic Information for Multilingual Plurifunctional Lexica
TAG	Tree Adjoining Grammar
TAL	Traitement Automatique des Langues
TALN	Traitement Automatique des Langues Naturelles
TC	Technical Committee
TEI	Text Encoding Initiative
TMF	Terminological Markup Framework
TSNLP	Test Suites for Natural Language Processing
WG	Work Group
W3C	World Wide Web Consortium
XML	Extended Markup Language
=	Dans le contexte de traduction, « = » dénote une traduction équivalente
$=_{litt}$	Dans le contexte de traduction, « $=_{litt}$ » dénote une traduction mot à mot

INTRODUCTION

Durant ces dernières décennies, le traitement automatique des langues (TAL) a sans nul doute fait des progrès considérables sur le plan de la diversité des outils disponibles et celui de la qualité des résultats qu'ils fournissent. Néanmoins, ces progrès sont jusqu'à une date récente restés limités à un nombre relativement restreint de langues, majoritairement occidentales, sur lesquelles se sont focalisées la plupart des recherches entreprises dans ce domaine. Le développement d'Internet et la globalisation de la « société de l'information » ont toutefois occasionné un début d'évolution de cette situation, d'une part, en mettant en avant la problématique du multilinguisme (cherchant par exemple des informations touristiques sur la Russie, un utilisateur d'Internet devrait idéalement pouvoir profiter des ressources en russe présentes sur la toile, même s'il ne parle pas cette langue), d'autre part, en favorisant la diffusion d'un grand nombre d'outils et de ressources langagières mono et multilingues. Ces dernières années ont ainsi vu l'aboutissement d'un nombre important de travaux de recherche et de développement réalisés sur des langues qui étaient encore récemment considérées injustement comme « exotiques » – tels le japonais, le chinois et plus récemment encore l'arabe. Toutefois, il persiste encore des lacunes importantes pour une quantité non négligeable de langues peu ou pas du tout étudiées dans la communauté du TAL. C'est en particulier le cas du vietnamien, qui fait l'objet du travail de recherche présenté dans ce document.

Le vietnamien, quoique pratiqué dans le monde par environ 80 millions de locuteurs, ce qui le place au 14ème rang mondial, fait partie des langues encore peu représentées au sein de la communauté scientifique internationale, tant du point de vue des sciences humaines que de celui du TAL. Au Vietnam, les travaux en TAL sont encore rares, et sont souvent entrepris sans la participation des linguistes, qui restent assez « traditionnels » – le terrain de la linguistique informatique est presque vierge. On peut citer deux groupes institutionnels menant des recherches en traduction automatique anglais-vietnamien (à l'Université des Sciences de Hồ Chí Minh ville, et au Centre d'Applications des Technologies à Hà Nội, qui commercialise le traducteur EVTRAN), dont les résultats restent encore modestes, et dont les ressources ne sont pas dans le domaine public. Quelques chercheurs (par exemple à l'Institut Polytechnique de Hà Nội) étudient l'analyse syntaxique, mais en se focalisant principalement sur l'aspect algorithmique. À l'université de Đà Nẵng, un groupe de chercheurs mène les études sur les lexiques multilingues en collaboration avec l'équipe GETA du CLIPS-IMAG à Grenoble. D'autres recherches plus fructueuses portent sur la reconnaissance d'écriture et de parole, qui prennent en compte encore peu de ressources en TAL.

Les récents travaux portant sur la gestion de ressources multilingues et l'alignement automatique de textes parallèles sont une opportunité pour la valorisation de la langue vietnamienne dans la société de l'information. C'est dans ce contexte multilingue que nous avons élaboré notre projet de thèse, qui consiste à construire les outils et ressources linguistiques indispensables pour l'analyse automatique des textes vietnamiens, potentiellement en relation avec d'autres langues.

Le projet initial à l'origine de cette thèse a porté sur l'alignement multilingue, c'est-à-dire la mise en évidence des éléments équivalents par traduction (paragraphes, phrases, expressions, mots, *etc*.) dans des textes traduisant une même source dans plusieurs langues distinctes. Ces textes alignés constituent une ressource importante pour la traduction (semi-) automatique basée sur les exemples ainsi que les applications comme la recherche d'information multilingue, l'étude de terminologie et de traduction, *etc*. Nous nous intéressons naturellement à l'analyse de corpus bilingues français-vietnamiens, afin de construire des ressources linguistiques françaises-vietnamiennes qui pourraient être distribuées et employées dans la communauté de la recherche en TAL. Pour cela, il est indispensable d'entreprendre le développement de ressources et d'outils pour l'analyse des textes vietnamiens, encore inexistants, en lien avec ce qui a été développé pour le français.

En conséquence, notre recherche porte sur les outils et ressources linguistiques pour l'annotation de corpus vietnamiens dans une perspective monolingue ainsi que multilingue. Nous nous limitons à trois sujets principaux concernant les corpus de texte : ***l'annotation morphosyntaxique, l'analyse syntaxique et l'alignement multilingue***. Nous n'avons pas la prétention de mener à bien toutes ces tâches, mais d'étudier leur faisabilité par de premières expériences, ainsi que de construire des cadres de travail pour chacune d'elles, ce qui nous permet de visualiser les pistes à suivre ultérieurement.

L'annotation (ou l'étiquetage) morphosyntaxique consiste à identifier la catégorie syntaxique ainsi que des informations morphosyntaxiques associées aux occurrences des mots d'un texte dans leur contexte d'énonciation (Paroubek et Rajman [PAR 00]). Les corpus annotés peuvent être utilisés dans des travaux d'analyse comme la lemmatisation, l'analyse syntaxique, l'extraction de terminologie, l'acquisition d'informations lexicales ou de modèles statistiques de la langue, pouvant à leur tour être employés par des application de « plus haut niveau » comme l'extraction d'information, les systèmes de dialogue homme-machine, *etc*. Comme nous pouvons le constater dans la définition, la tâche d'annotation morphosyntaxique fait l'appel à la détermination des mots, des catégories syntaxiques ainsi que des descriptions morphosyntaxiques. Les méthodes d'étiquetage nécessitent le plus souvent un *lexique* mettant en correspondance les mots avec leurs étiquettes possibles. Notre travail porte donc sur la définition de catégories grammaticales adaptées pour le vietnamien, l'enrichissement du lexique par des descriptions syntaxiques, et le développement d'outils réalisant la segmentation d'un texte en unités lexicales (mots) et l'étiquetage catégoriel de ceux-ci.

L'analyse syntaxique consiste à identifier des syntagmes et leurs fonctions dans un texte. La détermination des composants syntaxiques est essentielle pour de nombreuses applications. « On peut distinguer les applications demandant une analyse syntaxique complète (traduction) et celles qui peuvent se contenter d'une analyse partielle (indexation, extraction d'informations, *etc*.) ; on peut également distinguer celles qui demandent un analyseur robuste, tolérant aux fautes, comme la traduction ou l'indexation de gros volumes de textes, et celles qui reposent sur un analyseur « exigeant », c'est-à-dire capable de détecter toutes les agrammaticalités (correction d'orthographe, logiciels d'apprentissage d'une langue) » – Abeillé [ABE 00]. Les grammaires à large couverture et les corpus arborés (annotés syntaxiquement, *TreeBank* en anglais) sont les deux types de ressources les plus importants dans ce domaine. Nous ne sommes pas en mesure de construire complètement ces ressources, mais présentons une première tentative de modélisation de la grammaire vietnamienne et définissons un cadre pour la construction d'un lexique syntaxique et d'un corpus arboré pour le vietnamien.

L'alignement multilingue consiste à mettre au jour les liens entre les parties équivalentes d'un corpus parallèle (c'est-à-dire des textes accompagnés de leur traduction). Les applications des corpus parallèles alignés sont extrêmement diverses : constitution de mémoires de traduction, extraction de dictionnaires et de listes terminologiques bilingues, extraction de connaissances pour la recherche d'information multilingue, construction d'exemples pour l'enseignement assisté par ordinateur ou la linguistique contrastive, *etc*. – Véronis [VER 00b]. Nous avons pour but de construire un système d'alignement multilingue à deux niveaux de granularité : phrase et mot. L'application de ce système à des corpus multilingues français-vietnamiens et anglais-vietnamiens doit être évaluée pour chercher les pistes à suivre par la suite.

Ces trois tâches ont d'ores et déjà fait l'objet de recherches importantes pour les langues occidentales, ainsi que pour quelques langues « exotiques » déjà mentionnées, comme le japonais, le chinois, *etc*. Le TAL ayant encore au Vietnam un statut de domaine « nouveau né », nos efforts ne porteront pas en priorité sur le développement de nouvelles techniques, mais sur l'adaptation au vietnamien des méthodes et formalismes déjà mis au point pour d'autres langues. Cette adaptation est en particulier envisageable grâce à la structure de beaucoup de ces outils, qui se présentent sous la forme d'algorithmes génériques s'appuyant sur un ensemble de ressources caractéristiques de la langue étudiée afin de pouvoir analyser des textes écrits en cette langue. Notre travail porte donc en majorité sur la construction de ressources linguistiques pouvant permettre à ces outils de travailler sur le vietnamien : lexiques (nécessaires pour l'étiquetage catégoriel et l'analyse syntaxique), grammaire (pour l'analyse syntaxique) et divers types de corpus de référence pour l'acquisition de connaissances empiriques (en particulier statistiques) sur la langue.

Pour que ces ressources soient facilement extensibles et exploitables par la communauté de recherche en TAL, nous considérons tout au long de la thèse un autre sujet important, qui porte sur ***la normalisation des ressources*** (contenu, codage, *etc*., *cf*. Bonhomme [BON 00a]). En effet, si les ressources langagières jouent comme nous venons de l'évoquer un rôle essentiel dans le domaine de l'ingénierie des langues, leur coût de construction, financier autant qu'humain, est très important. La possibilité de les échanger facilement entre équipes de recherche ainsi qu'entre systèmes de traitement de la langue permet de réduire l'impact de ce coût, et est donc un facteur essentiel pour une progression rapide dans le domaine du TAL. Consciente de l'enjeu que cela représente, en particulier pour stimuler le développement d'une discipline naissante au Vietnam, nous mettons l'accent sur la normalisation de la représentation des ressources recueillies. Cette exigence trouve le cadre de sa mise en œuvre dans les activités internationales en cours pour la normalisation des ressources langagières d'un sous-comité du comité technique 37 de l'organisation internationale pour la normalisation (ISO/TC 37/SC 4). Les ressources qui retiennent notre attention par rapport aux objectifs que nous nous sommes fixés sont les lexiques morphosyntaxique et syntaxique, les corpus étiquetés morpho-syntaxiquement, les corpus arborés et les corpus parallèles alignés.

Ce document présente nos travaux introduits ci-dessus. Il est organisé selon les cinq chapitres suivants :

Chapitre 1 – ***Ressources linguistiques pour le TAL***. Dans ce chapitre, nous réalisons un état des lieux du domaine des ressources linguistiques au niveau international. Cette présentation est restreinte aux ressources pour le TAL qui nous intéressent dans le cadre de cette thèse, à savoir les lexiques, grammaires et divers types de corpus (bruts, annotés et parallèles). Nous abordons également la question de la normalisation de la gestion de ces ressources linguistiques, tâche entreprise en particulier par le sous-comité TC 37/SC 4 de l'ISO, aux travaux duquel nous avons pu participer. Ce bilan du contexte de recherche est suivi d'une présentation des tâches qu'il nous revient de traiter durant notre thèse.

Chapitre 2 – ***Notions élémentaires de vietnamien***. Ce chapitre a pour but de fournir au lecteur une connaissance des principes de base de la langue vietnamienne suffisante pour comprendre les difficultés particulières liées à l'exploitation informatique de cette langue, et ainsi les facteurs ayant guidé les choix que nous avons effectués au cours de nos travaux. Nous présentons les principales caractéristiques du vietnamien du point de vue de l'écriture, de la phonétique, du vocabulaire ainsi que d'autres attributs grammaticaux importants d'une langue isolante.

Chapitre 3 – ***Construction d'outils et ressources linguistiques pour l'analyse morphosyntaxique du vietnamien***. Nous présentons dans cette partie les travaux sur l'annotation morphosyntaxique des corpus vietnamiens. Il s'agit de la construction des ressources lexicales (lexique, corpus annotés) du vietnamien et des outils d'étiquetage morphosyntaxique. Nous insistons particulièrement sur le fait qu'il n'y a pas, jusqu'à présent, de consensus sur la question des parties du discours du vietnamien dans la communauté linguistique. Une partie importante de notre travail est donc de construire un lexique avec des descriptions lexicales qui nous permettent de définir ultérieurement des jeux d'étiquettes comparables pour la tâche d'étiquetage. Nous discutons ensuite des problèmes de segmentation des textes vietnamiens en unités lexicales et des solutions possibles. Enfin, nous présentons une méthode statistique simple fondée sur l'utilisation d'un modèle de Markov caché pour l'étiquetage automatique de corpus vietnamiens. Par ailleurs, toutes les ressources construites font l'objet d'une discussion sur leur représentation normalisée.

Chapitre 4 – ***Ressources linguistiques pour l'analyse syntaxique du vietnamien***. Ce chapitre discute de la modélisation de la grammaire vietnamienne à l'aide du formalisme TAG (Grammaire d'Arbres Adjoints), que nous avons expérimentée grâce au parseur LLP2 développé au LORIA. Dans le cadre de cette thèse, il n'est bien sûr pas question d'aboutir à une analyse syntaxique à large couverture, mais nous montrons que l'approche TAG permet de rendre compte de suffisamment de phénomènes observables sur le vietnamien. Nous finissons ce chapitre par une spécification de ce que pourrait être une TreeBank à la vietnamienne.

Chapitre 5 – ***Traitement de corpus multilingues français-vietnamiens***. Nous présentons dans ce chapitre les problèmes concernant l'alignement de textes multilingues. Nous structurons cette étude en deux parties : l'alignement au niveau des phrases et celui au niveau des mots (unités lexicales), pour lesquels nous avons développé deux outils spécialisés. Pour l'alignement au niveau des phrases, nous disposons d'un outil fondant son analyse sur la structure hiérarchique des documents, qui s'est montré d'une assez grande efficacité pour le couple de langues français-anglais dans le cadre de la campagne d'évaluation ARCADE (*cf.* Véronis et Langlais [VER 00a]). Notre première tâche est donc d'évaluer l'adaptation de cet outil aux textes français-vietnamiens. Nous proposons ensuite un outil d'alignement au niveau des unités lexicales. Dans le temps limité de la thèse, nous ne pouvons effectuer qu'une évaluation rapide de l'application de cet outil à chaque couple de langues d'un texte multilingue français – vietnamien – anglais, dont chaque texte est soumis à un prétraitement lexical, afin de donner un premier aperçu de l'efficacité de la technique mise au point. Nous avons également participé avec cet outil à la campagne d'évaluation ARCADE II (Chiao *et al* [CHI 06]), dont nous rappelons les resultats.

Les chapitres 3, 4 et 5 de ce document ont pour objectif principal de fournir des bases de travail pour une recherche à plus long terme permettant le développement d'outils de référence pour l'analyse automatique du vietnamien. En conséquence, chacun de ces chapitres se conclut sur un bilan du travail réalisé détaillant et « planifiant » les tâches restant à accomplir.

Chapitre 1

Ressources linguistiques pour le TAL

Dans ce chapitre, nous réalisons un état des lieux du domaine des ressources linguistiques au niveau international. Cette présentation est restreinte aux ressources pour le TAL qui nous intéressent dans le cadre de cette thèse, à savoir les lexiques, grammaires et divers types de corpus (bruts, annotés et parallèles). Nous abordons également la question de la normalisation de la gestion de ces ressources linguistiques, tâche entreprise en particulier par le sous-comité TC 37/SC 4 de l'ISO, aux travaux duquel nous avons pu participer. Ce bilan du contexte de recherche est suivi d'une présentation des tâches qu'il nous revient de traiter durant notre thèse.

- *Ressources linguistiques : état des lieux*
- *Normalisation de la gestion des ressources langagières*
- *Bilan et travail de thèse*

1.1. Ressources linguistiques : état des lieux

Aujourd'hui, les ressources linguistiques (RL) jouent un rôle très important dans les applications de la technologie des langues. En effet, d'une part les RL alimentent les différents processus des systèmes de TAL, et d'autre part ces ressources sont de plus en plus utilisées pour accompagner le travail de modélisation linguistique par des méthodes statistiques, qui tiennent une position de plus en plus importante dans les applications de TAL (*cf*. Romary [ROM 00b]).

Les conférences LREC[1] (*International Conference on Language Resources and Evalutaion*) sur les RL et leur évaluation ont montré l'intérêt réservé aux RL par les équipes de recherche dans tous les domaines : annotation morphosyntaxique, grammaires, corpus arborés, terminologie et connaissances, sémantique, web sémantique et ontologies, pragmatique, dialogue multimodal… Toutes les applications comme la traduction automatique, la recherche et l'extraction d'information, la classification de documents, la détection de thèmes, la fouille de données, l'e-éducation, le résumé automatique ou les systèmes de question-réponse nécessitent souvent des ressources volumineuses. Dans le monde de la recherche, les questions d'acquisition et d'utilisation de données, ainsi que d'évaluation et de normalisation des ressources, des outils et des systèmes ont ainsi acquis une importance centrale.

Les RL à grande échelle connaissent une diffusion croissante, notamment grâce à des structures comme le LDC[2] (*Linguistic Data Consortium*) aux Etats-Unis et l'ELRA[3] (*European Language Resources Association*) en Europe.

Dans cette section, nous donnons un aperçu des ressources linguistiques existant au niveau international, ce qui constitue le contexte de notre projet de recherche. Nous nous limitons aux ressources concernant l'analyse automatique des corpus textuels. Les sujets abordés sont donc les lexiques (section 1.1.1), les grammaires (section 1.1.2), les corpus de textes monolingues bruts et annotés (sections 1.1.3 et 1.1.4), et les corpus multilingues alignés (section 1.1.5). Dans chacune de ces catégories, nous nous concentrons, d'une part, sur les ressources consacrées aux langues indo-européennes, qui sont naturellement les plus étudiées, et, d'autre part, sur les langues asiatiques dites « isolantes »[4], famille à laquelle appartient le vietnamien. À titre d'exemple, nous présentons dans ces deux catégories les ressources développées pour le français et l'anglais, et le chinois (officiel, c'est-à-dire mandarin) et le thaï[5], respectivement. Nous présentons également, lorsqu'elles existent, les ressources existant pour le vietnamien.

La question de la normalisation des outils et ressources linguistiques est ensuite discutée à la section 1.2.

---

[1] http://www.lrec-conf.org/

[2] http://www.ldc.upenn.edu/

[3] http://www.elra.org/

[4] Nous définissons cette notion au chapitre 2, lors de la présentation des principes de base du vietnamien.

[5] On peut noter à ce sujet que dans le passé, le vietnamien a été considéré par les linguistiques comme appartenant à la même famille que le thaï (*cf*. Maspero [MAS 12]).

1.1.1. Lexiques

Il s'agit des lexiques opérationnels monolingues ou multilingues conçus pour servir de données dans des outils de TAL. Par exemple, les lexiques monolingues sont des ressources indispensables pour les analyses linguistiques (morphologique, syntaxique, sémantique) des documents. Dans un cadre multilingue, les lexiques multilingues sont essentiels pour les systèmes de traduction automatique.

Un lexique se compose d'une liste d'entrées lexicales auxquelles peuvent être associées des informations linguistiques comme la morphologie, la syntaxe, ou la sémantique de l'entité lexicale décrite, sa fréquence d'usage, des exemples d'emploi, *etc*. On distingue deux types d'informations lexicales : d'une part, les informations intralexicales (constituant la micro-structure du lexique) rassemblent les descriptions de type morphologique, syntaxique, sémantique et pragmatique de chaque entrée lexicale ; d'autre part, les informations interlexicales (constituant la macro-structure du lexique) représentent les relations entre entrées lexicales, qu'elles soient d'ordre morphologique (lien entre une forme fléchie et son lemme), syntagmatique (collocations) ou paradigmatique (synonymes, antonymes, hypéronymes, *etc*.). Les lexiques opérationnels peuvent être construits manuellement par des experts, ou de manière (semi-)automatisée, à partir de dictionnaires traditionnels ou de corpus annotés.

Nous présentons dans cette section quelques uns des lexiques les plus connus dans le domaine du TAL, en abordant dans un premier temps les lexiques monolingues pour nous focaliser ensuite sur les expériences visant la construction de lexiques multilingues.

1.1.1.1. Lexiques monolingues

L'étude des langues indo-européennes bénéficiant d'une plus longue expérience, c'est naturellement pour celles-ci que les lexiques les plus aboutis ont été constitués, couvrant la totalité du champ de la description lexicale, de la morphologie à la sémantique. Nous décrivons donc, à titre de référence, quelques lexiques développés pour les langues indo-européennes (en nous limitant pour l'exemple au français et à l'anglais), avant de présenter les travaux en cours pour les langues asiatiques isolantes.

Langues indo-européennes

De nombreux modèles de lexiques ont été définis avec plus ou moins de généralité (*cf*. Francopoulo [FRA 03], Romary *et al.* [ROM 04]). Nous introduisons ici, par ordre de « complexité » croissante (de la morphologie à la sémantique), ceux d'entre eux qui peuvent prétendre au statut de « standard », en ce sens qu'ils sont devenus des références largement reconnues dans le domaine du TAL.

Le lexique français BDLEX et les bases lexicales multilingues Européennes CELEX, MULTEXT et « MULTEXT goes East » traitent principalement de morphologie. BDLEX (*cf*. De Calmès et Pérennou [CAL 98]), conçu pour le traitement morphologique et également phonologique, contient 440 000 formes fléchies générées à partir d'environ 50 000 formes canoniques (entrées lexicales) avec les informations sur la prononciation et la morpho-syntaxe (*cf*. Figure 1-1). BDLEX dispose par ailleurs de statistiques lexicales représentées par un ensemble d'indices de fréquences d'origine diverses. CELEX (*cf*. Burnage [BUR 90]) est une large base contenant des informations lexicales de plusieurs types (lemme, formes fléchies, abréviations et corpus) pour l'anglais, l'allemand et le néerlandais. CELEX dispose également de l'information concernant la prononciation des formes. Les projets MULTEXT (*cf*. Ide et Véronis [IDE 94]) et « MULTEXT goes East » (*cf*. Erjavec *et al.* [ERJ 96]) visent le développement de systèmes d'analyse morphologique comparables grâce à un modèle de représentation de descriptions grammaticales ayant un noyau commun pour les langues européennes. Ce modèle de descriptions grammaticales (*cf*. Figure 1-2) permet de définir et comparer les jeux d'étiquettes morphologiques. Le lexique MULTEXT du français a servi de base pour l'évaluation des systèmes d'analyse morphosyntaxique du français dans le cadre du projet GRACE (*cf*. 3.2.5.2).

Graphie	Prononciation		Morpho syntaxe			
ORTHO	PHONO	FPH	CS	VS	M	LIEN
prendre	pRa~dR	@	V		inf	=
prennent	pREn	@t"	V	3P	pi	prendre
petites	p@tit	@z"	J	FP		Petit
Un	9~	n"	D	MS	di	=
Avion	avjo~		N	MS		=

PHONO : représentation phonologique, FPH : fonctionnement phonologique de la finale, CS : catégorie syntaxique, VS : variation syntaxique, M : mode, LIEN : entrée lexicale (lemme) dont la forme est dérivée.

Figure 1-1 Structure lexicale des entrées de BDLEX

Attribute	Value	Example	Code
Type	main	partir	m
	auxiliary	avoir	a
Mood/Vform	indicative	viens	i
	subjunctive	vienne	s
	imperative	viens	m
	conditional	viendrais	c
	infinitive	venir	n
	participle	venu	p
Tense	present	viens	p
	imperfect	venais	i
	future	viendrai	f
	past	vins	s
Person	first	suis	1
	second	es	2
	third	est	3
Number	singular	viens	s
	plural	venons	p
Gender	masculine	venu	m
	feminine	venue	f
Clitics	///	///	-

Figure 1-2 Attributs spécifiés des verbes du modèle MULTEXT

aimer	
CB	P0 PSelf (P1)
SELF	catgram VERB trait_l [aux:avoir]
P0	NP PRONOUN[lex:quelqu'un]
PSelf	V[aux:avoir]
P1	NP S[introd:le fait que] S[mood:infinitive] S[mood:infinitive][prep:à] S[mood:infinitive][prep:de] S[sbcat:complementizer][mood:subjunctive] PRONOUN[lex:le] PRONOUN[lex:quelqu'un] PRONOUN[lex:quelque chose]

Figure 1-3 Exemple de la description syntaxique d'une unité lexicale dans GENELEX

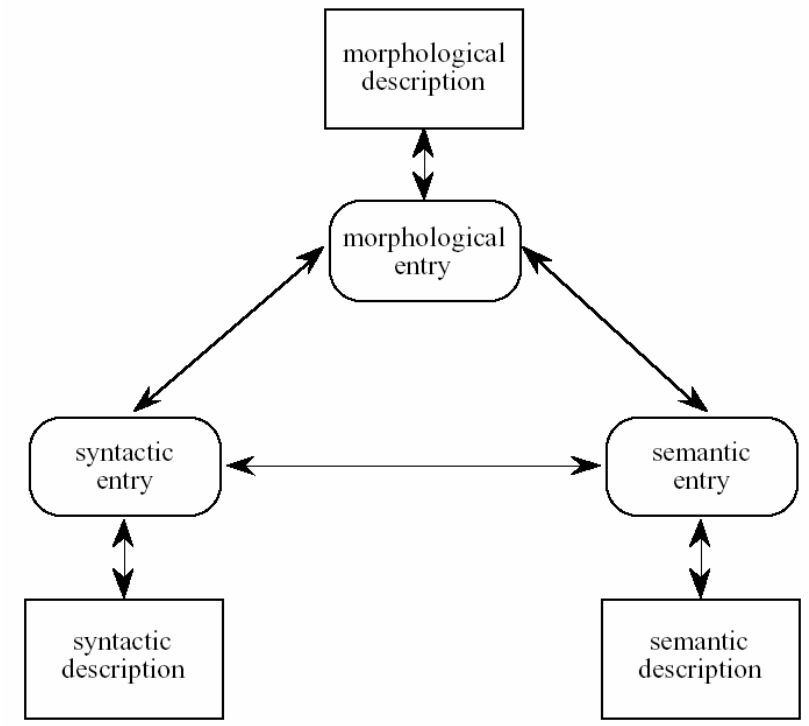


Figure 1-4 Vue réduite du modèle GENELEX

Les modèles Européens complexes[6], dont l'original est GENELEX (*GENEric LEXicon*, projet EUREKA, Antoni-Lay *et al.* [ANT 93]) fournissent pour chaque lemme (unité lexicale) une information très riche : le comportement syntaxique et la sous-catégorisation (*cf.* Figure 1-3), ainsi que la sémantique. Ils sont puissants en terme de généricité et de possibilité d'usage multiple. Dans le modèle GENELEX, chaque entrée est représentée sous forme d'un graphe de relation entre entités lexicales (morphologique, syntaxique, sémantique, *cf.* Figure 1-4 – Sérasset [SER 93]). GENELEX n'est directement lié à aucune application de TAL, mais une application de TAL peut extraire pour ses besoins particuliers une partie de l'information contenue dans cette base lexicale très large. Le développement de modèles dérivés de GENELEX pour les langues européennes fait l'objet de nombreux projets, notamment dans le cadre du groupe EAGLES (*Expert Advisory Group for Language Engineering Standards*).

Plus spécifiquement orienté vers la semantique, le thésaurus WordNet (anglo-américain, construit depuis 1985 à l'Université de Princeton – Miller *et al.* [MIL 90b]) contient environ 200 000 paires de mot-sens. Les mots sont organisés en classes de synonymes, ou *synsets*, dont chacun représente un concept lexical (*cf.* Figure 1-5). Ces synsets sont eux-mêmes organisés en une arborescence ontologique structurée par la relation d'hyponymie (*cf.* Figure 1-6), ainsi que par des liens transversaux marquant d'autres types de relations syntaxiques (antonymie, métonymie…). WordNet joue un rôle important dans de nombreux travaux en étiquetage sémantique ou qui visent l'accès aux textes par le sens.

Une autre base lexicale (anglaise) orientée sémantique est en cours de développement dans le cadre du projet FrameNet à Berkeley – Baker *et al.* [BAK 03]. L'objectif de FrameNet est de documenter les liens entre les unités lexicales (paires mot-sens) et leur cadre sémantique, en se basant sur des usages observés en corpus (principalement le *British National Corpus* – BNC). Chaque unité lexicale est accompagnée de ses définitions et des exemples annotés sensés illustrer toutes ses possibilités combinatoires (*cf.* Figure 1-7), et liée à un cadre sémantique, qui peut-être partagé par d'autres unités lexicales. FrameNet contient actuellement plus de 8 900 unités lexicales, dont plus de 6 100 sont complètement annotées dans 625 cadres sémantiques, et exemplifiés dans plus de 135 000 phrases annotées. FrameNet contient également un réseau de relations entre les cadres (*cf.* Figure 1-8, Fillmore *et al.* [FIL 04]). La base est disponible sous licence par le biais de son site Internet. D'autres projets dérivés pour l'allemand, l'espagnol et le japonais sont également en cours.

Langues asiatiques isolantes

Le premier lexique utilisé pour le traitement du chinois est le lexique syntaxique du groupe CKIP (*Chinese Knowedge Information Processing* [CKIP 93]), qui rassemble environ 80 000 entrées de mots chinois. À chaque entrée lexicale sont associées sa catégorie syntaxique et ses rôles thématiques dans la théorie ICG (*Information-based Case Grammar*, *cf.* 1.1.2.2).

Au niveau sémantique, plusieurs réseaux de concepts chinois ont été développés. On peut citer, en particulier, la base SKCC (*Semantic Knowledge base of Contemporary Chinese*) de l'Institut d'Informatique Linguistique de l'université de Pékin. Cette base (*cf.* Wang et Yu [WAN 03]) contenant 66 539 mots chinois est construite suivant le modèle du thésaurus WordNet.

Dans le cadre du ChineseLDC (*Chinese Linguistic Data Consortium* – Zhao *et al.* [ZHA 04], http://www.chineseldc.org), deux lexiques ont été développés :

- Un premier lexique construit contient environ 100 000 mots, accompagnés d'informations précisant leur transcription « pinyin » (pseudo-phonétique en alphabet occidental) et leur fréquence. Les fréquences des mots sont évaluées en se basant sur deux statistiques : les fréquences calculées à partir d'un corpus segmenté de 5 millions de caractères chinois, et les fréquences de chaînes d'un corpus brut d'un milliard de caractères.

[6] Les consortiums les ayant développé ont, pour satisfaire les exigences de tous leurs partenaires, réalisé l'union de nombreux mécanismes de représentation, ce qui a rendu complexe la structure de ces modèles.

- Une deuxième base lexicale est la base de connaissances grammaticales chinoises concernant les mots courants. Cette base se compose d'environ 30 000 mots chinois fréquemment utilisés, couvrant tous les mots de catégories grammaticales ambiguës, et tous les mots outils. À chaque mot d'entrée sont associées ses étiquettes morphosyntaxiques possibles, leurs fréquences relatives, une suite d'attributs grammaticaux décrivant l'usage du mot et un ensemble de phrases d'exemple. Les mots de la base sont extraits du corpus journalistique chinois *1998 People Daily*.

Pour le thaï, le lexique monolingue NAiST[7] Lexibase (*Kasetsart University*) contient 15 000 mots accompagnés par des informations syntaxiques et sémantiques. Le NAiST Lexibase est fondé sur un modèle relationnel (*cf.*Figure 1-9, Kawtrakul *et al.* [KAW 95]).

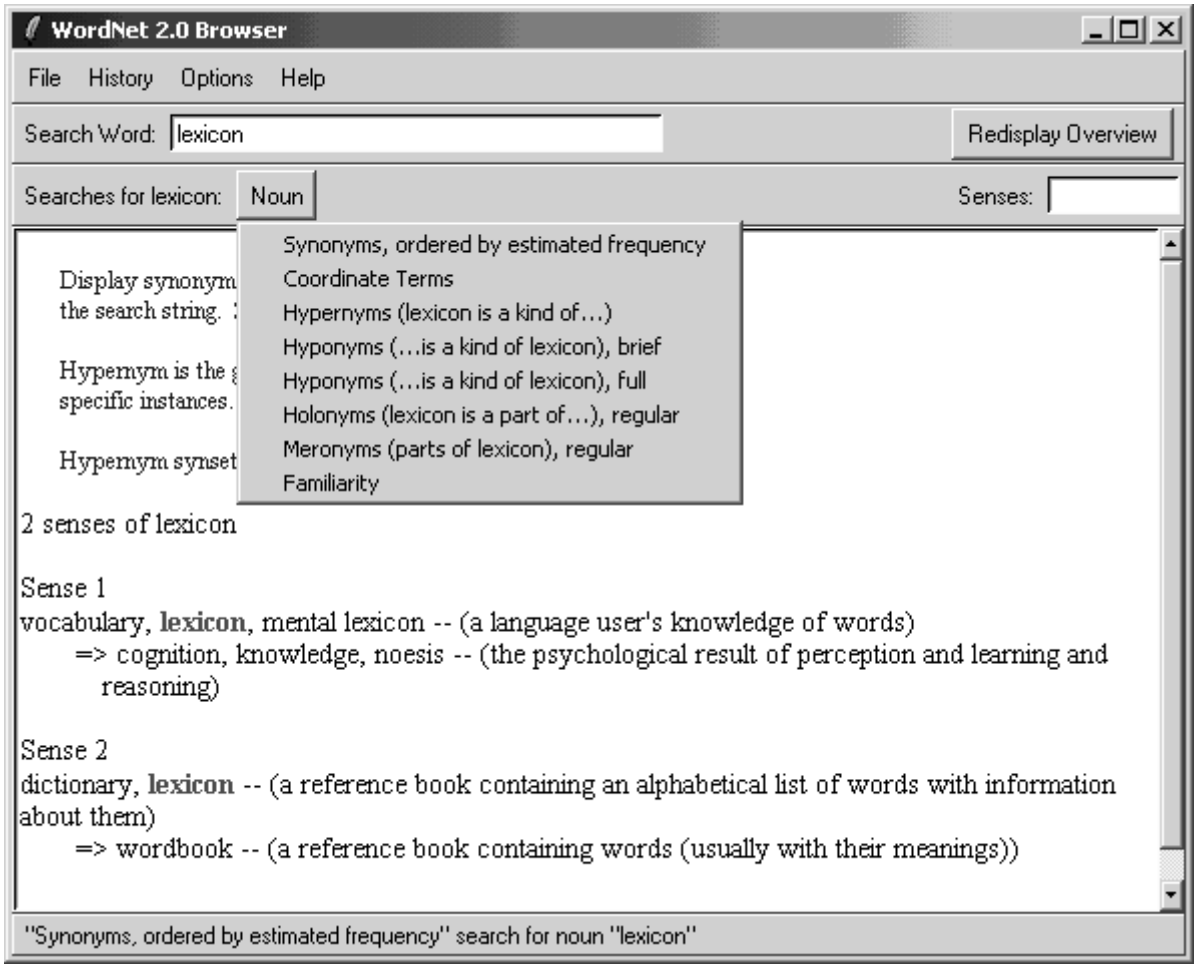


Figure 1-5 Exemple de consultation de WordNet

[7] Natural Language Processing and Intelligent Information System Technology Research Laboratory

Sense 2
dictionary, lexicon
=> wordbook
=> reference book, reference, reference work, book of facts
=> book<<<<
=> publication
=> work, piece of work
=> product, production
=> creation
=> artifact, artefact
=> object, physical object
=> entity
=> whole, whole thing, unit
=> object, physical object
=> entity

Figure 1-6 Exemple de hiérarchie hyperonymique dans WordNet

[SPEAKER We] **informed** [ADDRESSEE the press]
[MESSAGE that the prime minister has resigned]

[SPEAKER We] **informed** [ADDRESSEE the press]
[MESSAGE of the prime minister's resignation]

Figure 1-7 FrameNet – Exemples annotés du cadre sémantique du verbe « *inform* » [FIL 04]

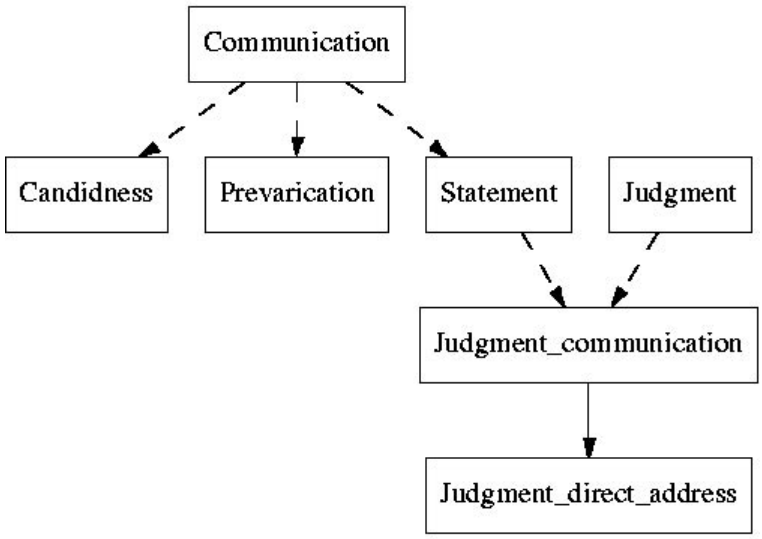


Figure 1-8 FrameNet – Exemples de relations de cadres sémantiques [FIL 04]

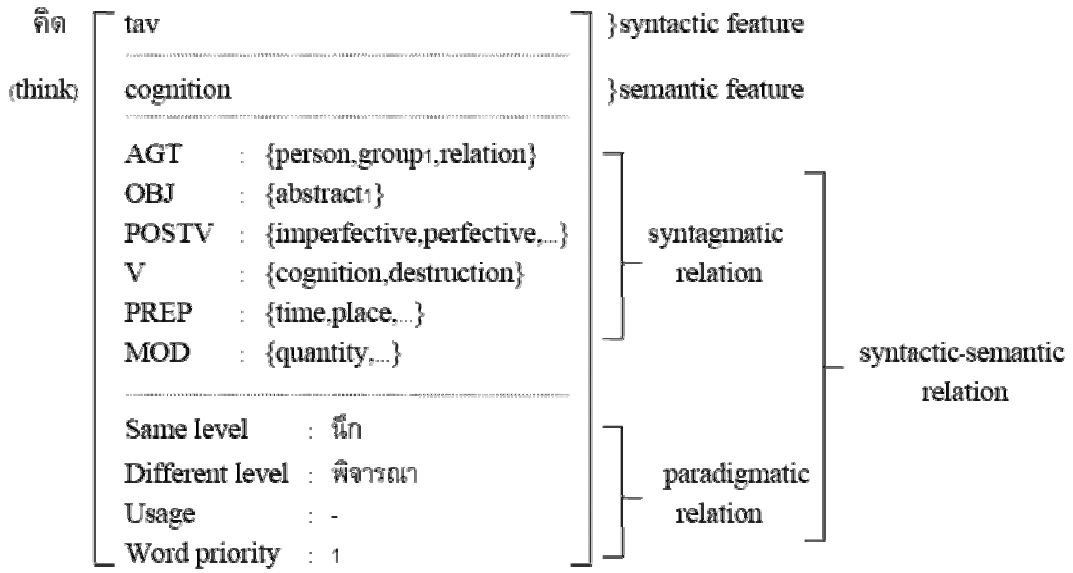


Figure 1-9 Matrice de lexique pour le NAiST Lexibase (thaï)

1.1.1.2. Lexiques multilingues

On peut distinguer parmi les lexiques multilingues ceux qui s'intéressent en particulier à la mise en correspondance de deux langues, souvent dans un objectif précis (lexiques bilingues), et ceux dont l'objectif plus ambitieux est de développer un mécanisme générique pouvant permettre la mise en parallèle d'informations lexicales pour un nombre *a priori* arbitraire de langues.

Lexiques bilingues

Le modèle bilingue EDR (*Electronic Dictionary Research*) est spécifiquement destiné au couple japonais-anglais. Ce modèle consiste en différents dictionnaires : de mots, de concepts, de co-occurrences, et bilingue. L'architecture du dictionnaire bilingue EDR (détaillée à la Figure 1-10) se base principalement sur un dictionnaire de concepts où des concepts indépendants des langues sont décrits et reliés aux entrées lexicales monolingues dans chaque langue. Les entrées lexicales monolingues sont enregistrées dans deux dictionnaires de mots (pour l'anglais et pour le japonais) qui fournissent leur information grammaticale (représentée comme une liste d'attributs) et un lien à un concept du dictionnaire de concepts. Chaque entrée lexicale est une forme fléchie de mot, ce qui n'est pas très efficace pour les langues fortement flexionnelles comme le français (qui compte en moyenne environ 10 formes par lemme – Sérasset [SER 93]).

Concernant les ressources lexicales pour les langues asiatiques isolantes considérées (le chinois et le thaï), on peut mentionner deux lexiques mis à disposition du public : le lexique de traduction chinois-anglais distribué par le LDC et le lexique pour le couple de langues thaï et anglais. Le lexique chinois-anglais a été rendu public pour la première fois en 1999 ; sa dernière version date de 2002 et contient 54 170 entrées du chinois. Le format du lexique est simple :

<mot chinois> /<traduction anglaise 1>/…/<traduction anglaise *n*>/

Le lexique LEXITRON (voir Figure 1-11 et Figure 1-12), dont la première publication date de 1996, se focalise pour sa part sur le couple thaï-anglais, mettant en correspondance 53 000 entrées lexicales anglaises et 35 000 entrées thaïs, extraites d'un large corpus (*cf.* Charoenporn *et al.* [CHA 04]). Il s'agit néanmoins plus d'un dictionnaire bilingue informatisé, destiné à une consultation humaine, que d'un lexique conçu pour le TAL ; il ne présente donc qu'un intérêt limité de notre point de vue.

```
<English-Japanese Bilingual Dictionary>
                ::= <English-Japanese Bilingual Dictionary Record>BBB
<English-Japanese Bilingual Dictionary Record>
                        ::=     <Record Number>
                                \t<Headword Information>
                                \t<Grammatical Information>
                                        \t<Semantic Information>
                                        \t<Correspondence Information>
                                \t<Management Information>\n
<Record Number>         ::= <Character String>
<Headword Information>  ::= <Headword>
<Headword>              ::= <Character String>
<Grammatical Information>       ::= <Part of Speech>
<Part of Speech>                ::= <Character String>
<Semantic Information>          ::= <Concept Identifier>\t<Headconcept>
                                    \t<Concept Explication>
        <Concept Identifier>    ::= <Hexadecimal Integer>
        <Headconcept>           ::= <English Headconcept>
                                    \t<Japanese Headconcept>
                <English Headconcept>
                                ::= <Character String>
                <Japanese Headconcept>
                                ::= <Character String>
        <Concept Explication>   ::= <English Concept Explication>
                                    \t<Japanese Concept Explication>
                <English Concept Explication>
                                ::= <Character String>
                <Japanese Concept Explication>
                                ::= <Character String>
<Correspondence Information>    ::=<Correspondence Word Information>
                                   | <Correspondence Information>
                                   // <Correspondence Word Information>
        <Correspondence Word Information>
                                ::=<Correspondence Word Category>
                                  '|'<Correspondence Word Notation>
                                        '|' x
                <Correspondence Word Category>
                                ::=<Number>
                <Correspondence Word Notation>
                                ::=<Character String>
<Management Information>        ::=<Management History Record>
<Management History Record>     ::=<Attribute Name>=<Attribute Value>
                                   | <Management History Record>
                                ; <Attribute Name>=<Attribute Value>
        <Attribute Name>        ::=<Character String>
        <Attribute Value>       ::=<Character String>
```

Figure 1-10 Structure de données des entrées du dictionnaire Anglais-Japonais (EDR)

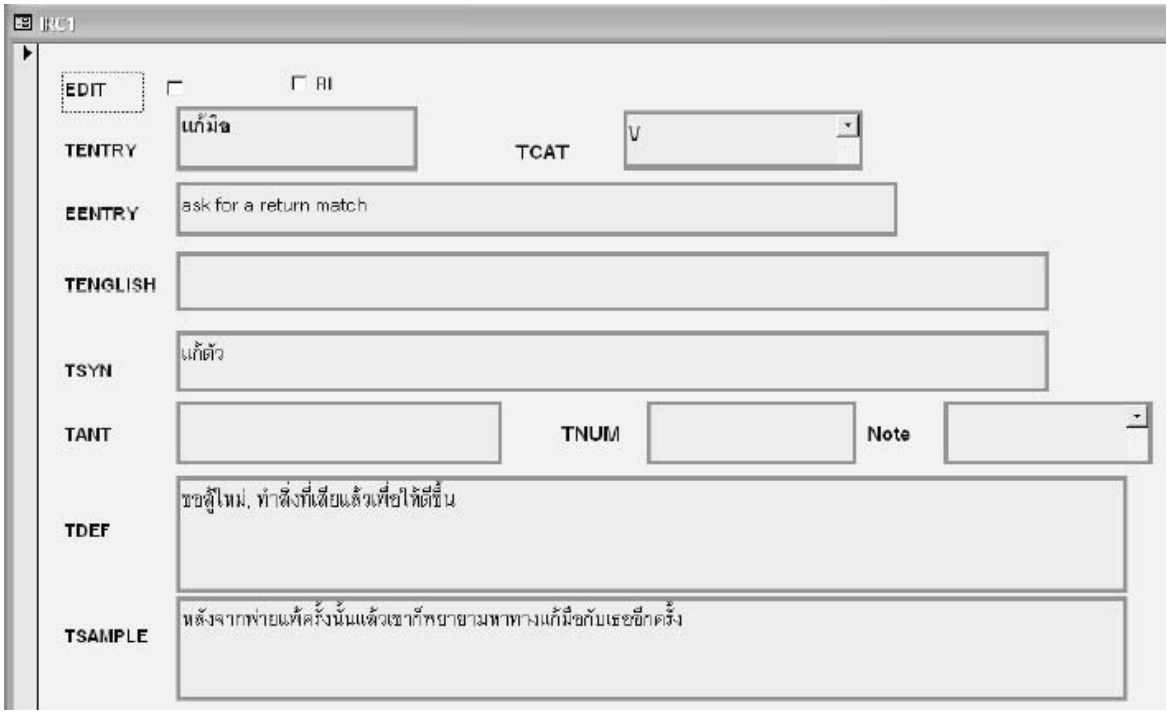


Figure 1-11 Exemple d'édition d'une entrée dans Lexitron

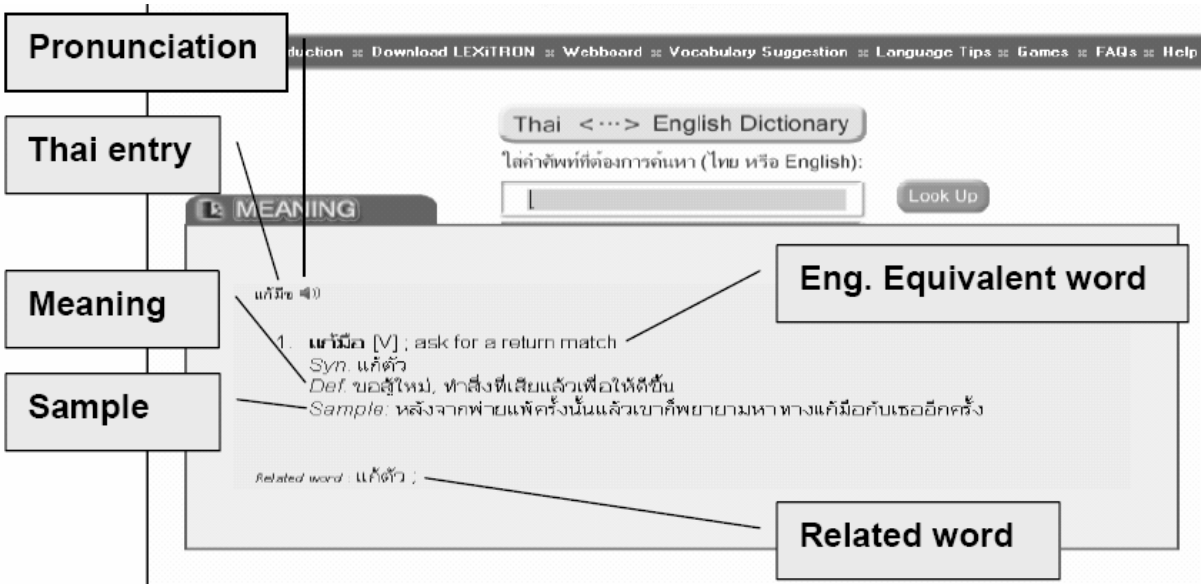


Figure 1-12 Exemple de données dans Lexitron

Projets multilingues

Une piste souvent empruntée pour le développement de lexiques multilingues consiste à étendre un modèle monolingue existant afin de permettre l'encodage de liens entre langues. Ainsi, le principe de description des entrées lexicales développé dans le projet GENELEX a donné lieu à plusieurs extensions visant à permettre la mise en parallèle de lexiques construits pour différentes langues. C'est en particulier le cas du projet Européen SIMPLE (*Semantic Information for Multilingual Plurifunctional Lexica*) qui permet le rapprochement de lexiques en différentes langues en définissant un vocabulaire commun pour la description d'informations lexicales sémantiques. Un autre « produit dérivé » de GENELEX est le projet ISLE (*International Standards for Language Engineering*) / MILE (*Multilingual ISLE Lexical Entry*) du groupe EAGLES.

Dans le même ordre d'idées, de nombreux lexiques dérivés de WordNet (EuroWordNet pour les langues d'Europe de l'Ouest, ItalWordNet pour l'italien, IndoWordNet pour l'Asie et BalkaNet pour les langues de l'Europe de l'Est) visent le développement d'une ontologie de haut niveau qui puisse être commune à toutes les langues qu'ils traitent. Ils définissent en outre un index interlingual des synsets permettant la mise en correspondance directe de ceux-ci d'une langue à l'autre.

Le projet Papillon (*cf*. Boitet [BOI 01], Mangeot *et al.* [MAN 03]), en revanche, est conçu dès l'origine avec le multilinguisme comme objectif. Il a pour but de créer une base lexicale multilingue ouverte et coopérative comprenant entre autres l'anglais, le français, le japonais, le malais, le lao, le thaï et le vietnamien. L'idée est de permettre aux utilisateurs un accès libre à la base sur Internet, et la possibilité de participer à son enrichissement. La macrostructure du dictionnaire est composée d'un volume monolingue pour chaque langue et d'un volume pivot contenant des liens interlinguaux reliant les sens des mots composant les volumes monolingues (*cf*. Figure 1-13). Pour chacune des langues étudiées, la microstructure des articles (*cf*. Figure 1-14) est fondée sur la lexicographie combinatoire extraite de la théorie sens-texte de Mel'cuk (DEC *Dictionnaire explicatif et combinatoire*, *cf*. Mel'cuk *et al.* [MEL 84, 88]).

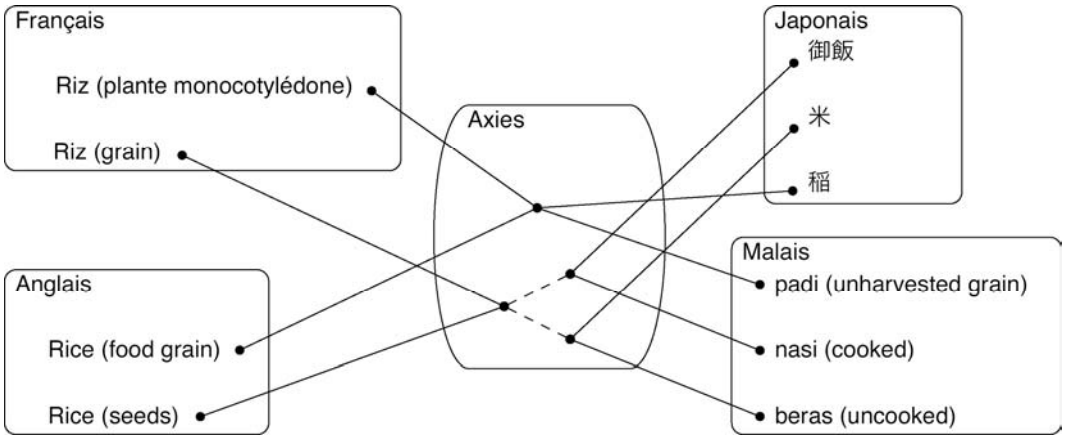


Figure 1-13 Liens entre la traduction du mot « riz » dans quatre langues de la base Papillon [MAN 03]

regretter ,

v.tr.

sentiment LA personne X ~ SON action Y

GOVERNMENT PATTERN

X = I Y = II
1 . N
1 . N
2 . de V-inf

LEXICAL FUNCTIONS

QSyn : se repentir

S0 : regret#1

Able2 : (*Que l'on peut R.*) regrettable

Magn : (*Intensément*) beaucoup

Y étant grave, Magn : amèrement , cruellement ; _se mordre les doigts_

EXAMPLES

1 . *C'est une décision qu'il va regretter cruellement.*

2 . *Il ne regrette pas d'avoir investi 4 000 F dans ce nouveau programme.*

Figure 1-14 Forme inspirée du DEC pour la lexie « regretter.1 » du dictionnaire Papillon

1.1.2. Grammaires à large couverture

Les grammaires permettent d'analyser et de modéliser la structure syntaxique de la phrase, et ainsi de préciser les relations existant entre ses composants, ce qui constitue un apport d'information très important pour l'accès au sens. Les recherches menées en TAL ont déclenché le besoin de grammaires à large couverture (*cf.* Abeillé et Blache [ABE 00]), c'est à dire permettant de modéliser sinon toute la langue, du moins une part importante de celle-ci. Ces grammaires doivent décrire, pour une langue comme le français :

- les dépendances locales : accord, sous-catégorisation des prédicats, expressions semi-figées, restrictions modifieur-modifié, clitiques, *etc*.,
- les dépendances « moyennes » : pronominalisation, contrôle des infinitives, association négative, quantifieurs flottants, *etc*.,
- les dépendances à distance : questions, relatives, constructions disloquées, *etc*.,
- les alternances syntaxiques : passif, impersonnel, causatives, *etc*.,
- les phénomènes de coordination et de comparaison.

Plusieurs projets ont été entrepris et menés à bien pour la création de grammaires, principalement pour les langues indo-européennes, qui sont les plus étudiées. Nous présentons ici quelques exemples de celles-ci, puis décrivons rapidement la grammaire du chinois développée par le Sinica, qui constitue à notre connaissance le seul exemple d'une telle ressource pour les deux langues asiatiques auxquelles nous nous intéressons.

1.1.2.1. Langues indo-européennes

Pendant longtemps, les seules grammaires électroniques de qualité représentaient le travail d'un petit groupe de chercheurs pendant de nombreuses années et étaient étroitement dépendantes de l'application envisagée (par exemple la traduction automatique) et des programmes de traitement.

La stabilisation des modèles syntaxiques basés sur l'unification a permis le développement de plates-formes d'implantation multilingues, librement accessibles, qui incorporent des outils de test et de mise à jour et permettent le développement de grammaires neutres quant à leur application.

Pour l'anglais, le projet XTAG (*cf.* [XTA 01]) a développé et distribué (sous licence GPL[8]) une grammaire électronique « réutilisable » à grande échelle. XTAG utilise le formalisme LTAG (*Lexicalized Tree Adjoining Grammar*, *cf.* section 4.3.1.2). La grammaire se compose des familles d'arbres initiaux représentant les cadres de sous-catégorisation, ainsi que des arbres auxiliaires permettant l'adjonction des modifieurs dans les syntagmes.

Pour le français, on peut citer les lexiques et grammaires de l'équipe de Maurice Gross [GRO 75] qui ne sont pas directement formalisés pour l'analyse syntaxique, mais constituent une base de connaissances lexico-syntaxiques irremplaçable. Une autre référence est la grammaire FTAG du français fondée sur l'architecture XTAG, développée par l'équipe Talana (université de Paris VII) et utilisée en analyse et en génération (*cf.* Abeillé [ABE 02]).

1.1.2.2. Langues asiatiques isolantes

Pour le chinois, depuis 1993, l'équipe CKIP de l'Académie Sinica (Taiwan) a développé un analyseur syntaxique basé sur le formalisme ICG (*Information-based Case Grammar*). Il s'agit d'un formalisme de grammaire d'unification (*cf.* section 4.2.1.2). Les traits considérés pour les éléments lexicaux dans la grammaire CKIP correspondent à des informations syntaxiques mais aussi sémantiques, le chinois n'ayant pas de flexion. La relation de précédence de constituants est définie par les rôles thématiques.

---

[8] GNU General Public License, *cf.* http://www.gnu.org/copyleft/gpl.html

En ce concerne l'analyse syntaxique du thaï, aucune grammaire publiquement disponible n'existe aujourd'hui à notre connaissance.

1.1.3. Corpus de textes bruts et étiquetés

« Un corpus, ou collection de textes, peut être vu comme un échantillon d'une langue et c'est à ce titre que les corpus sont utilisés pour le traitement automatique des langues naturelles. Plus le corpus est étendu et varié, plus l'échantillon est représentatif. » – Laporte [LAP 00]. Les corpus de textes représentent un usage réel de la langue, et fournissent donc une référence objective pour vérifier ou même acquérir des descriptions formelles de la langue. Les corpus de référence doivent satisfaire deux caractéristiques : une taille suffisamment grande et la diversité des usages présentés. Leur utilisation comme source d'information apporte une aide irremplaçable à la construction de dictionnaires et de grammaires.

Les corpus de textes étiquetés associent à chaque mot des textes qu'ils rassemblent des informations grammaticales ou morphologiques. Ces ressources sont cruciales pour les études ultérieures comme le découpage du texte en groupes syntagmatiques, son analyse syntaxique, l'élaboration de concordances, *etc*. ; elles peuvent également être employées par des applications « finales » comme le résumé automatique de textes ou l'extraction de terminologie.

On peut distinguer deux grandes catégories de corpus suivant le type de langue représenté : les corpus de spécialités tentent de refléter l'usage de la langue dans un domaine particulier (corpus techniques, médicaux), tandis que les corpus généralistes s'intéressent à l'ensemble d'une langue et rassemblent souvent des textes plus variés, représentatifs de sa diversité. Nous ne nous intéressons ici qu'à cette seconde catégorie, et présentons une fois encore des exemples pour les langues indo-européennes, d'une part, et asiatiques isolantes de l'autre.

1.1.3.1. Langues indo-européennes

Les premiers corpus étiquetés ont été construits pour l'anglais américain, le plus ancien et plus connu étant le corpus de Brown (Kucera et Francis [KUC 67]), qui rassemble un million de mots étiquetés manuellement. Par sa mise dans le domaine public, ce corpus a joué un rôle moteur pour les recherches basées sur les corpus. Son équivalent pour l'anglais britannique est le corpus de *Lancaster-Oslo-Bergen* (LOB).

Le BNC (*British National Corpus*) contenant 100 millions de mots (dont 90% relèvent de la langue écrite et 10% de la langue orale) fournit une ressource de grande échelle pour l'anglais britannique. Le corpus contient des textes de fiction et des textes informatifs venant de livres, périodiques, discours, *etc*. Le corpus BNC a également été étiqueté (Leech *et al.* [LEE 94], Burnard [BUR 95]).

L'ANC (*American National Corpus*, *cf.* Ide et MacLeod [IDE 01a]) en est l'équivalent pour l'anglais américain. Le but est d'obtenir un corpus d'au moins 100 millions de mots, comme le BNC, équilibré du point de vue des types de textes rassemblés. La première édition de l'ANC est un corpus de 10 millions de mots (dont plus de 8 millions de mots relèvent de la langue écrite et le reste de la langue orale), annotés pour le lemme et la partie du discours. Les textes sont automatiquement étiquetés sans validation humaine (en employant un étiqueteur automatique standard, dont la précision est d'au moins 95%). Cette première édition a pour but de recevoir les critiques sur la structure et l'annotation du corpus.

Pour le français, le corpus le plus volumineux est la base FRANTEXT « Trésor de la langue française », constituée depuis les années soixante, qui contient 3 737 textes, soit environ 210 millions d'occurrences de mots. Ce corpus est à portée principalement littéraire et historique : 80% de ses textes correspondent à des œuvres littéraires, et 20% à des publications scientifiques ou techniques, du XVIe au XXe siècle.

En ce qui concerne les corpus étiquetés, la campagne d'évaluation d'étiqueteurs automatiques GRACE-Multitag a été l'occasion du développement du plus volumineux corpus de référence pour le français (un million de mots). Les étiquettes sont principalement celles définies par le projet MULTEXT (qui a pour sa part occasionné le développement d'un corpus étiqueté de 200 000 mots). D'autres projets ont permis la construction de corpus partiellement étiquetés, comme par exemple PAROLE (250 000 mots sur environ 2 millions) ou CLIF (300 000 mots sur 20 millions).

1.1.3.2. Langues asiatiques isolantes

Pour le chinois, le corpus équilibré Sinica[9] (*Academia Sinica Balanced Corpus of Modern Chinese*), contruit depuis 1995, constitue la première base (10 millions mots de texte brut, dont un million de mots étiquetés). Sa version 3.0, distribué sur le web en 1997, rassemble 5 millions de mots (corpus étiqueté avec 46 étiquettes réduites de 178 catégories syntaxiques du lexique syntaxique CKIP, *cf*. 1.1.1.1). La Figure 1-15 présente deux exemples de phrases étiquetées extraites du corpus Sinica, accompagnées de leur traduction anglaise[10].

En 2002, l'Institut d'Informatique Linguistique de l'université de Pékin a achevé l'étiquetage d'un corpus spécialisé de 26 millions caractères chinois (*1998's People's Daily*). Un sous-corpus étiqueté de plus de 2 millions de caractères a été distribué gratuitement sur leur site web en 2001.

Le corpus étiqueté proposé par le ChineseLDC (Zhao *et al.* [ZHA 04]) est un corpus de 5 millions de caractères chinois, contenant des articles de journaux, des oeuvres littéraires, des livres scientifiques, *etc*. Ce corpus est segmenté en mots et étiqueté morpho-syntaxiquement de manière semi-automatique, de manière suivante :

- Collecte et classification d'un corpus de 5 millions de mots ;
- Spécification de la segmentation et de l'étiquetage lexical des textes chinois (20 grandes classes et 51 sous-classes) ;
- Collecte de la liste de mots pour la segmentation et l'étiquetage ;
- Développement d'outils pour la segmentation et l'étiquetage ;
- Développement d'un système d'aide à la vérification manuelle du corpus automatiquement traité.

Pour le thaï, le corpus NAiST (*Kasetsart University*) contient environ 60 millions de mots.

Le corpus étiqueté thaï, nommé ORCHID, est construit en Thaïlande depuis 1996 par le NECTEC[11], en collaboration avec le CRL[12] japonais. Le corpus est annoté en trois niveaux : paragraphes, phrases et mots. La segmentation en paragraphes et phrases est manuelle, tandis que la segmentation en mots et l'étiquetage lexical sont automatiques mais suivis d'un contrôle manuel. Le corpus contient 2 560 000 mots, et le jeu d'étiquettes consiste en 14 catégories et 47 sous-catégories. Les balises utilisées dans ce corpus sont exposées à la Figure 1-16, et la Figure 1-17 présente un texte extrait de ce corpus (*cf*. Charoenporn *et al.* [CHA 04]).

Pour le vietnamien, le Centre de Lexicographie du Vietnam (Vietlex) a construit une base de textes anciens et modernes contenant environ 50 millions de syllabes (2 millions de phrases). Cette base rassemble des textes appartenant à tous les genres : littérature (40,5%), articles de journaux (53,7%), droit, sciences sociales, sciences naturelles (5,8%). Ce corpus a notamment été employé afin de constater les contextes effectifs d'emploi des mots pour le travail de construction d'un dictionnaire du vietnamien. Des négociations sont à l'heure actuelle en cours afin de rendre ces ressources disponibles à la communauté de recherche publique en TAL au Vietnam.

---

[9] http://godel.iis.sinica.edu.tw/CKIP/engversion/20corpus.htm

[10] Merci à Chu-Ren Huang (Académie Sinica) de nous avoir fourni ces exemples.

[11] National Electronics and Computer Technology Center.

[12] Communications Research Laboratory.

1. 五色鳥**(Na)**喜歡**(VK)**像**(P)**詩人**(Na)**般**(Ng)**的**(DE)**在**(P)**森林**(Na)**中**(Nc)**鳴唱**(VC)**。

Muller's barbet**(Na)** fond**(VK)** like**(P)** poet**(Na)** sort**(Ng)** DE**(DE)** in**(P)** forest**(Na)** middle**(Nc)** sing**(VC)**

Muller's barbets are fond of singing like a poet in the forest.

2. 爲什麼**(D)**有**(V_2)**「包種茶」**(Na)**這**(Nep)**個**(Nf)**名稱**(Na)**呢**(T)**？

Why**(D)** have**(V_2)** " Formosa Oolong Tea"**(Na)** this**(Nep)** piece**(Nf)** name**(Na)** NE**(T)** ?

Where does the name "Formosa Oolong Tea" derive from?

Figure 1-15 Deux exemples du corpus étiqueté SINICA

Tag	Description
%TTitle:	Title written in Thai
%ETitle:	Title written in English
%TAuthor:	Authors' name written in Thai
%EAuthor:	Authors' name written in English
%TInbook:	Book name of the article written in Thai
%EInbook:	Book name of the article written in English
%TPublisher:	Publisher's name written in Thai
%EPublisher:	Publisher's name written in English
#Pn	Paragraph number counted from the beginning to the end of the article
#n	Sentence number counted from the beginning to the end of the article
\\	Line break within a sentence
//	Sentence end
word/POS	Word, delimiter ("/") and the corresponding POS

Figure 1-16 Schéma de balisage du corpus ORCHID

%TTitle: คาร์บอนไดออกไซด์เลเซอร์กำลังสูงแบบไหลเวียนตามแนวแกน

%ETitle: High-Power Compact Axial Flow CO2 Laser

%TAuthor: ผศ.พิพัฒน์ โชคสุวัฒนสกุล

%EAuthor: [Asst. Prof. Pipat Choksuwatanasakul]

%TInbook: การประชุมทางวิชาการ ครั้งที่ 6 โครงการวิจัยและพัฒนาอิเล็กทรอนิกส์และคอมพิวเตอร์ ปีงบประมาณ 2536

%EInbook: The 6th NECTEC Annual Conference

%TPublisher: ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ กระทรวงวิทยาศาสตร์ เทคโนโลยีและสิ่งแวดล้อม

%EPublisher: National Electronics and Computer Center, Ministry of Science Technology and Environment

:

#P5

:

#4

ดังนั้นกำลังของเลเซอร์ต่อหน่วยความยาวจึงสามารถเพิ่มขึ้นได้ค่อนข้างมาก//

ดังนั้น/JSBR

กำลัง/NCMN

ของ/RPRE

เลเซอร์/NCMN

ต่อ/RPRE

หน่วย/NCMN

ความ/FIXN

ยาว/VATT

จึง/XVBM

สามารถ/XVAM

เพิ่ม/VATT

ขึ้น/XVAE

ได้/XVAE

ค่อนข้าง/ADVN

มาก/ADVN

//

#5

ในการวิจัยครั้งนี้เราได้ลองศึกษาการเกิดดิสชาร์จจากลักษณะของรูปทรงของแคโถดที่ใช้ต่างๆ กัน\\

พบว่าการใช้แคโถดเป็นรูปทรงกระบอกกลวงทำให้เกิดกระแสในการดิสชาร์จ//

ใน/RPRE

การ/FIXN

วิจัย/VACT

:

การ/FIXN

ดิสชาร์จ/VACT

//

:

Figure 1-17 Extrait d'un texte étiqueté du corpus thaï ORCHID

1.1.4. Corpus arborés : *Treebanks*

Les corpus arborés (*Treebanks*) sont des corpus de référence annotés syntaxiquement ; pour les plus petits d'entre eux, cette tâche peut être réalisée manuellement, pour les plus volumineux, une première analyse automatique doit être suivie d'une validation manuelle. Ces corpus peuvent avoir des utilisations multiples en TAL (*cf.* Abeillé [ABE 00]), par exemple pour l'évaluation des étiqueteurs et des parseurs, pour l'induction de grammaires ou de préférences syntaxiques, pour la construction automatique d'étiqueteurs ou de parseurs probabilistes, ou pour l'enrichissement des dictionnaires électroniques (extraction de collocations, extraction de cadres de complémentation). De tels corpus permettent également d'obtenir une documentation détaillée sur les annotations attendues pour les principales constructions rencontrées dans les textes mais négligées dans les grammaires.

L'ouvrage édité par Abeillé [ABE 03a] présente amplement un état de l'art de la construction et de l'utilisation de corpus arborés.

1.1.4.1. Langues indo-européennes

Le développement de l'informatique linguistique a commencé plus tôt aux États-Unis, ce qui explique une avance considérable en matière de corpus pour l'anglais. Les projets Treebank (http://www.cis.upenn.edu/~treebank/home.html) sont des exemples de création de grands corpus annotés. L'université de Pennsylvanie aux États-Unis a fait construire le *Penn Treebank*, distribué par le LDC, qui contient environ 4 millions de mots, dont 2 millions de mots sont analysés pour la structure prédicat-argument. Un autre corpus de taille moins importante est également mis à disposition des acteurs de la recherche en TAL : le corpus Suzanne, constitué de 128 000 mots extraits du corpus Brown (1995), accompagné également d'une annotation sémantique.

Le corpus *Bank of English*, qui rassemble 200 millions de mots annotés par une grammaire de dépendance est également une référence importante depuis 1995 (*cf.* Järniven [JAR 03]), mais ce corpus n'est malheureusement pas disponible publiquement.

Pour le français, l'équipe TaLaNa dirigée par A. Abeillé à l'Université de Paris 7 a entrepris depuis 1997 le projet CORFRANS de construction d'un corpus textuel annoté syntaxiquement, en collaboration avec l'équipe LATL (Laboratoire de Genève) et l'équipe RALI (laboratoire de Montréal). Ce projet a donné lieu à la constitution d'un corpus journalistique d'un million de mots (*cf.* Abeillé *et al.* [ABE 03b]).

Devant la faible quantité de corpus arborés existant pour le français et la difficulté de leur exploitation, le projet FREEBANK (*cf.* Salmon *et al.* [SAL 04]) a l'ambition de construire une base ouverte de corpus du français annotés à plusieurs niveaux : structurel, morphologique, syntaxique, coréférentiel. En particulier, cette base est conçue dans un esprit de libre accès, de respect des normes de codage, et de possibilité d'enrichissement progressif par une communauté d'utilisateurs. La base initiale est constituée d'environ un million de mots de différentes sources : littéraires (oeuvres libres de droits de FRANTEXT, *cf.* 1.1.3.1), journalistiques (du quotidien *l'Est Républicain*), scientifiques (thèses, articles scientifiques), techniques (journal du CNRS), *etc*.

1.1.4.2. Langues asiatiques isolantes

En 2000, deux corpus arborés du chinois sont simultanément publiés : le *Penn Chinese Treebank* et le *Sinica Treebank*.

Le *Penn Chinese Treebank*, projet de construction d'un corpus arboré journalistique du chinois mené par l'université de Pennsylvanie, a vu sa version 4.0 diffusée en 2004 par le *Linguistic Data Consortium*. Elle contient plus de 400 000 mots, et est structurée suivant le modèle du *Penn Treebank* pour l'anglais.

Le *Sinica Treebank* adopte le modèle ICG qui a été développé précédemment pour l'analyse syntaxique du chinois. La première version du *Sinica Treebank* rassemble environ 240 000 mots, extraits du corpus Sinica (*cf.* 1.1.3.2). Une partie du corpus est disponible sur le web, alors que le corpus complet peut être obtenu sous licence.

Le but du ChineseLDC (Zhao *et al.* [ZHA 04]), est de créer un corpus *Chinese Treebank* qui se compose d'environ un million d'idéogrammes. Le corpus sélectionné est enrichi par une annotation syntaxique arborescente et est construit en 4 étapes :

- Spécification de la construction de Treebank : 16 étiquettes de fonction syntaxique et 33 étiquettes de structure syntaxique ;
- Collecte du corpus segmenté et étiqueté morpho-syntaxiquement, avec vérification humaine ;
- Développement d'outils d'étiquetage automatique des tronçons (*chunks*) fonctionnels syntaxiques, avec validation humaine ;
- Développement d'un logiciel d'étiquetage automatique des tronçons de structure syntaxique, avec validation humaine.

Pour le thaï, aucune ressource n'est, à notre connaissance, publiée.

1.1.5. Corpus multilingues alignés

Les corpus multilingues parallèles sont constitués de plusieurs exemplaires d'un même texte dans diverses langues. Accompagnés par une annotation des équivalences de traduction, ces corpus peuvent être utilisés pour la création de bases de données multilingues, la consolidation des lexiques, la construction et la validation de mémoires de traduction, *etc*.

La construction de corpus parallèles est relativement facile pour les langues occidentales beaucoup pratiquées, comme l'anglais et le français : une assez longue « histoire » de numérisation de documents textuels et de constitution de corpus monolingues pour les langues considérées rend souvent possible l'exploitation de données existante – même si subsiste dans ces conditions les difficultés liées à l'existence de traduction inexactes ou incomplètes. Pour les langues moins pratiquées, ou pratiquées dans les pays moins développés technologiquement, la constitution de corpus parallèles peut demander beaucoup plus d'efforts (*cf.* Singh *et al.* [SIN 00]) : la quantité de textes parallèles (numérisés ou non) peut être beaucoup moins importante, les documents sont moins couramment numérisés, et enfin, dans le cas des langues n'employant pas l'alphabet latin, les codages des textes dans les différentes langues considérées peuvent s'avérer incompatibles.

1.1.5.1. Langues indo-européennes

Il existe de nombreux projets visant à la compilation de corpus de textes parallèles, dont les principaux sont (*cf.* Véronis [VER 00b]):

- le corpus *Hansard* (français-anglais), construit en 1980 et rassemblant 50 millions de mots extraits des délibérations du Parlement canadien ; c'est l'un des premiers corpus de textes parallèles, et le plus connu ;
- le corpus de textes trilingues (français, anglais, espagnol) de l'*International Telecommunications Union CCITT Handbook* (13,5 millions de mots) et de l'*International Labour Organisation* (5 millions de mots) ;
- le corpus construit dans le cadre des projets MULTEXT-MLCC des textes du Parlement de la Communauté Européenne (9 langues : danois, allemand, anglais, espagnol, français, grec, italien, néerlandais, portugais), contenant environ 70 millions de mots.

Pourtant, les corpus déjà alignés et vérifiés sont beaucoup moins nombreux. Trois millions de mots du *CCITT Handbook* sont alignés par le projet CRATER (*cf.* Garside *et al.* [GAR 94]), un million de mots sont alignés au niveau des phrases en cinq langues dans le cadre du projet MULTEXT (*cf.* Ide et Véronis [IDE 94]). Dans le cadre du projet ARCADE I (*cf.* 0), deux corpus multilingues ont été mis au point pour l'évaluation :

- Corpus BAF (*cf.* Simard [SIM 98]), bi-texte anglais-français comportant environ 400 000 mots dans chaque langue, élaboré au CITI (Montréal), dans le cadre d'une ARC (Action de Recherche Concertée) financée par l'AUPELF-UREF (aujourd'hui AUF, « Agence Universitaire de la Francophonie »). Les textes du BAF sont classés en 4 catégories : institutionnels, scientifiques, techniques et littéraires.
- Corpus JOC : questions écrites posées par des membres du Parlement Européen, et réponses données par la Commission, publiées en 1993, dans les Séries C du Journal officiel de la Communauté européenne. Ces questions concernent des sujets très variés : agriculture, économie, environnement, institutions, droits de l'homme, transports, *etc*. Le corpus JOC comporte des textes parallèles en 9 langues, avec environ 1,1 millions de mots par langue. Un cinquième des textes français et anglais a été aligné sous forme de bi-texte, avec validation manuelle.

Le projet ARCADE II [CHI 06], continuation du précédent, a été l'occasion du développement de nouveaux corpus, étendant le champ d'investigation à de nombreuses autres langues, y compris non indo-européennes. Ayant participé à ce projet, nous le présentons plus en détail à la section 5.7.

1.1.5.2. Langues asiatiques isolantes

Les corpus parallèles chinois-anglais constituent la grande majorité des corpus parallèles construits autour du chinois. En effet, plusieurs corpus de textes parallèles sont distribués sur le site du LDC, la plupart venant de Hong Kong.

Le corpus multilingue proposé par le ChineseLDC est un corpus chinois-anglais contenant des textes de journaux, des oeuvres littéraires, des livres scientifiques, *etc*. Ce corpus parallèle contenant 270 000 paires de phrases alignées est construit selon les étapes suivantes :

- Spécification du corpus parallèle, comprenant les informations de l'entête et du corps des textes, avec référence au jeu de méta-données Dublin Core (DCMS - *Dublin Core Metadata Set*) ;
- Collecte du corpus parallèle et étiquetage des informations d'en-tête ;
- Développement d'outils d'alignement automatique des phrases ;
- Alignement automatiquement des phrases ;
- Vérification manuelle du résultat.

Pour le thaï, aucun corpus parallèle n'est publiquement disponible.

Nous avons jusqu'ici présenté de nombreux types de ressources linguistiques dans le domaine de TAL. Pour faciliter l'échange de données et économiser ainsi le coût de construction de ressources énormes, se pose évidemment la question de normaliser la représentation de ces ressources. C'est le sujet de la section suivante.

1.2. Normalisation de la gestion des ressources langagières

La normalisation de la gestion des ressources langagières consiste à définir des méthodes et des modèles pour faciliter l'échange de données, l'interopérabilité entre des composants logiciels, et l'évaluation des résultats. D'un point de vue technologique, la normalisation vise à la stabilisation des pratiques existantes, et à l'anticipation des évolutions technologiques. D'un point de vue organisationnel, elle doit obtenir un consensus international, assurer une disponibilité à long terme et une maintenance des normes.

Les normes et standards sont introduits par les instances principales suivantes :

- instances officielles de normalisation au niveau national (AFNOR en France, ANSI aux États-Unis, TCVN au Vietnam, *etc*.) ou international (ISO – *International Standardization Organization*, W3C – *World Wide Web Consortium*, spécialisé en réglementation et contrôle des aspects techniques liés à l'utilisation d'Internet, *etc*.) ;
- forums académiques et industriels (par exemple : LISA – *Localization Industry Standards Association*, TEI – *Text Encoding Inititative*, *cf*. 1.2.1.2) ;
- projets à visées pré-normatives (par exemple les projets EAGLES et MULTEXT déjà évoqués dans ce chapitre).

Deux visions de la normalisation sont distinguées : horizontale et verticale. La normalisation horizontale fournit les briques de base nécessaires à un champ d'activité, par exemple les échanges de données sur la Toile ; alors que la normalisation verticale définit un ensemble des spécifications nécessaires à un domaine d'application, par exemple le domaine du TAL. Voici quelques exemples des normes existantes concernant le domaine des langues.

Les normes horizontales :

- Codage des caractères – ISO 10646 / Unicode : codage universel et uniforme pour chacun des caractères utilisés dans l'ensemble des langues vivantes.
- Codes de langue – ISO 639 : système de codage des noms de langues par deux caractères (par exemple `fr` pour le français, `vn` pour le vietnamien, *etc*.).
- Codage des documents structurés : SGML (ISO), XML (W3C), Recommandations de la TEI (présentés ci-après).

Les normes verticales :

- Activités liées aux services web (W3C) : WSDL (*Web Services Description Language*), SOAP (*Simple Object Access Protocol*).
- Activités liées au web sémantique (W3C) : RDF (*Resource Description Framework*), RDFS (*RDF Schema*), OWL (*Web Ontology Language*), *etc*.
- Codage de l'information multimédia (ISO) : série MPEG (*Moving Picture Expert Groups*) – MPEG-1 (base des formats Video CD et MP3), MPEG-2 (base de la télévision numérique et du DVD), MPEG-4 (multimédia pour les web fixe et mobile), MPEG-7 (description et recherche du contenu audio et visuel) et MPEG-21 (Cadre de multimédia), *etc*.
- Terminologie (au sein du comité technique 37 de l'ISO : ISO/TC 37/SC 3 – *Terminology and other language resources/Computer application in terminology*) : ISO 12200 (MARTIF – *MAchine-Readable Terminology Interchange Format*), ISO 12620 (Catégories de données), ISO 16642 (Cadre de balisage terminologique).

Nous procédons maintenant à une présentation plus précise des normes existantes et en cours de définition qui sont impliquées dans notre projet, à savoir les normes/recommandations pour le codage structuré de ressources (méta-langages SGML et XML, directives de la TEI), et pour la gestion des ressources langagières.

1.2.1. Codage des documents structurés

1.2.1.1. SGML (*Standard Generalized Markup Language*)

SGML (norme ISO 8879:1986, *cf.* Goldfarb [GOL 91]) est un méta-langage de balisage qui spécifie des règles permettant la définition de systèmes de balises pour le codage de divers types de document électronique et d'éventuelles informations associées.

Trois concepts fondamentaux permettent de définir un document SGML :

- Éléments : SGML représente les données textuelles avec des éléments de contenu, de types différents et encapsulés les uns dans les autres. Le modèle de base pour la représentation de données au format SGML correspond donc à un arbre hiérarchique. Chaque nœud correspond à un élément SGML, et chaque feuille correspond ainsi à un contenu élémentaire (*cf.* Figure 1-18 et l'explication qui la suit). En règle générale, les éléments du texte sont encadrés par des balises ouvrantes et fermantes, du type `<balise> … </balise>`. Ces balises peuvent contenir des attributs fournissant une description de l'élément textuel concerné, et qui se placent sur la balise ouvrante : `<balise attribut = "valeur"> … </balise>`.
- Types de documents : SGML impose que toute instance de document soit conforme à une DTD (*Document Type Definition*) qui lui est systématiquement associée. Cette DTD précise les balises autorisées et les agencements légaux et hiérarchiques de ces balises. La DTD fournit également, pour chaque type d'élément, la liste des attributs qu'il est possible d'utiliser, ainsi que le type de leur valeur, et éventuellement, une liste de valeurs prédéfinies. Cela assure la consistance du codage de ressources.
- Entités : Il s'agit de chaînes de caractères nommées qui sont lors de l'interprétation du document remplacées par leur définition, à la manière des *macros* ou *alias* en programmation informatique. Une entité peut remplacer un seul caractère spécial, mais aussi une chaîne contenant des éléments SGML complexes (à condition que ces éléments soient entièrement définis dans la chaîne).

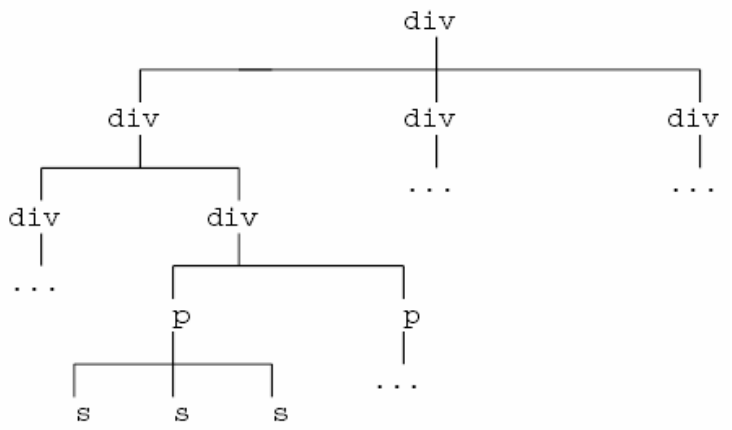


Figure 1-18 Structure arborée d'un document simple

L'exemple de la Figure 1-18 expose la structure logique des documents textuels ordinaires pouvant être représentés en SGML. En effet, un texte se compose de phrases (noeuds `s` dans l'arbre), groupées en paragraphes (noeuds `p`). Ces paragraphes peuvent être groupés en sections, puis à leur tour, les sections en chapitres, parties, *etc*. Ces éléments de texte (section, chapitre, partie, *etc*.) sont classées dans un même type d'élément récursif : division (noeuds `div` dans l'arbre).

Le balisage de documents textuels électroniques en SGML permet d'une part de distinguer l'interprétation des parties de texte et leur format d'impression, et d'autre part d'extraire facilement des parties de contenu de texte à la demande. Il assure également l'échange et la pérennité de ces ressources grâce à un codage explicite de l'information, indépendant de tout périphérique ou application.

En pratique, la norme SGML est souvent critiquée pour sa lourdeur et sa complexité de mise en œuvre, principalement en raison de l'exigence de définition rigoureuse d'une DTD et de conformité du document à celle-ci, ainsi qu'à cause du manque de flexibilité de sa syntaxe. La recommandation XML, conçue originalement pour la distribution de documents structurés sur Internet, est devenue une solution alternative largement employée.

1.2.1.2. XML (*eXtensible Markup Language*)

La recommandation XML (Bray *et al.* [BRA 98]), dérivée du format SGML, a été développée par un groupe de travail XML formé sous les auspices du W3C en 1996. Il était présidé par Jon Bosak de Sun Microsystems avec la participation active d'un groupe d'intérêts particuliers XML (connu précédemment sous le nom de groupe de travail SGML), également sous l'égide du W3C.

Par construction, les documents XML sont des documents SGML conformes. À la différence de SGML, un document XML peut ne pas contenir de référence à une DTD, d'où se distinguent les notions de document *valide* (respectant un schéma ou une DTD donné) et de document *bien formé* (respectant la structure d'un document SGML/XML). Il est donc possible, à partir d'un document primaire d'origine (validé au regard d'une DTD connue), de n'en transmettre qu'une partie qui soit pertinente pour un traitement donné ou suite à une requête d'un utilisateur. À l'inverse, des documents ou parties bien formées de documents issus de sources différentes peuvent être re-combinées pour former un nouveau document (*cf*. Romary [ROM 00b]).

Les éléments d'un document structuré, représenté par le format SGML/XML, peuvent être identifiés grâce aux mécanismes de référencement des applications manipulant les documents SGML/XML. Cette identification se fait soit directement par des identifieurs uniques spécifiés dans un attribut porté par l'élément cible ; soit de façon relative en utilisant des localisations dans l'arbre SGML/XML pour pointer vers des éléments spécifiques en utilisant la structure du document (par exemple *le 2^{e} paragraphe du 4^{e} chapitre*) – Bonhomme [BON 00a].

Le W3C propose également les recommandations annexes de XML qui permettent l'accès à la structure des documents XML :

- *XPath*, *XPointers* et *XLink* – mécanismes de localisation et de lien de fragments de document XML
- *XSLT* – langage de transformation de feuilles de style XSL, permet d'exprimer des requêtes ou sélections de contenu des éléments.

La force de XML est de devoir son succès à l'Internet et de disposer ainsi d'une communauté extrêmement vaste d'utilisateurs travaillant avec des conventions terminologiques communes. De nombreux développements logiciels accompagnent le déploiement de ce standard. Cela permet au XML de devenir un format puissant, capable de représenter tout type de ressource.

La possibilité de définir des DTD adaptées à chaque tâche particulière est l'une des principales sources de puissance et de flexibilité des métalangage SGML et XML. Le revers de cette médaille est la possibilité de multiplication de DTD incompatibles pour représenter les mêmes types de documents. Ainsi apparaît la problématique de définir des DTD « normalisées » pour chaque usage, permettant l'échange et la mise en commun d'informations de sources variées. C'est le travail entrepris par la TEI (*Text Encoding Initiative*), dont nous présentons maintenant les travaux.

1.2.1.3. TEI (Text Encoding Initiative)

La TEI (Ide et Sperberg-McQueen [IDE 95a]) est un consortium académique international, créé en 1987, dans le but de développer les recommandations pour le codage et l'échange de données linguistiques et littéraires. En mai 1994, le travail effectué par les différents comités a été publié sous forme de « Recommandations pour le codage et l'échange des textes informatisés » (*Guidelines for the Encoding and Interchange of Machine-Readable Texts*), aussi connues sous le nom de TEI P3, reposant sur les DTD du SGML.

Ces directives proposent un ensemble de conventions de codage utilisables dans une grande variété d'applications : publication électronique, analyse littéraire et historique, lexicographie, traitement automatique des langues, recherche documentaire, hypertexte, *etc*. Les directives concernent les textes écrits ou parlés, sans restriction de langue, de période, de genre ou de contenu et répondent aux besoins fondamentaux de nombreux d'utilisateurs : lexicographes, linguistes, philologues, bibliothécaires, et de manière générale, de tous ceux qui sont concernés par l'archivage et l'accès à des documents électroniques.

Trois aspects du codage des textes sont particulièrement mis en avant par la TEI :

- documentation de textes : les documents TEI doivent fournir obligatoirement les informations bibliographiques sur le texte lui-même et son codage. Ces informations sont balisées dans la partie en-tête « TEIheader » se trouvant au début de chaque document codé en TEI (*cf*. Figure 1-19).
- représentation de textes : la TEI propose un système de balises pour coder la description de structure logique de différents types de document (textes écrits ou parlés, prose littéraire, poésie, théâtre, dictionnaires, données terminologiques, hypermédia *etc*.)
- analyse et interprétation de textes : les directives de la TEI contiennent des jeux de balises pour le codage des références croisées ou des index dans les textes, des analyses linguistiques et des informations concernant l'étude littéraire.

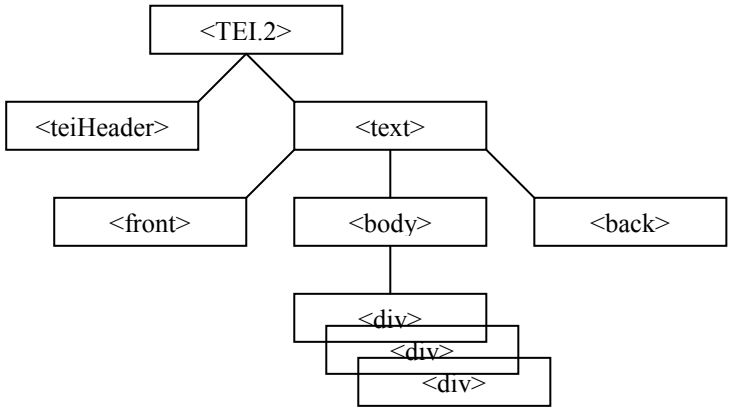


Figure 1-19 Structure TEI de base de textes courants [BON 00a]

Depuis sont sorties la version TEI P4 en 2002 reposant sur le schéma XML, et récemment la version encore plus modulaire TEI P5 en 2004.

1.2.2. Gestion des ressources langagières

Avec la maturité de développement des standards dans le domaine de langues (TEI, EAGLES/ISLE, LISA[13], *etc*.), l'ISO a validé en août 2002 la création d'un sous-comité TC 37/SC 4[14] entièrement dédié à la normalisation de la gestion des ressources linguistiques, sous la présidence de Laurent Romary. Le SC 4 a pour but de développer des principes et méthodes pour la création, l'encodage, le traitement et la gestion des ressources langagières comme des corpus écrits ou oraux, des lexiques, des schémas de classification. Les centres d'intérêts sont : la modélisation de données, le balisage, l'échange de données et l'évaluation des ressources langagières à part des terminologies (traitées précédemment par d'autres sous-comités du TC 37).

Le SC 4 est organisé en cinq groupes de travail (WG – *Work Group*) :

- WG 1 : Descripteurs et mécanismes de base pour les ressources linguistiques – Ce groupe a pour but de définir le vocabulaire commun au SC 4, d'établir la méthodologie générale à la définition de schémas d'annotation ou de représentation de données linguistiques, et de décrire un schéma de méta-données pour les ressources linguistiques.
- WG 2 : Schémas de représentation – L'objectif est de fournir des cadres normatifs pour la représentation de données morphosyntaxiques et syntaxiques, pour la description de contenus sémantiques multimodaux, ainsi que pour les annotations aux niveaux discursifs ou dialogiques.
- WG 3 : Représentation de données textuelles multilingues – Ce groupe vise à fournir des standards pour la représentation des mémoires de traduction et des textes parallèles alignés, la segmentation et la numération des textes en unités lexicales, la gestion de systèmes hétérogènes d'annotation dans des contextes d'activités de globalisation, d'internationalisation et de localisation.
- WG 4 : Bases lexicales – Le but est de fournir les cadres de description des différents formats de représentation de données lexicales pour des applications de TAL.
- WG 5 : Environnement de gestion de ressources linguistiques – Ce groupe a la tâche de définir des directives pour la création, la validation et la distribution de ressources linguistiques.

Comme on peut constater par leurs objectifs, ces groupes de travail sont déterminés selon les types de ressources considérés : le WG 2 pour les corpus annotés, le WG 3 pour les ressources multilingues, le WG 4 pour les ressources lexicales.

À l'heure actuelle, plusieurs activités sont en cours. Nous présentons rapidement les principes généraux des projets du SC 4.

La méthodologie générale pour le codage des ressources linguistiques repose sur les principes génériques proposés par N. Ide et L. Romary [IDE 03] : un format linguistique est spécifié par la combinaison d'un méta-modèle et d'une sélection de catégories de données spécifiques à la description linguistique considérée. Ces principes ont été implémentés dans le domaine de la terminologie avec la norme ISO 16642:2003 (TMF – *Terminological Markup Framework*). Nous précisons maintenant la définition d'un méta-modèle et d'une sélection de catégories de données.

---

[13] Le groupe de travail LISA/OSCAR a proposé des standards concernant, par exemple, l'échange de données de mémoire de traduction (TMX – *Translation Memory eXchange*), l'échange de données terminologiques (TBX – *TermBase eXchange*).

[14] Le site officiel de ce sous-comité se trouve au http://www.tc37sc4.org. Les documents de travail sont mis sur le site au fur et à mesure des activités du comité.

Un méta-modèle exprime indépendamment de toute préoccupation de format la structure sous-jacente d'un modèle de données (bases lexicales, corpus monolingues annotés de différents niveaux : morphosyntaxique, syntaxique, sémantique, discours, *etc*., corpus multilingues, *etc*.). Par exemple, dans le cadre de SC 4, le projet MAF (*Morpho-syntactic Annotation Framework*) définit un méta-modèle pour l'annotation morphosyntaxique (*cf*. [ISO 05a]) ; le projet LMF (*Lexical Markup Framework*) définit un méta-modèle pour la représentation des lexiques opérationnels.

Une catégorie de données (*cf*. Ide et Romary [IDE 04]) est un descripteur élémentaire dans une structure linguistique ou un schéma d'annotation, par exemple « partie du discours », « genre grammatical », « féminin », *etc*. Afin de permettre l'échange et la réutilisabilité des ressources au sein d'un domaine, et d'augmenter la cohérence des systèmes, il est nécessaire de définir des catégories de données « consensuelles », d'usage généralisé. Un *répertoire de catégories de données* (DCR : *Data Category Registry*) rassemble ainsi les catégories de référence pour un domaine donné, en précisant pour chacune d'elles leur sémantique précise dans une ou plusieurs langues. La définition d'une *sélection de catégories de données* est l'opération consistant à circonscrire le sous-ensemble d'un DCR pertinent pour une tâche donnée.

Le DCR défini pour les RL par le comité TC 37 a pour but de servir de référence pour tous les standards (existants ou futurs) concernant la modélisation et l'échange de données dans le TC 37. Sa définition se base sur :

- l'utilisation et l'extension des catégories de référence pour la représentation de terminologie (ISO 12620:1999) et pour la représentation des langues (ISO 639) ;
- la prise en considération et l'extension des catégories définies dans le cadre des projets à visée pré-normatives comme EAGLES/ISLE ou MULTEXT, déjà évoqués.

Le TC 37 implémente pour l'instant un seul DCR[15] central qui couvre toutes les applications dans le domaine de création et d'usage de RL, publié par la norme ISO 12620 (*Terminology and other language resources – Data categories*). La construction de ce DCR est réalisée en deux étapes :

- sélection : rassembler les catégories définies par les experts dans chaque sous-domaine de RL,
- harmonisation : assurer la cohérence dans le DCR.

Au niveau de création et d'accès, le DCR est organisé suivant les vues thématiques (domaines d'activité). On peut envisager la définition des catégories au moins pour ces domaines : collection de données terminologiques, différents types d'annotation linguistique (morphosyntaxique, syntaxique, discours, *etc*.), représentation de lexique, méta-données de RL et codes de langues. Il existe bien évidemment des catégories propres à un domaine, et d'autres communes à plusieurs domaines. La définition du DCR doit assurer la possibilité de spécifier, à long terme, une ontologie de relations entre les catégories.

[15] Une interface prévue à la contribution du DCR de l'ISO/TC 37/ SC 4 est disponible à l'adresse http://syntax.inist.fr.

1.3. Bilan

De nombreuses ressources linguistiques ont été construites ces dernières années, fournissant les conditions d'un développement accéléré des activités de recherches dans les domaines concernés. C'est particulièrement le cas pour la langue anglaise, qui concentre l'attention d'une part très importante de la communauté scientifique.

En ce qui concerne le vietnamien, au moment où nous avons entrepris cette thèse, il n'existait aucun outil ni ressource linguistique pour le traitement automatique. Dans ce contexte, nous avons pour objectif de mettre en place les premières briques pour la construction de ces ressources fondamentales : lexique morphosyntaxique, corpus primaires, corpus étiquetés, grammaire, corpus multilingues. Il s'agit de concevoir un environnement de développement de ces ressources en termes de modèle/format et de contenu.

Nous devons donc développer à partir de rien tous les outils et ressources nécessaires à l'analyse lexicale et syntaxique du vietnamien. Conscient du travail gigantesque des linguistes ainsi que des informaticiens pour la construction de ressources linguistiques, nous tentons de créer une base de connaissances linguistiques ouverte au monde de la recherche en TAL. Pour faciliter l'échange et la maintenance de données, la question de la normalisation de la gestion de ressources est bien évidemment importante.

Nous présentons rapidement dans la sous-section suivante nos tâches de recherche, puis nous finissons ce chapitre en introduisant les projets de recherche dans lesquels nous nous avons été impliquée, et qui ont été l'occasion à la fois d'un enrichissement et d'une valorisation de notre travail.

1.3.1. Travail de thèse

La première tâche de notre thèse est la création des ressources lexicales et construction des outils pour la construction de corpus de textes vietnamiens annotés morpho-syntaxiquement (*cf*. Chapitre 3). Il s'agit, dans une première étape, de construire un lexique avec les descriptions lexicales qui nous permet de définir ultérieurement les jeux d'étiquettes comparables pour la tâche d'étiquetage morphosyntaxique. Ainsi, nous étudions les caractéristiques de la langue vietnamienne afin de choisir une définition convenable des unités lexicales, c'est-à-dire les entrées du lexique, et des descriptions lexicales appropriées à retenir dans le lexique. Dans la deuxième étape, nous développons des outils pour la segmentation et l'étiquetage morphosyntaxique de corpus vietnamiens, en évaluant l'application des méthodes simples sur les textes vietnamiens. D'autres outils d'assistance à la gestion du lexique et des corpus annotés du vietnamien sont également conçus. En particulier, la normalisation de codage des ressources est insistée pour la réutilisabilité des ressources construites.

La deuxième tâche est la recherche en vue du développement d'une grammaire et d'un analyseur syntaxique du vietnamien (*cf*. Chapitre 4). Il existe plusieurs approches, plus probabilistes ou plus linguistiques, pour l'analyse syntaxique, qui, à son tour, peut être de surface ou en profondeur en fonction des applications visées. Etant donné que les recherches en linguistique informatique n'ont pas été menées au Vietnam, nous avons pour but de fonder une base de ressources grammaticales pour une usage et développement à long terme. Nous choisissons un formalisme de grammaire parmi des formalismes conçu comme théorie syntaxique les plus courants et développons une grammaire, contenant un lexique syntaxique et des règles grammaticales, servie à l'analyse syntaxique du vietnamien. Nous analysons les structures syntaxiques du vietnamien et essayons de les modéliser avec le formalisme choisi. Ici encore, la normalisation de représentation des ressources nous préserve la possibilité de convertir nos ressources pour sa réinitialisation dans un autre système d'analyse syntaxique.

La troisième tâche est la recherche en vue de la réalisation d'un système d'alignement multilingue et de la construction d'un corpus multilingue de référence ayant le vietnamien pour langue pivot (Chapitre 5). Nous développons deux outils spécialisés pour l'alignement au niveau des phrases et celui au niveau des mots (unités lexicales). Pour l'alignement au niveau de phrase, nous disposons d'un outil fondant son analyse sur la structure hiérarchique des documents, qui s'est montré d'une grande efficacité pour le couple de langues français-anglais dans le cadre de la campagne d'évaluation ARCADE I. Notre première tâche est donc d'évaluer l'adaptation de cet outil aux textes français-vietnamiens. Nous développons ensuite un outil d'alignement au niveau des unités lexicales et évaluons celui-ci vis à vis de chaque couple de langues d'un texte multilingue français - vietnamien – anglais, afin de montrer la perspective de la technique utilisée. Une autre évaluation a été effectuée du fait de notre participation au programme ARCADE II, dont nous présentons également les résultats.

Ces travaux n'auraient pas pu être réalisés sans les conditions favorables des efforts collectifs : les projets de recherche auxquels nous avons participé tout au long de la thèse nous ont apporté des soutiens très importants. La section suivante présente ces projets.

1.3.2. Intégration dans les projets de recherche

Durant cette thèse, nous avons été impliquée dans le projet national vietnamien KC01-03 « Recherche et Développement de la technologie de reconnaissance, de synthèse et du traitement automatique du vietnamien » (d'octobre 2001 à mai 2004). Ce premier projet national concernant le traitement automatique du vietnamien comprend trois parties :

- Reconnaissance et synthèse de la parole du vietnamien
- Reconnaissance de l'écriture du vietnamien
- Traitement automatique de la langue vietnamienne

Cette troisième composante inclut notre projet de thèse. L'intégration de celui-ci dans ce projet national nous permet d'avoir les moyens financiers et institutionnels d'une collaboration étroite avec les linguistes vietnamiens du Centre de Lexicographie du Vietnam, au niveau non seulement des compétences mais aussi de l'utilisation de ressources lexicales.

Nous suivons également les activités de normalisation de la gestion des ressources linguistiques de l'ISO/TC 37/SC 4 et de l'évaluation d'outils d'alignement de la campagne ARCADE II.

Concernant l'ISO/TC 37/SC 4, nos travaux dans le cadre de cette thèse nous ont amenée à participer (en tant que représentante du Vietnam, membre « observateur » de ce comité) aux activités suivantes :

- catégories de données (DCR) : catégories grammaticales ;
- schéma de représentation des lexiques opérationnels (LMF) ;
- annotation morphosyntaxique (MAF).

La campagne ARCADE II est la suite logique de l'action ARCADE I (Action de Recherche Concertée sur l'Alignement de Documents et son Évaluation) financée par l'AUPELF-UREF (*cf.* Véronis et Langlais [VER 00a]). ARCADE II se propose d'identifier les évolutions récentes de l'état de l'art, mais également d'approfondir l'évaluation sur des axes qui n'avaient pas été traités ou qui avaient seulement été effleurés par l'action précédente : identification des ruptures de parallélisme, alignement de tri-textes, élargissement à des langues présentant de fortes dissimilarités avec le français, identification des cognats, appariement lexical. Nous faisons partie des six participants officiels du projet, en présentant notre système d'alignement multilingue.

Nous introduisons au chapitre suivant les notions élémentaires du vietnamien, avant de présenter nos travaux sur cette langue.

Chapitre 2

Notions élémentaires de vietnamien

Ce chapitre a pour but de fournir au lecteur une connaissance des principes de base de la langue vietnamienne suffisante pour comprendre les difficultés particulières liées à l'exploitation informatique de cette langue, et ainsi les facteurs qui ont guidé les choix que nous avons effectués au cours de nos travaux. Nous présentons les principales caractéristiques du vietnamien du point de vue de l'écriture, de la phonétique, du vocabulaire ainsi que d'autres attributs grammaticaux importants d'une langue isolante.

- *Généralité : origine et typologie*
- *Écriture et phonétique*
- *Lexique*
- *Grammaire*
- *Bilan*

2.1. Généralités : origine et typologie

Nous présentons dans cette section les caractéristiques générales du vietnamien d'un point de vue taxonomique (famille, classification), et esquissons ses principales spécificités par rapport aux langues occidentales « classiques ». Celles-ci seront développées dans les sections suivantes, qui approfondissent la description fonctionnelle de la langue vietnamienne.

2.1.1. Origine de la langue vietnamienne

Le vietnamien appartient au groupe Viet-Muong, branche Mon-Khmer de la famille Austro-Asiatique. Cette classification de l'origine du vietnamien est dérivée de l'hypothèse d'André-Georges Haudricourt [HAU 53], qui est, à l'heure actuelle, acceptée par le monde de la recherche. Selon A.G. Haudricourt, le vietnamien ressemblait originellement aux langues non toniques du groupe Mon-Khmer. Le caractère tonique du vietnamien fut ultérieurement ajouté grâce aux échanges culturels du voisinage avec le thaï (*cf*. Haudricourt [HAU 54]). Le vietnamien avait déjà une riche littérature orale au moment de la conquête chinoise, vers le IIe siècle ap. J.C., puis, pendant une dizaine de siècles, sous la domination de la Chine, le chinois devint la langue administrative, et toutes les œuvres « savantes » furent écrites en chinois. Le vietnamien a, durant cette période, été enrichi par un nombre important de mots chinois prononcés « à la vietnamienne » et appelés des mots sino-vietnamiens. Au XII^{e} siècle apparaissent les caractères «nôm» (démotiques) basés sur l'écriture idéographique, qui permettent une transcription purement vietnamienne. Au XVII^{e} siècle, le missionnaire jésuite français Alexandre de Rhodes met au point une romanisation de l'écriture vietnamienne, dite « quốc ngữ », toujours en usage, après être devenue officielle au XIX^{e} siècle en Indochine française. Sous la colonisation française, le vietnamien a également évolué par emprunt de mots et de constructions grammaticales françaises.

2.1.2. Type de langue et classification du vietnamien

2.1.2.1. Présentation des types de langues recensés

On distingue 4 types de langues : flexionnelles, agglutinantes, analytiques (isolantes) et polysynthétiques.

Les ***langues flexionnelles*** sont des langues dans lesquelles les lemmes (« mots ») changent de forme selon leur rapport grammatical aux autres lemmes. On dit d'eux qu'ils subissent le jeu de la flexion et que l'ensemble des formes différentes d'un même mot fléchi forme son paradigme. Chaque forme d'un même paradigme peut transmettre un ou plusieurs types de traits grammaticaux (genre, nombre, fonction syntaxique, classe lexicale, temps, mode, *etc*.) pouvant s'opposer (singulier contre pluriel, masculin contre neutre, première personne du singulier contre première personne du pluriel, *etc*.). La flexion nominale est souvent nommée *déclinaison* tandis que celle du verbe est la *conjugaison*. Un représentant de ce type de langues est le latin.

Les ***langues agglutinantes*** sont des langues qui présentent la caractéristique structurelle de l'agglutination, c'est-à-dire, l'accumulation après le radical d'affixes distincts, pour exprimer les rapports grammaticaux. Les mots d'une langue agglutinante sont analysables en une suite de morphèmes nettement distincts. Un exemple représentatif est le turc.

Opposées aux langues agglutinantes et aux langues flexionnelles, les ***langues isolantes*** (ou ***analytiques***) expriment les divers rapports grammaticaux par des mots et des signes isolés. Les « mots » sont ou tendent à être invariables et il est, par conséquent, impossible de distinguer le radical des éléments grammaticaux. Les fonctions syntaxiques sont manifestées par l'ordre des mots dans la phrase. Le vietnamien et le chinois appartiennent à ce type de langues.

Dans les ***langues polysynthétiques***, des suffixes et des morphèmes grammaticaux se « synthétisent » sur un radical donné aboutissant à la formation de longs « mots-phrases ». Par exemple, l'inuit est une langue polysynthétique.

2.1.2.2. Le vietnamien

Le vietnamien est une langue typiquement isolante. Ses propriétés isolantes principales sont manifestées par les phénomènes suivants.

- Les mots sont morphologiquement invariables, c'est-à-dire qu'il n'y a ni conjugaison des verbes, ni accord des noms, adjectifs, *etc*.
- Le vietnamien a une unité qui s'appelle « tiếng ». Cette unité constitue à la fois une syllabe, un morphème et un mot simple. Par exemple, la phrase « Tôi đi đến trường » = « *Je vais à l'école* » est une phrase qui contient quatre mots, quatre syllabes et quatre morphèmes.
- Les sens grammaticaux[16] sont déterminés par les mots outils et l'ordre des mots. Par exemple, pour dire « *Je le bats* », on dit « Tôi đánh nó » =$_{litt}$ « *Je battre il* », alors que pour dire « *Il me bat* », on dit « Nó đánh tôi » =$_{litt}$ « *Il battre je* ». Ou encore, pour dire « *Je **vais** le battre* », on dit « Tôi **sẽ** đánh nó », « sẽ » étant un mot outil représentant le futur.

[16] c'est-à-dire le temps grammatical, le sens de la détermination définie/indéfinie, *etc*.

2.2. Écriture et phonétique

Le vietnamien possède 41 phonèmes, dont 23 consonnes, 13 voyelles simples, 3 diphtongues et 2 semi-voyelles, représentés par 29 lettres dans l'alphabet comme suit (en majuscule et minuscule):

Aa Ăă Ââ Bb Cc Dd Đđ Ee Êê Gg Hh Ii Kk Ll Mm Nn Oo Ôô Ơơ Pp Qq Rr Ss Tt Uu Ưư Vv Xx Yy

De plus, 5 accents sont utilisés pour représenter 6 tons, qu'on peut retrouver dans leur nom : ngang (sans accent), huyền (accent grave), hỏi (accent retombant), ngã (accent remontant), sắc (accent aigu), nặng (accent intensif).

Le vietnamien est une langue monosyllabique. Chaque syllabe a une structure phonétique rigoureuse, représentée dans le Tableau 2-1. Dans l'écriture, les syllabes sont séparées par des espaces.

Ton			
	Rime		
[premier son] Consonne (1)	[son intercalé] /w/ (2)	<son noyau> voyelle (3)	[dernier son] consonne/semi-voyelle (4)

Tableau 2-1 Composition phonétique d'une syllabe en vietnamien

Les numéros (1), (2), (3), (4) dans le Tableau 2-1 sont les numéros de positions que nous réutilisons dans les tableaux Tableau 2-2 et Tableau 2-3. Ces deux tableaux montrent la représentation des phonèmes par les lettres de l'alphabet.

Numéro	Phonème	Écriture	Positions possibles	Exemples
1	| f |	ph	1	**ph**ổi, **ph**áo
2	| t' |	th	1	**th**u, **th**ôi
3	| ʈ |	tr	1	**tr**ăng, **tr**ời
4	| z |	gi / d	1	**gi**ếng, **d**ao
5	| c |	ch	1, 4[17]	**ch**ơi, **ch**o, **ch**uộng, thí**ch**[1]

[17] D'après l'opinion d'un certain nombre de phonéticiens, le "ch" à la 4e position doit être transcrit en phonétique par | k |, et le "nh" à la 4e position par | ŋ |.

6	| ɲ |	nh	1, 4[1]	**nh**à, **nh**ảy, **nh**ững, mi**nh**[17]
7	| ŋ |	ng / ngh	1, 4	**ng**ười, **ngh**ĩ, **ngh**ề, **ng**a**ng**
8	| χ |	kh	1	**kh**uya, **kh**ông
9	| ɣ |	g / gh	1	**g**à, **g**ọi, **gh**i, **gh**e
10	| k |	c / q / k	1, 4	**c**à **k**ê, **c**á **q**uả, **c**á**c**
11	| t |	t	1, 4	**t**a, **t**ôi, **t**ức, **t**ố**t**
12	| ɾ |	r	1	**r**ổ, **r**á
13	| h |	h	1	**h**oa, **h**ọc **h**ành
14	| b |	b	1	**b**ằng, **b**ơi, **b**iết
15	| m |	m	1, 4	**m**ồ**m**, **m**ôi, **m**ắt, **m**ũi
16	| v |	v	1	**v**ui, **v**ắng, **v**ụt
17	| d |	đ	1	**đ**ang, **đ**ợi, **đ**ói
18	| n |	n	1, 4	**n**ăm, **n**àng, **n**ê**n**
19	| l |	l	1	**l**ên, **l**òng, **l**ợi
20	| s |	x	1	**x**e, **x**uống, **x**ua
21	| p |	p	4	bắ**p**, bị**p**, chắ**p**
22	| ʂ |	s	1	**s**ay **s**ưa, **s**ắp **s**ửa
23	| ʔ |	zero	1, 4	ăn uống, ỉ eo, ồn ào

Tableau 2-2 Liste des 23 phonèmes consonnes utilisés en vietnamien

Par ailleurs, le vietnamien importe les quatre lettres *f*, *j*, *w*, *z* pour écrire les mots étrangers empruntés par transcription phonétique (surtout les termes scientifiques). Par exemple `flo-rua = fluorure`, `juđô = judo`.

La structure phonique stricte des syllabes vietnamiennes semble être la raison de l'absence d'annotation phonétique dans tous les dictionnaires vietnamiens existants. Cette propriété remarquable résulte naturellement de l'origine artificielle du système d'écriture développé par A. de Rhodes, qui est une forme de transcription phonétique spécialisée. Une transcription automatique des syllabes vietnamiennes en alphabet phonétique international pourrait aisément être réalisée pour enrichir les bases lexicales pour le TAL.

Numéro	Phonème	Écriture	Positions possibles	Exemples
1	| i |	i / y	3	**i**m **ỉ**m, **ý** ch**í**
2	| e |	ê	3	**ê** ch**ề**, **ê**m đ**ề**m
3	| ɛ |	e	3	**e** d**è**, **e** th**ẹ**n
4	| ɛ̆ | [18]	a {nh, ch}	3	**a**nh **á**ch, x**a**nh x**a**nh
5	| a |	a	3	**a** h**a**, l**a** đ**à**, t**a**i
6	| ă |	ă / a{u, y}	3	**ă**n n**ă**n, **ă**n ch**ặ**n, r**au** đ**ay**
7	| ɤ |	ơ	3	b**ơ** ph**ờ**, t**ờ** m**ờ**
8	| ɤ̆ |	â	3	**â**n c**ầ**n, l**ấ**n b**ấ**n
9	| ɯ |	ư	3	t**ự** t**ử**, th**ư** t**ừ**
10	| o |	ô	3	**ô** h**ô**, h**ồ** đ**ồ**
11	| ɔ |	o / oo	3	c**o** r**o**, l**ò** d**ò**, x**oo**ng
12	| ɔ̆ |	o	3	v**õ**ng l**ọ**ng, t**ó**c, h**ọ**c
13	| u |	u	3	t**u** h**ú**, l**ù** m**ù**
14	| i◡e |	ia / ya / iê / yê	3	k**ia** k**ìa**, khu**ya**, **yê**u chi**ề**u
15	| u◡o |	uô / ua	3	t**uố**t t**uồ**n t**uộ**t, t**ua** r**ua**
16	| ɯ◡ɤ |	ươ / ưa	3	l**ướ**t th**ướ**t, l**ưa** th**ưa**
17	| i̯ |	i / y	4	ta**i** t**á**i, cà**y** c**ấ**y
18	| u̯ |	o / u	2, 4	t**oá**n, đà**o** hà**o**, t**uầ**n, đ**au**

Tableau 2-3 Liste des 13 voyelles simples, 3 diphtongues et 2 semi-voyelles utilisées en vietnamien

[18] Ce phonème est proposé par les phonéticiens qui n'acceptent pas les consonnes | ɲ | et | c | à la fin des syllables.

2.3. Lexique

Les unités de base du vietnamien sont des syllabes séparées, que nous définissons ici dans un premier temps plus précisément (section 2.3.1). Les mots à proprement parler peuvent être constitués d'une unique syllabe, ou construits par une composition plus ou moins complexe (section 2.3.2). Échappent en partie à cette règle les mots empruntés à des langues étrangères (section 2.3.3).

2.3.1. Unité de base : la syllabe (« tiếng »)

Comme nous l'avons évoqué à la section 2.1.2, la langue vietnamienne possède une unité spéciale appelée « tiếng » qui correspond à la fois à une syllabe du point de vue phonologique, à un morphème du point de vue syntaxique, à un sémantème du point de vue de la structure du mot, et à un mot du point de vue des constituants de la phrase. Il existe trois types de « tiếng » :

1. « tiếng » ayant un sens réel comme sông = *rivière*, núi = *montagne*, đi = *aller*, đứng = *tenir debout*, nhớ = *se souvenir*, thương = *aimer, compatir* … Il peut constituer à lui seul un constituant de phrase complet du point de vue syntaxique et sémantique. Ce type de mot est appelé **mot lexical**.
2. « tiếng » comme nhưng = *mais*, mà = *que*, tuy = *bien que*, nên = *alors* …, qui ne peut pas être un constituant de phrase à lui seul, mais qui est utilisé pour composer un constituant de phrase lexical, est appelé **mot vide**.
3. « tiếng » employable uniquement pour la création de mots composés. Il peut s'agir de mots d'origine chinoise comme sơn = *montagne*, thuỷ = *eau*, gia = *maison*, bất = [*négation*]… ou de mots dont le sens n'est pas défini indépendamment comme cộ dans xe cộ = *véhicule*, đẽ dans đẹp đẽ = *beau*, vẻ dans vui vẻ = *joyeux*…

2.3.2. Unités lexicales

Parmi les diverses définitions du mot en vietnamien, les linguistes sont parvenus à un accord et considèrent comme un mot la plus petite unité ayant un sens spécifié et une structure stable, et utilisée pour composer des constituants de phrase. Une grande partie des mots vietnamiens sont des mots simples repérés exactement comme des morphèmes et des syllabes morphologiquement invariables. Mais il existe également un nombre important de mots composés dont nous présentons ici la structure.

La structure d'un mot est décrite en se basant sur le nombre de syllabes (pour distinguer des mots mono- ou multi-syllabiques), et la manière dont sont composés les mots complexes (pour déterminer leurs éléments et la relation entre ces éléments).

Le dictionnaire vietnamien contient donc :

1. Des **mots simples** ou mots monosyllabiques correspondant aux catégories 1 et 2 de « tiếng ». Nous considérons également comme mots simples quelques mots multi-syllabiques soit empruntés à des langues étrangères, soit composés de syllabes dont le sens n'est pas reconnaissable.

   Exemples :

   nhà = *maison*, cửa = *porte*, bàn = *table*, ghế/ *chaise*, cười = *rire*, nói = *parler*, hát = *chanter*, đi = *aller*…

   cà rốt = *carotte*, cà phê = *café*, …

bồ câu = *pigeon*, bù nhìn = *épouvantail*, ễnh ương = *kaloula* (espèce de crapaud), …

Il convient d'apporter une attention particulière aux « tiếng » sino-vietnamiens, parce que certains d'entre eux ont été tout à fait vietnamisés (par exemple : tuổi = *âge*, buồm = *voile*, đầu = *tête*), alors que d'autres ne peuvent être qu'élément d'un mot composé (par exemple : nhân = *personne, humain* ne peut se situer que dans des mots composés comme nhân tài = *homme de talent*, nhân lực = *main-d'œuvre*, nhân loại = *genre humain* …).

Il existe également des syllabes d'origine vietnamienne qui étaient autrefois des mots simples mais ne peuvent en vietnamien moderne être employées seules (par exemple trốc = *tête*, tác = *âge*…).

2. Des **mots complexes** qui ont plus d'une syllabe. Ce sont des mots redoublés et des mots composés sémantiquement.

   a. *Mots redoublés* : Les mots redoublés sont construits par combinaison phonique, généralement de deux syllabes. On entend par combinaison phonique, un phénomène de répétition et de symétrie. La répétition peut être totale ou (plus fréquemment) partielle, et on distingue ainsi des mots répétés totalement ou partiellement.

      Les mots redoublés totalement sont généralement des mots décrivant le son, par exemple : oe oe dans la phrase Em bé khóc oe oe =$_{litt}$ *enfant pleurer [le son]* = *Le nouveau-né vagit*. On trouve aussi dans cette catégorie des noms d'animaux ou de plantes, par exemple cu cu = *coucou*, chôm chôm = *ramboutan*.

      Les mots redoublés partiellement sont créés par la répétition partielle des syllabes suivant certaines règles de combinaison phonique :

      - Quand le premier son (*cf.* 2.2) est redoublé, il est possible de constater une combinaison phonique modifiant par un mécanisme dit de « symétrie » une partie de la syllabe, qui peut être :
        - le son noyau : nhúc nhích = *pincer légèrement les lèvres* ;
        - le dernier son : chằm chặp = *fixement* ;
        - le son intercalé (répétés) : xuềnh xoàng = *simplement (en parlant d'une tenue vestimentaire)* ;
        - le ton : may mắn = *chanceux* ;
      - Il est également possible de redoubler la rime et le ton, alors que les premiers sons sont symétriques dans le système phonique du vietnamien : lí nhí = *très petit*, càu nhàu = *grommeler*.

      Parmi les mots de cette catégorie, nous pouvons distinguer une classe de mots dont aucune des deux syllabes n'a de sens (par exemple : trục trặc = *détraqué, en panne*), et une classe de mots dont une syllabe a un sens qui se trouve spécialisé par la répétition (par exemple : nhỏ = *petit* => nhỏ nhắn / nhỏ nhoi = *menu, mignon / minime, modique*).

   b. *Mots composés* : Ces mots sont composés à partir des syllabes en suivant une combinaison sémantique. On distingue deux classes de mots composés, en fonction de la composition sémantique des syllabes composantes.

La première classe contient les mots dits ***composés parallèles*** (ou *composés copulatifs*), dont la composition représente une coordination sémantique des syllabes. Chaque composant a son propre sens et joue un rôle égal, et l'ordre des mots est donc généralement variable (à l'exception de quelques mots composés dont une ou toutes les syllabes sont d'origine chinoise, pour lesquels l'ordre des syllabes est fixe). Par exemple :

quần = *pantalon*, áo = *chemise* => quần áo / áo quần = *vêtement*

giang = *rivière*, sơn = *montagne* [19] => giang sơn = *pays natal*

Pour les mots composés de deux syllabes, il existe habituellement une symétrie entre ces syllabes et leur sens devient symbolique. Par exemple :

vuông = *carré*, tròn = *rond* => vuông tròn = *à souhait*

cười = *rire*, cợt = *taquiner* => cười cợt = *rigoler, badiner*.

La deuxième classe constitue des mots dits ***composés majeur/mineur***, dont la composition manifeste une subordination sémantique de normalement deux composants. Un composant jouant le rôle de majeur porte un sens général, et l'autre composant jouant le rôle de mineur limite le domaine du majeur. Dans ce cas, l'ordre des éléments est important, habituellement majeur-mineur. Par exemple :

nhà = *maison*, khách = *invité*, tù = *prisonnier* => nhà khách = *maison de réception* ; nhà tù = *prison*.

làm = *faire*, việc = *travail*, giàu = *riche* => làm việc = *travailler* ; làm giàu = *s'enrichir*.

nhanh = *rapide*, tay = *main*, ý = *idée, pensée, esprit* => nhanh tay = *se dépêcher* ; nhanh ý = *avoir la présence d'esprit*.

Pour les composés dont les composants sont d'origine chinoise, l'ordre devient mineur-majeur. Par exemple :

quốc = *national*, kì = *drapeau*, ca = *chanson*[20] => quốc kì = *drapeau national*, quốc ca = *hymne national*.

Le mot jouant le rôle majeur peut porter un sens généralisé ou symbolique. Par exemple :

chân = *pied*, trời = *ciel*, vịt = *canard* => chân trời = *horizon*, chân vịt = *hélice (de bateau)*.

Pour les adjectifs, le mot jouant le rôle mineur indique l'intensité de la qualification. Par exemple :

đen = *noir* => đen sì = *trop noir*; đen ngòm = *très sombre, très obscur*.

3. Enfin, **des expressions figées et des locutions**, qui sont généralement aussi considérées comme des unités lexicales.

La Figure 2-1 synthétise les différentes formes de mots du vietnamien.

[19] tous ces deux composants sont des mots d'origine chinoise, utilisés uniquement dans les mots composés

[20] Il faut noter que pour plusieurs composés sino-vietnamiens, la plupart des vietnamiens ne reconnaissent plus le sens d'origine des composants.

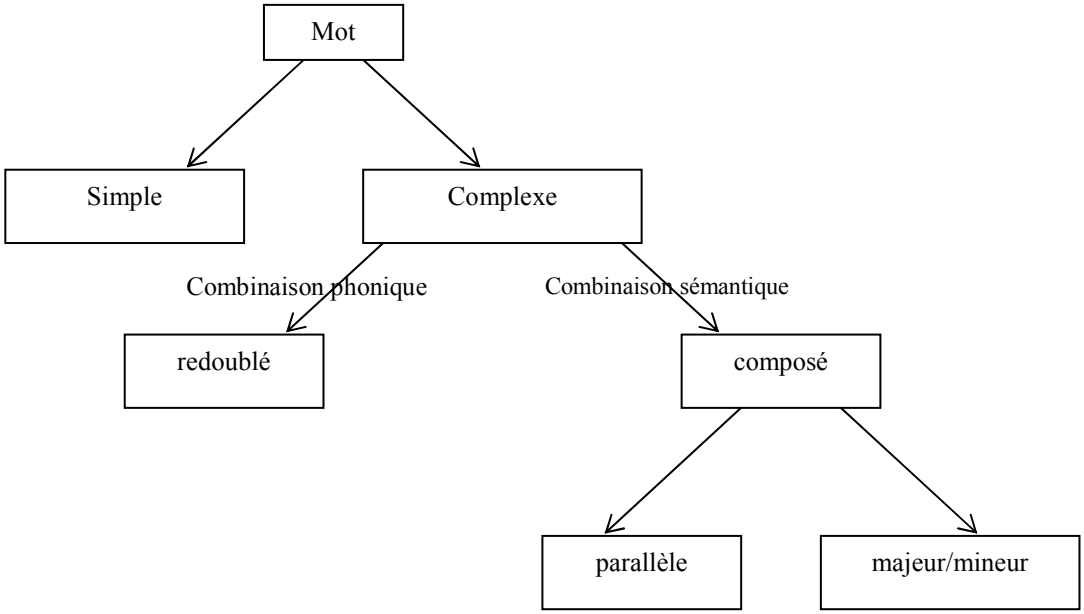


Figure 2-1 Formes des mots en vietnamien

Cette complexité des formes des mots vietnamiens constitue une difficulté majeure pour la tâche de segmentation automatique des textes en unités lexicales. Ce sujet est concrètement discuté à la section 3.4.3.

Avant de nous concentrer sur le sujet de la grammaire vietnamienne, nous émettons encore quelques remarques importantes concernant l'emprunt par le vietnamien de mots étrangers.

2.3.3. Mots empruntés

Dans l'histoire de son développement, le vietnamien a été principalement en contact avec le chinois (mille ans de domination chinoise et de multiples guerres ensuite), puis avec le français (une centaine d'années de colonisation). Cela explique l'emprunt très important de mots chinois et de plusieurs mots français. Le vietnamien a également emprunté des mots d'autres langues, soit directement (mais de manière non systématique), soit indirectement par l'intermédiaire du chinois. Nous discutons seulement ici des mots empruntés du chinois et du français.

2.3.3.1. Mots empruntés au chinois

Une grande partie du vocabulaire vietnamien est composée de mots empruntés au chinois, avec une prononciation vietnamisée. À côté des mots complexes stabilisés que nous présentons dans la section précédente, nous attirons l'attention sur les syllabes sino-vietnamiennes qui peuvent ou non être utilisées indépendamment, mais qui, quand elles participent à la formation de nouveaux mots ou termes, suivent l'ordre de combinaison employé en chinois, qui ne coïncide pas toujours avec celui du vietnamien (*cf*. l'ordre des syllabes dans les mots complexes à la section 2.3.2). Par exemple :

tin = *information*, họ c = *science* (seulement dans les mots composés[21]), hoá = *devenir* => tin họ c = *informatique*, tin họ c hoá = *informatiser*.

[21] On retrouve ce sens dans les autres mots désignant des sciences et disciplines : toán họ c = *mathématiques,* văn họ c= *littérature*, *etc*.

2.3.3.2. Mots empruntés au français et aux langues occidentales

Les mots empruntés au français sont classés en trois catégories :

- Transcription phonétique directe, par exemple : gác đờ bu = *garde-boue*, xà phòng = *savon*, pênixilin = *pénicilline*.
- Transcription phonétique réduite, par exemple : phanh = *frein*, len = *laine*, tôn = *tôle*.
- Traduction : máy = *machine*, kéo = *trainer* => máy kéo = *tracteur*.

Un problème à noter concerne les transcriptions phonétiques directes. Parmi les mots appartenant à cette classe, nous distinguons les mots de la vie quotidienne comme *savon*, *garde-boue*, *etc*. dont l'écriture est stabilisée avec une espace entre les syllabes, et les mots plutôt scientifiques ainsi que les noms propres occidentaux dont l'écriture est encore variable : leurs syllabes peuvent être attachées (pênixilin), séparées par des espaces (pê ni xi lin), ou séparées par des tirets (pê-ni-xi-lin). La dernière forme est encouragée mais cette recommandation n'est pas toujours respectée en réalité. De plus, il n'existe pas réellement de consensus concernant ces transcriptions, et la tendance moderne tend à maintenir ces mots empruntés sous leur forme originale ou anglaise. En bref, l'écriture non standardisée des mots phonétiquement transcrits reste actuellement un problème pour l'analyse automatique des textes.

2.4. Grammaire

Dans cette section, nous discutons des catégories grammaticales (section 2.4.1) et de la syntaxe (section 2.4.2) de la langue vietnamienne.

2.4.1. Classification des mots

Le vietnamien est une langue isolante (*cf*. section 2.1.2), dans laquelle chaque mot a une forme unique qui ne peut pas être modifiée par dérivation ou flexion. Les relations grammaticales ne se manifestent pas par la flexion mais par l'ordre des mots. La classification des parties du discours n'est donc pas morphologiquement évidente.

Il n'existe pas à l'heure actuelle de standard pour la classification des mots en vietnamien. Nous présentons dans cette section une classification proposée par le Comité des Sciences Sociales du Vietnam [UYB 83], sur laquelle s'appuie le jeu d'étiquettes pour l'étiquetage morphosyntaxique du vietnamien que nous présentons au chapitre suivant (*cf*. section 3.3.2).

Une classification classique et très générale des mots consiste à les séparer en deux classes. La première classe se compose des **mots autonomes** qui portent un sens réel, c'est-à-dire auxquels est associé un objet ou un phénomène. Autrement dit, ce sont des mots ayant un sens lexical. Par contre, la deuxième classe rassemble les **mots outils** qui jouent seulement un rôle grammatical (par exemple : rất = *très*, với = *avec*, thì = *alors*, mà = *que / dont / où*, *etc*.).

Nous nous intéressons ici à une catégorisation plus détaillée qui peut nous aider à analyser la composition des phrases. Cette classification contient les 8 parties du discours qui sont généralement universelles et obtient donc un haut consensus des linguistes (Tableau 2-4).

Partie du discours	Description	Exemples
Nom	désigne des entités	nhà = *maison*, xe đạp = *vélo*
Pronom	remplace un nom	tôi = *je*, họ = *ils/elles*, vậy = *ça, cela*
Verbe	désigne un procès, un état ou un devenir	là = *être*, đi = *aller*, uống = *boire*
Adjectif	désigne une propriété	đỏ = *rouge*, cẩu thả = *sans soin*
Mot complément	désigne un sens grammatical concernant le temps, le degré, la négation, *etc*.	đã = temps passé, sẽ = temps futur, rất = *très*
Conjonction	désigne une relation entre des constituants de la phrase	của = *de*, thì = *alors*, và = *et*
Mot introductif (mot modal)	ajouté pour exprimer l'opinion du locuteur, attaché à la structure de la phrase	nhỉ = *n'est-ce pas* / *comme* (comme c'est bon)
Mot émotif (interjection)	ajouté pour exprimer l'opinion du locuteur, indépendant de la structure de la phrase	ái chà = *fichtre* ! / *Ouais* !, vâng / dạ = *oui*

Tableau 2-4 Les parties de discours du vietnamien

Nous présentons dans ce qui suit, pour chacune des catégories du discours énumérées, des catégories plus fines également proposées dans [UYB 83] permettant de caractériser plus précisément les mots. Cette section se conclut ensuite sur une brève description du phénomène de mutation grammaticale, assez courant en vietnamien.

2.4.1.1. Noms

Les noms peuvent être répartis dans des sous-catégories suivantes :

1. **Noms comptables** : Ce sont des noms désignant des choses qui peuvent exister sous forme individuelle, par exemple : áo = *chemise*, xe = *véhicule*, người = *personne*, *etc*. Au pluriel, ces noms doivent être accompagnés par un classificateur (*cf*. nom individuel).

   Exemple : nhà = *maison* => *deux maisons* = 2 **ngôi** nhà

2. **Noms collectifs** : Les noms collectifs désignent des choses qui n'existent pas sous forme d'individu mais uniquement en tant qu'ensemble. Dans les groupes nominaux, ils ne peuvent être précédés directement que par un quantificateur universel ou un nombre « géant ».

   Exemple : giấy tờ = *papiers*, tất cả = *tout* => Tất cả giấy tờ = *tous les papiers*, hàng nghìn giấy tờ = *un millier de papiers*

3. **Noms individuels** (ou **classificateurs**) : Les classificateurs jouent le rôle d'affixe des noms, dans les syntagmes ils se trouvent avant des noms comptables[22] (et parfois des noms abstraits).

   Exemple : cái = [*affixe des noms d'objet*], nhà = *maison*, bút = *stylo* => một cái nhà = ***une*** *maison*, hai cái bút = ***deux*** *stylos*

4. **Noms concrets (**ou **noms d'unité de mesure**) : Ce sont des noms qui désignent une unité déterminant une quantité. On distingue quatre différents types d'unité :

   a. unité conventionnelle exacte, par exemple : Lít = *litre* => Một lít nước = *un litre d'eau*

   b. unité conventionnelle inexacte, par exemple : nắm = *poignée* => Một nắm muối = *une poignée de sel*

   c. unité de temps, par exemple : giờ = *heure* => Một giờ nghỉ = Une heure de pause

   d. unité d'organisation, par exemple : tổ = *groupe* => Một tổ công nhân = un groupe d'ouvriers

5. **Noms abstraits** : Ce sont des noms qui désignent un concept abstrait. Au pluriel, ces noms n'ont pas besoin d'être précédés par un classificateur. Par exemple,

   ý nghĩ = *pensée* => Những ý nghĩ = *des pensées*

   thái độ = *comportement* => Hai thái độ = *deux comportements*

6. **Noms de quantité (**ou **numéraux)** : Ce sont des noms qui désignent une quantité, par exemple : hai = *deux*, vài = *quelques*, phần lớn = *la plupart*, những = <*marque de pluriel*>. La proposition [UYB 83] prévoit également que ces mots puissent être considérés non pas comme des noms, mais comme constituant une classe à part entière.

7. **Noms locatifs** : Ce sont des noms qui désignent une position relative. Ces mots sont également utilisés comme des prépositions. Par exemple : trên = *sur / supérieur (n.m)*.

[22] Chaque nom comptable peut avoir plusieurs classificateurs

Par ailleurs, comme dans toutes les langues, les noms peuvent être également sous-catégorisés en noms communs et noms propres. Comme en français, les noms propres se distinguent à l'écrit par une initiale majuscule.

2.4.1.2. Verbes

Les verbes sont classés comme suit :

1. **Verbes transitifs** : Ce sont des verbes qui font passer directement l'action d'un sujet sur un objet, par exemple : ăn = *manger*, viết = *écrire*, cải tiến = *améliorer*
2. **Verbes intransitifs** : Ce sont les verbes qui n'ont pas de complément d'objet, par exemple ngủ = *dormir*, làm việc = *travailler*, nghỉ ngơi = *se reposer*.
3. **Verbes d'impression** : Ce sont des verbes qui expriment un état ou processus psychologique, par exemple : hiểu = *comprendre*, yêu = *aimer*. Ces verbes ont aussi le caractère « transitif » comme les verbes transitifs ci-dessus, la différence est qu'ils peuvent s'associer à un adverbe de degré, par exemple : rất = *très* => rất yêu =$_{litt}$ *très aimer* = *aimer beaucoup*.
4. **Verbes d'orientation** : Ce sont des verbes de mouvement qui incluent une certaine direction. Ces verbes peuvent suivre d'autres verbes pour indiquer la « direction » de ces derniers. Par exemple : vào = *entrer* => Anh ta vào = *Il entre* ; Anh ta chạy vào =$_{litt}$ *Il court à l'intérieur* = Il entre en courant.
5. **Verbes d'état** : Ce sont des verbes qui indiquent l'existence des objets, et qui peuvent être utilisés sous forme impersonnelle. Par exemple : có = *avoir*, còn = *rester*, *avoir encore* => Tôi còn gạo = *j'ai encore du riz* ; Còn gạo = *il reste du riz*.
6. **Verbes de transformation** : Ce sont des verbes correspondant à une transformation comme nên, trở nên = *devenir*. Ces verbes sont suivis par un complément indiquant le résultat de transformation.
7. **Verbes volitifs** : Ce sont des verbes qui désignent une volonté, comme muốn = *vouloir*, dám = *oser*. Ces verbes sont généralement suivis par d'autres verbes.
8. **Verbes de réception (passifs [23])** : Ce sont des verbes qui désignent un état de réception. Contrairement aux verbes volitifs, le sujet n'exerce pas l'action. Ces verbes doivent être suivis par un verbe ou un nom. Par exemple : bị, phải, được = état passif (dans un sens négatif ou positif) => bị phạt = *être puni* (phạt = *punir*); được khen = *être félicité* (khen = *féliciter*) ; được giải thưởng = *gagner le prix* ; phải gió = *être frappé d'un courant d'air* (gió = *vent*).
9. **Verbes comparatifs** : Ce sont les verbes qui expriment une comparaison entre le sujet et un objet, par exemple : bằng = *égaler*, hơn = *être supérieur*.
10. **Verbe « là »** : C'est la copule qui peut se traduire en français par « être ».

2.4.1.3. Adjectifs

Les adjectifs sont classés comme suit :

1. **Adjectifs qualitatifs** : Ce sont des adjectifs qui désignent une qualité, et qui peuvent être précédés par un adverbe de degré comme rất = *très* (comme les verbes d'impression). Ces adjectifs peuvent être suivis par un nom ou un groupe verbal qui limite le « domaine » de la qualité exprimée. Par exemple : tốt, giỏi = *bon* => giỏi toán = *bon en mathématiques*.

[23] Noter que la passivation n'est pas habituelle en vietnamien, *cf.* section 2.4.2.2.

2. **Adjectifs quantitatifs** : Ce sont des adjectifs qui désignent un propriété quantifiable, par exemple cao = *haut*. Ces adjectifs peuvent être utilisés avec soit un adverbe de degré, soit un complément de quantité et/ou un repère concernant l'évaluation de quantité, par exemple : cao 2 mét = *de 2 mètres de hauteur*.

2.4.1.4. Pronoms

Les pronoms peuvent se classer dans les catégories suivantes.

1. **Pronoms personnels** : Les pronoms personnels se distinguent non seulement par la personne et le nombre, mais aussi par la relation sociale ou familiale entre le locuteur et la personne désignée, ainsi que le sentiment. Par exemple, tous ces mots « tôi, tao, ta, cháu, con, em, anh, chị, cô, bác, … » peuvent être utilisés pour dire « *je* ». En outre, plusieurs de ces mots peuvent être également utilisés pour dire « *tu/vous* ».
2. **Pronoms déictiques et démonstratifs** : Ce sont les pronoms utilisés pour indiquer le temps ou l'espace, par exemple đây = *ici*, này = *ce*, bây giờ = *maintenant*.
3. **Pronoms de quantité** : Comme leur nom l'indique, ces pronoms sont utilisés pour indiquer la quantité, par exemple : bấy nhiêu = *tant, autant*.
4. **Pronoms de qualité** : Ce sont les pronoms utilisés pour référencer une action ou une qualité, par exemple : thế, vậy = *comme ceci, comme cela*
5. **Pronoms interrogatifs** : Ce sont les pronoms utilisés pour l'interrogation, par exemple : ai = *qui* ; gì = *quoi, que* ; đâu = *où* ; bao giờ = *quand* ; bao nhiêu, mấy = *combien* ; sao, thế nào = *comment.*

2.4.1.5. Mots compléments

Les mots compléments (MC) sont catégorisés comme suit :

1. **MC de temps** : Ce sont des mots qui expriment le sens grammatical de temps, par exemple đã = *déjà* / [*temps passé*], sẽ = [*temps futur*]. Ces mots-outils jouent le rôle dévolu dans les langues flexionnelles comme le français à la conjugaison des verbes aux divers temps grammaticaux.
2. **MC de degré** : Ces mots peuvent se trouver éventuellement avant ou après les adjectifs ou les verbes d'impression pour exprimer un sens grammatical de degré, par exemple : rất / lắm = *très / beaucoup* (rất đẹp / đẹp lắm = *très belle*) ; hoàn toàn = *absolument*, cực = *extrêmement*.
3. **MC de rapport** : Ces compléments expriment un rapport de constance d'une action ou d'un état dans un contexte temporaire, par exemple cũng = *aussi, également* ; vẫn, còn = *encore* ; liên tục = *continuellement*
4. **MC de négation, d'affirmation** : Ces compléments sont utilisés pour la négation ou l'affirmation d'une préposition, par exemple : không = *négation*, có = *affirmation soulignée*.
5. **MC impératifs** : Ce sont des mots compléments utilisés pour exprimer l'impératif, par exemple : đừng = *suggestion négative*, hãy = *suggestion positive*, phải = *devoir*, …

2.4.1.6. Conjonctions

Les conjonctions sont classées dans deux catégories suivantes :

1. **Conjonctions de subordination** : Ces conjonctions expriment un rapport de subordination entre deux constituants de phrase, par exemple : do = *due à*, của = *de (possession)*, để = *pour*, từ = *de (origine)*, bằng = *par*.

2. **Conjonctions de coordination** : Ces conjonctions expriment un rapport de coordination entre deux constituants de phrase, par exemple : và = *et* ; cùng = *ensemble* ; nếu … thì = *si … alors*.

2.4.1.7. Mots modaux et interjections

Ces catégories consistent en des particules qui ajoutent une nuance d'émotivité ou de modalité à une phrase. La différence entre eux est leur position dans la phrase. Les mots modaux s'attachent à la composition des phrases, alors que les interjections sont plus indépendantes (*cf.* description et exemples dans le Tableau 2-4).

2.4.1.8. Bilan

Comme nous l'avons évoqué en introduction à cette section, le problème de classification grammaticale des mots vietnamiens est encore en débat dans la communauté linguistique vietnamienne, tant du point de vue théorique que pratique.

Au niveau théorique, le problème est celui de la caractérisation plus fine des parties du discours. Notons que la classification présentée ci-dessus est encore assez « grossière », sorte de compromis permettant d'avoir une base de travail raisonnable en évitant les discussions trop profondes. Cette classification est donc loin de bien refléter la distribution syntaxique des mots. Pour les études plus détaillées, les différents auteurs proposent différentes classifications même s'ils ont la même approche[24], car une propriété lexicale peut-être mise en avant pour un auteur mais pas pour un autre. Cela est dû au fait que la classification n'est pas guidée par des signes morphologiques, mais notamment par des signes sémantiques et le jeu des cooccurrences.

D'un point de vue technique, l'association d'une catégorie précise à chaque occurrence de mot n'est pas toujours évidente, à cause du phénomène de mutation grammaticale des mots, très fréquent en vietnamien. Voici quelques cas typiques :

Mots autonomes et outils : Plusieurs mutations sont reconnues, comme du nom en conjonction, du verbe en conjonction. Par exemple :

> Anh **cho** em cuốn sách này =$_{litt}$ *Je donner tu [classificateur] livre ce = Je te* ***donne*** *ce livre*. (cho = *donner*)
>
> Anh gửi cuốn sách này **cho** em =$_{litt}$ *Je envoyer [classificateur] livre ce* ***à*** *tu = Je t'envoie ce livre*. (cho = conjonction « *à* »)

Verbes et noms : On peut convertir systématiquement les verbes en noms en les faisant précéder de classificateurs (*cf.* 2.4.1.1) correspondant au français « le fait de ». Par ailleurs, les verbes d'impression sont souvent mutables en nom sans ajouter de classificateur. Par exemple :

> Suy nghĩ = *réfléchir* => những suy nghĩ = *les réflexions*

Adjectifs et noms : À l'instar des verbes d'impression, certains adjectifs sont directement mutables en noms. Comme les verbes, les adjectifs peuvent souvent systématiquement être convertis en noms en les faisant précéder d'un classificateur. Par exemple :

> mệt mỏi = *fatigué* => **sự** mệt mỏi = *la fatigue*

Un ensemble de descriptions lexicales plus détaillées, prenant en compte plusieurs points de vue, est donc nécessaire pour la tâche d'analyse morphosyntaxique du vietnamien. Nous discutons de ce sujet au Chapitre 3.

---

[24] Ils se basent notamment tous sur 3 critères : sémantique catégorielle, cooccurrence dans un même syntagme, et fonction syntaxique

2.4.2. Syntaxe

Dans cette section, nous réalisons une brève introduction des structures syntagmatiques et syntaxiques du vietnamien.

2.4.2.1. Procédés syntaxiques

Comme nous l'avons vu à la section 2.1.2, les sens grammaticaux se manifestent par l'ordre des mots, les mots outils, le redoublement des mots ainsi que, pour l'oral, l'intonation du locuteur.

L'ordre des mots permet de distinguer les différents rapports entre les constituants des phrases. En vietnamien, les constituants se mettent toujours dans l'ordre tête - complément et thème - rhème. Par exemple :

- Nom - modificateur : nhà = *maison*, gạch = *brique*, đẹp = *beau*, nghỉ = *se reposer*
    - nhà gạch = *maison en brique*
    - nhà đẹp = *belle maison*
    - nhà nghỉ = *maison de repos*
- Verbe - complément d'objet : đọc = *lire*, sách = *livre*
    - đọc sách = *lire un livre*
- Verbe - adverbe de manière : đi = *aller*, nhanh = *vite*
    - đi nhanh = *aller vite*
- Thème - rhème : gió = *vent*, thổi = *souffler*
    - gió thổi = *le vent souffle*.

Les mots outils sont utilisés pour exprimer le pluriel dans les groupes nominaux, le temps dans les groupes verbaux, la conjonction dans les structures de coordination ou de subordination. Voici quelques exemples.

- Le pluriel : những = *pluriel indéfini*, các = *pluriel défini*, gà = *poulet*, con = *classificateur (pour les animaux)*
    - những con gà = *des poulets*
    - các con gà = *les poulets*
- Le temps : đã = temps passé, anh ấy = *il*, về = *rentrer*
    - Anh ấy về = *Il rentre*
    - Anh ấy đã về = *Il est rentré*.
- Conjonction de coordination et de subordination : gà = *poulet*, mẹ = *mère*, và = *et*, của = *de (possession)*
    - gà mẹ = *la mère poule*
    - gà và mẹ = *le poulet et la mère*
    - gà của mẹ = *le poulet de la mère*

La forme redoublée des mots permet notamment de modifier l'intensité des adjectifs. Le redoublement peut être combiné avec des mots outils pour souligner un constituant. Par exemple :

- Redoublement des mots d'une syllabe : vàng = *jaune*
    - vàng vàng = *jaunâtre*
- Redoublement des mots de deux syllabes : lúng túng = *perdre contenance*

- lúng ta lúng túng = *perdre contenance (sens plus fort)*

- Redoublement des mots en ajoutant des mots outils : đẹp = *beau*, là = *mot introductif*
  - đẹp đẹp là = *très beau (exclamation)*

En outre, dans les dialogues, on remarque également le redoublement d'un mot quelconque en ajoutant des mots qui ont une relation phonique ou sémantique avec le mot redoublé. Considérons l'exemple suivant comme l'illustration :

- phòng est synonyme de ngừa = *prévenir*, xà phòng[25] = *savon* => forme redoublée : xà phòng xà ngừa.

L'intonation du locuteur peut changer le sens de la phrase, par exemple transformer une affirmation en négation.

2.4.2.2. Structure « thème - rhème »

Du point de vue de la grammaire fonctionnelle, le vietnamien appartient aux langues avec préférence du thème (*topic prominent languages*, *cf.* Li et Thompson [LIT 76]). Cette propriété se manifeste en vietnamien par les phénomènes suivants.

- Le sujet ne peut pas être identifié par la morphologie (il n'y a pas de variation morphologique en vietnamien), ni par sa position dans la phrase, alors que le thème, qui est un groupe nominal quelconque et qui peut n'avoir aucun lien syntaxique avec le prédicat de la phrase, est toujours à la position initiale dans la structure phrastique. La Figure 2-2 illustre cette structure « thème - rhème » du vietnamien.

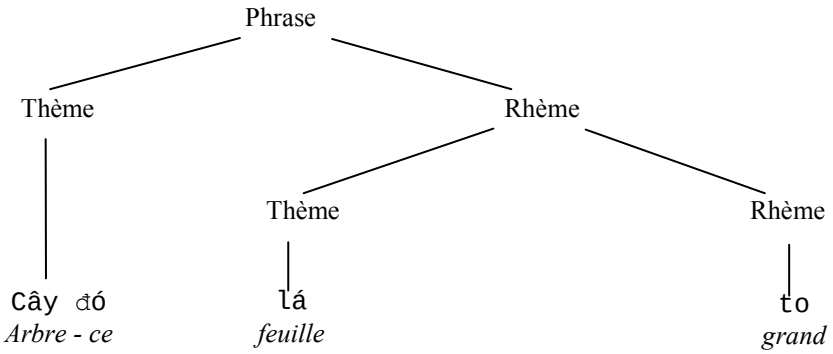


Figure 2-2 Structure « thème - rhème » de la phrase « Cet arbre, les feuilles sont grandes »

- La passivation n'est pas une construction naturelle, car c'est le thème et non pas le sujet qui joue le rôle plus important dans la construction de la phrase.
- Il n'y a pas de sujet impersonnel en vietnamien, car le sujet n'est pas obligatoirement présent. Par exemple, pour dire « *il fait très froid ici* », on dit tout simplement « ở đây rất lạnh » = « *ici très froid* ».
- La construction des phrases à double sujet est familière en vietnamien. L'exemple de la Figure 2-2 est un cas très courant.
- C'est le thème mais pas le sujet qui contrôle la co-référentialité dans la phrase. Considérons l'exemple suivant :
  - Cây đó lá to nên tôi không thích =$_{litt}$ *Arbre ce feuille grand donc je non aimer = Cet arbre, les feuilles sont grandes, donc je ne* ***l****'aime pas.*

[25] mot français d'écriture vietnamisée.

Dans cet exemple, le constituant supprimé de la fin de la phrase ne fait pas référence au sujet « *feuille* » mais au thème « *arbre* ».

2.4.2.3. Grammaire formelle

Dans l'état actuel des recherches au Vietnam, la notion de grammaire formelle est encore restreinte à la communauté Informatique pour les langages de programmation. La problématique de formalisation de la grammaire vietnamienne est discutée au Chapitre 4.

2.5. Bilan

Ce chapitre nous a permis de présenter les bases de la langue vietnamienne : origine, type des langues, composition (graphique et sémantique) des mots, catégorisation grammaticale, structure syntaxique.

En nous basant sur ces connaissances, nous présentons aux chapitres suivants le travail que nous avons mené afin de construire une banque de données du vietnamien pour les recherches en TAL. Les outils et ressources discutés concernent l'annotation morphosyntaxique (Chapitre 3), l'analyse syntaxique (Chapitre 4), ainsi que l'alignement multilingue (Chapitre 5).

Outre les nombreux travaux existant sur l'anglais et sur le français, grâce auquel nous bénéficions d'un important héritage de méthodes et d'outils, nous prenons comme références les travaux menés sur le chinois et le thaï, qui ont sans doute beaucoup de points communs avec le vietnamien. De plus, le chinois est étudié par une communauté importante, et les recherches en TAL pour le thaï ont commencé bien plus tôt que pour le vietnamien.

Chapitre 3

Construction d'outils et ressources linguistiques pour l'analyse morphosyntaxique du vietnamien

Nous présentons dans cette partie les travaux sur l'annotation morphosyntaxique des corpus vietnamiens. Il s'agit de la construction des ressources lexicales (lexique, corpus annotés) du vietnamien et des outils d'étiquetage morphosyntaxique. Nous insistons particulièrement sur le fait qu'il n'y a pas, jusqu'à présent, de consensus sur la question de parties de discours du vietnamien dans la communauté linguistique. Une partie importante de notre travail est donc de construire un lexique avec des descriptions lexicales qui nous permettent de définir ultérieurement les jeux d'étiquettes comparables pour la tâche d'étiquetage. Nous discutons ensuite du problème de segmentation des textes vietnamiens en unités lexicales, et des solutions possibles. Enfin, nous présentons une méthode statistique simple fondée sur l'utilisation d'un modèle de Markov caché pour l'étiquetage automatique de corpus vietnamiens. Par ailleurs, toutes les ressources construites font l'objet d'une discussion sur leur représentation normalisée.

- *Introduction*
- *Méthodes pour l'étiquetage morphosyntaxique*
- *Construction de ressources lexicales*
- *Annotation morphosyntaxique de textes vietnamiens*
- *Bilan et perspectives*

3.1. Introduction

Chaque mot d'une langue appartient potentiellement à une ou plusieurs parties du discours selon son contexte d'utilisation. L'étiquetage lexical consiste à attribuer une étiquette morphosyntaxique reflétant des informations morphologiques et grammaticales (la catégorie syntaxique, le lemme, le genre, le nombre, *etc*.) à chaque mot d'un texte. Cette tâche est essentielle pour tout traitement ultérieur comme l'analyse syntaxique, sémantique ou même pragmatique d'une langue. Le problème d'étiquetage d'un corpus consiste à :

- définir les unités lexicales (« mots ») et le jeu d'étiquettes ;
- segmenter le corpus en suite de mots et appliquer une méthode d'étiquetage sur le corpus.

Il existe aujourd'hui différents outils pour l'étiquetage morphosyntaxique, ainsi que d'immenses ressources de corpus annotés destinées à des traitements variés dans nombreuses langues (*cf*. section 1.1.3). Cela suppose également l'existence de définitions variées d'unités lexicales et de jeux d'étiquettes selon l'objectif visé. Dans le cadre de MULTEXT (*cf*. Ide et Véronis [IDE 94]) et de « MULTEXT goes East » (*cf*. Erjavec *et al*. [ERJ 96]), des jeux d'étiquettes ont été définis pour une dizaine de langues avec un haut niveau de consensus au sujet de la structure de la description.

Se posent également les questions cruciales du caractère réutilisable de ces ressources linguistiques pour un nombre croissant d'applications, leur réutilisation combinée dans un contexte multilingue, et l'adaptation d'un outil à d'autres langues. De multiples projets ont vu le jour dans cette optique, visant à l'évaluation des outils, à la normalisation et à la représentation des structures de description morphosyntaxique (*cf*. Ide et Romary [IDE 01b]).

Dans le cas des textes vietnamiens, le travail d'étiquetage est une tâche nouvelle et difficile pour les informaticiens, essentiellement du fait du désaccord existant sur la classification linguistique traditionnelle des mots au sein de la communauté linguistique. À ce jour il n'existe aucun standard reconnu pour les catégories des mots en vietnamien. Un autre problème concerne la segmentation du corpus en unités lexicales, les mots composés (comme en français « pomme de terre ») étant très fréquents en vietnamien. Notre recherche vise deux objectifs principaux : en premier lieu créer des outils et des ressources linguistiques pour les applications de traitement automatique des textes vietnamiens, mais aussi assurer la disponibilité de ces outils pour les linguistes travaillant sur le vietnamien.

Dans ce chapitre, après un bref état de l'art du domaine de l'étiquetage morphosyntaxique (section 3.2), nous abordons deux sujets :

- ***Construction d'un lexique vietnamien*** : dans un premier temps, partant d'un dictionnaire sur papier, nous construisons un lexique du vietnamien, en créant des descriptions lexicales pour chaque entrée (section 3.3). Le lexique obtenu sert à la définition des jeux d'étiquettes et à la tâche de segmentation de textes vietnamiens, ainsi qu'aux tâches ultérieures de traitement de la langue vietnamienne.
- ***Développement d'un système de construction et de gestion de corpus étiquetés*** du vietnamien permettant d'effectuer les tâches d'analyse morphosyntaxique (traitement automatique et validation manuelle) sur les textes bruts (section 3.4).

Nous prêtons également une grande attention aux questions de normalisation de la gestion des ressources (lexiques, corpus annotés), en vue de faciliter l'échange de données avec d'autres systèmes et de les mettre à la disposition de la communauté de recherche. Ce sujet est discuté aux sous-sections 3.3.4 (pour les bases lexicales) et 3.4.2 (pour les corpus annotés). Nous concluons ce chapitre par un bilan et un plan de travail concret à mener par la suite pour doter le vietnamien d'outils et ressources de base en TAL du niveau de ceux des langues plus étudiées comme l'anglais et le français.

3.2. Méthodes pour l'étiquetage morphosyntaxique

Un état de l'art de l'étiquetage morphosyntaxique est présenté dans Paroubek et Rajman [PAR 00]. L'étiquetage lexical automatique s'effectue usuellement en trois étapes :

- segmentation du texte en unités lexicales,
- étiquetage *a priori*, c'est-à-dire association à chaque occurrence de mot de toutes ses étiquettes possibles, et
- désambiguïsation, ou sélection parmi ces étiquettes possibles de la seule correcte.

Dans les sous-sections suivantes nous présentons la méthodologie de chaque tâche, ainsi que les mesures d'évaluation des systèmes d'étiquetage automatique. Mais avant tout, nous discutons de la définition de l'unité lexicale et du jeu d'étiquettes, sans lesquels l'étiquetage ne peut avoir de sens. Nous finissons par un bilan permettant de définir les tâches que nous concrétisons pour le vietnamien dans les sections suivantes (3.3 et 3.4).

3.2.1. Définition d'unité lexicale et d'étiquettes

La notion de l'unité « mot » dans une tâche d'étiquetage lexical ne correspond pas nécessairement à un mot traditionnel (*cf.* Vergnes et Giguet [VER 98]). Un mot traditionnel peut être divisé en plusieurs unités ou morphèmes (dans le cas d'amalgames ou de mots composés, par exemple). Au contraire, plusieurs mots en séquence peuvent être groupés en une seule unité, comme par exemple les locutions, les mots ou noms propres composés, les nombres, les dates, *etc*. Pour les langues asiatiques, l'unité « mot » est sujet d'une décision plus difficile à cause de la fréquence des mots composés (*cf.* section 2.3 pour le cas du vietnamien). Ainsi, dans la communauté de recherche en traitement automatique du chinois, il existe deux standards légèrement différents de segmentation d'un texte en « mots » (un en Chine et l'autre à Taiwan). Un projet visant à définir une norme internationale pour la segmentation en mots est mené dans le cadre du sous-comité TC 37/SC 4 (*cf.* section 1.2.2).

Quant au jeu d'étiquettes, se pose la question de la pertinence des parties du discours traditionnelles. Dans [PRZ 03], Przepiórkowski et Woliński défendent l'idée que plusieurs jeux d'étiquettes existant pour le polonais sont linguistiquement naïfs du fait de l'adoption directe, sans analyse critique préalable, des classes traditionnelles de parties du discours, ce qui entraîne en particulier un manque de réutilisabilité. Ils proposent une nouvelle classification purement morphosyntaxique. Pour une langue isolante comme le vietnamien, l'application directe des parties du discours traditionnelles des langues occidentales n'est pas évidente, et l'analyse morphologique n'a pas beaucoup de sens. La définition des étiquettes demande donc une démarche reposant notamment sur les capacités de combinaison des mots.

Les étiquettes grammaticales reflètent des oppositions diverses dans le système syntaxique. Le principal critère pour la définition du jeu d'étiquettes est donc la distribution syntaxique. On devrait donc *a priori* avoir un important jeu d'étiquettes pour refléter exactement toutes les relations syntaxiques. Cependant, plus le jeu d'étiquettes est important, plus la tâche d'annotation est difficile. Aussi, on a souvent besoin d'un compromis pour parvenir à un jeu d'étiquettes assez précis et de taille acceptable. Tufiş [TUF 99] propose un étiquetage à deux passes dans le but de réduire les coûts de temps et de mémoire dans le processus d'étiquetage exploitant un jeu de plus de 700 étiquettes. Ainsi, un texte est d'abord étiqueté avec un jeu d'étiquettes plus grossières, puis le résultat est raffiné avec des étiquettes plus fines.

Dans un souci de réutilisabilité et de compatibilité des corpus étiquetés, on peut constater des efforts importants de normalisation de plusieurs projets. Dans le cadre du projet EAGLES, des jeux d'étiquettes de granularité variable ont été définis pour plusieurs langues européennes. Ces descriptions ont été reprises et affinées dans le cadre du projet MULTEXT (*cf.* 3.3.1), qui a également proposé le principe de définir les étiquettes du corpus en créant une application mathématique de l'ensemble de descriptions lexicales (qui sont en générale stables) au jeu d'étiquettes. Les étiquettes de deux jeux d'étiquettes ainsi définis peuvent alors être facilement mises en correspondance.

3.2.2. Segmentation

Le problème de segmentation n'est pas le même pour toutes les langues. Dans cette section, nous présentons les approches pour la segmentation des textes de langues indo-européennes flexionnelles, puis de langues isolantes asiatiques.

3.2.2.1. Langues indo-européennes : les langues flexionnelles

Pour les langues comme l'anglais ou le français, les unités lexicales (*token*) sont dans la plupart des cas reconnaissables par une simple analyse orthographique, en s'appuyant sur les caractères de séparateur (espaces, ponctuations, *etc*.) dans les textes.

Cependant, cette segmentation aveugle sans information syntaxique ou sémantique ne permet pas l'identification des unités composées, par exemple des segments « `pomme de terre` », « `avant que` », *etc*. du français. Une approche possible et formellement définie dans le cadre du projet GRACE (*cf.* 3.2.5.2) est de représenter les ambiguïtés de segmentation sous la forme d'ambiguïtés d'étiquetage portant sur les tokens minimaux produits par une segmentation de référence aussi fine que possible (Paroubek et Rajman [PAR 00]). Dans cette approche, la composition est indiquée par l'adjonction d'une étiquette supplémentaire à chacun des constituants de la forme composée, de la forme *X / i.j*, où *i* est la position du constituant dans la forme composée de type *X* et de taille *j*. Par exemple, si l'on assigne l'étiquette *Adv* (adverbe) à l'unité « `avant` », l'étiquette *SConj* (conjonction subordonnée) à l'unité « `que` », et l'étiquette *SConj* composée à l'unité composée de deux unités simples « `avant que` » ; alors l'étiquetage complexe donne *avant*[Adv | SConj/1.2] *que*[SConj | SConj/2.2].

3.2.2.2. Langues asiatiques : les langues isolantes

Les langues isolantes comme le chinois, le vietnamien ou le thaï sont en principe monosyllabiques, ce qui signifie que chaque syllabe peut être une unité lexicale. Il est également possible et courant dans ces langues de former des mots complexes à partir de plusieurs syllabes (*cf.* section 2.3), ce qui rend difficile le problème de segmentation. En fonction de l'écriture des langues, la segmentation est plus ou moins difficile. Pour le thaï, une syllabe est transcrite par plusieurs caractères de l'alphabet, et il n'y a pas de frontière entre les syllabes dans le texte (*cf.* Kawtrakul *et al.* [KAW 02]). Cela augmente la difficulté de la segmentation par rapport au chinois et au vietnamien, où il y existe une frontière entre les unités de base. Pour le chinois, chaque caractère représente une unité de base (*hanzi*, équivalente à « tiếng » du vietnamien, *cf.* Sproat *et al.* [SPR 96] et 2.3.1), alors que pour le vietnamien, les unités de base se sont séparées par des espaces.

Comme résumé dans Ha L. A. [HAL 03], le problème de segmentation des textes pour ces langues peut être divisé en deux sous-problèmes avec les approches suivantes :

- Désambiguïsation des séquences de mots avec l'aide d'un lexique et de méthodes statistiques (*cf.* Sproat *et al.* [SPR 96], Wong et Chan [WON 96] pour le chinois). Généralement, un dictionnaire est utilisé pour reconnaître des mots en s'appuyant sur les algorithmes gloutons et *maximal matching* (recherche des correspondances les plus longues). En cas d'ambiguïté, des calculs statistiques sont employés pour la désambiguïsation.
- Identification des mots inconnus par la détection de la collocation des unités de base en reposant sur des mesures statistiques comme l'information mutuelle ou le *t*-score (*cf.* Sun *et al.* [SUN 98] pour le chinois, Sornlertlamvanich *et al.* [SOR 00] pour le thaï).

Plusieurs systèmes d'analyse morphosyntaxique utilisent également des informations de catégorie syntaxique pour la désambiguïsation de la segmentation, et intègrent en conséquence la tâche de segmentation à celle d'étiquetage morphosyntaxique (*cf*. Feng *et al.* [FEN 04]).

3.2.3. Étiquetage *a priori*

L'étiquetage *a priori* consiste à affecter à un mot l'ensemble des étiquettes qui peuvent lui être associées hors contexte à l'aide d'un lexique ou d'une analyse morphologique (Paroubek et Rajman [PAR 00]).

À cette étape, l'étiqueteur peut chercher dans le dictionnaire la liste des étiquettes possibles de chaque unité lexicale. En cas de mot inconnu, la solution la plus naïve est d'associer à ces mots toutes les étiquettes existantes. Pour les langues flexionnelles, une solution possible est d'ajouter un module de prédiction en se basant sur des règles de généralisation morphologique (*cf*. Chanod et Tapanainen [CHA 95]).

3.2.4. Désambiguïsation

L'entrée de cette étape est une séquence unique de segments dont chacun des éléments est associé à une ou plusieurs étiquettes morphosyntaxiques. L'objectif de la désambiguïsation est alors de sélectionner la séquence d'étiquettes qui correspond le mieux à la séquence des mots à étiqueter.

Il existe deux approches principales pour la tâche de désambiguïsation : les méthodes à base de règles et les méthodes probabilistes.

3.2.4.1. Méthodes à base de règles

Les méthodes à base de règles exploitent un ensemble de règles grammaticales pour résoudre le problème de l'étiquetage.

Les méthodes non supervisées utilisent des contraintes produites par les linguistes et un lexique contenant pour chaque mot ses étiquettes possibles. Un tel étiqueteur s'apparente à un parseur, comme par exemple les systèmes GREYC et SYLEX présentés dans l'évaluation GRACE. La principale difficulté pour le développement de systèmes à base de règles réside dans deux aspects : la quantité d'information devant être encodée par des experts – au moins de l'ordre du millier de règles (*cf*. Samuelsson et Voutilainen [SAM 97]) – et la définition des contraintes régissant l'ordre d'application des règles.

Les méthodes supervisées construisent les étiquettes et les règles de transformation à partir de corpus étiquetés manuellement. L'étiqueteur de Brill [BRI 95] est l'exemple le plus connu de telles méthodes. Lors de l'étiquetage, à chaque mot est attribuée l'étiquette la plus fréquente enregistrée dans le lexique. Ensuite, les règles de transformation produites à partir du corpus d'entraînement servent à la correction itérative de cet étiquetage préalable. Brill a montré que sa méthode est théoriquement plus puissante que celles utilisant des arbres de décision (*cf*. Schmid [SCH 94b]). D'après la comparaison présentée dans Manning et Schütze [MAN 99], par rapport aux méthodes probabilistes et les automates, la méthode de Brill offre une meilleure robustesse (moindre sensibilité au « sur-ajustement » par rapport aux données d'entraînement) et une grande lisibilité des informations manipulées par le système, mais ceci au détriment d'une perte de finesse par rapport aux systèmes probabilistes qui manipulent des quantités numériques permettant des discriminations plus fines.

Un avantage des méthodes à base de règles est qu'elles peuvent être mises en œuvre de façon efficace. Par exemple, Shabes et Roche [SHA 95] montrent comment transformer automatiquement les règles apprises par le système de Brill en un transducteur déterministe à états finis (complexité linéaire en temps).

3.2.4.2. Méthodes probabilistes

Les méthodes probabilistes (dont les systèmes utilisant le modèle de Markov caché) utilisent la distribution de probabilité sur l'espace des associations possibles entre les séquences de mots et les séquences d'étiquettes. Cette distribution est produite à partir d'un corpus d'apprentissage étiqueté ou non. La désambiguïsation entre les étiquettes d'un mot s'opère par le choix de la séquence d'étiquettes qui maximise la probabilité conditionnelle de l'association avec la séquence de mots courante.

Ces méthodes reposent sur deux hypothèses :

(1) La probabilité d'association entre un mot et une étiquette est entièrement conditionnée par la connaissance de l'étiquette.

(2) La probabilité d'occurrence d'une étiquette est exhaustivement conditionnée par la connaissance d'un nombre fixe d'étiquettes voisines.

Cela peut se traduire mathématiquement comme suit : Soient m_i est le mot à la position i de la séquence des mots $S_m = m_1 m_2 \ldots m_N$ à étiqueter. On doit chercher une séquence d'étiquettes $S_e = e_1 e_2 \ldots e_N$, dont e_i est l'étiquette assigné au mot m_i qui maximise la probabilité conditionnelle $P(e_1 e_2 \ldots e_N \mid m_1 m_2 \ldots m_N)$. Les deux hypothèses ci-dessus correspondent aux formules suivantes :

(1) $P(m_i \mid m_1 \ldots m_{i-1}, e_1 \ldots e_N) = P(m_i \mid e_i)$

(2) $P(e_i \mid e_1 \ldots e_{i-1}) = P(e_i \mid e_{i-k} \ldots e_{i-1})$

où k est la taille du voisinage mentionné dans la deuxième hypothèse.

De (1), (2) et la formule de Bayes, nous obtenons :

$$P(e_1 e_2 \ldots e_N \mid m_1 m_2 \ldots m_N) = P(m_1 m_2 \ldots m_N \mid e_1 e_2 \ldots e_N) * P(e_1 e_2 \ldots e_N)$$
$$= P(e_1)P(m_1 \mid e_1)P(e_2 \mid e_1)\ P(m_2 \mid e_2) \ldots P(e_N \mid e_{N-k} \ldots e_{N-1})P(m_N \mid e_N)$$

Le modèle probabiliste obtenu est donc équivalent à un modèle de Markov caché dont les états cachés sont les séquences de k étiquettes morphosyntaxiques possibles et les états observables sont les mots du vocabulaire. En fonction de la valeur de $k = 0, 1, 2, \ldots$ nous avons les modèles d'étiquetage probabiliste uni-gramme, bi-gramme, tri-gramme…, dont les modèles bi-gramme et tri-gramme sont les plus utilisés.

Les probabilités définissant le modèle peuvent être calculées directement à partir d'un corpus de textes étiquetés manuellement, ou estimées sur les données non désambiguïsées à l'aide d'un algorithme itératif de type Baum-Welch (*cf.* Belaïd [BEL 92]). La deuxième approche, dans la plupart des cas, aboutit à des valeurs de paramètres de qualité médiocre. La solution la plus efficace consiste à utiliser une démarche incrémentale au cours de laquelle une petite quantité de texte étiqueté manuellement sert à initialiser un étiqueteur probabiliste dont les résultats sont soumis à correction manuelle afin d'augmenter progressivement la quantité de texte étiqueté disponible (*cf.* Marcus *et al.* [MAR 93], Merialdo [MER 94]).

3.2.4.3. Autres méthodes

Outre les méthodes présentées ci-dessus, on peut encore citer d'autres approches comme par exemple :

- Méthodes connexionnistes, à base de réseaux de neurones artificiels, pour lesquels très peu d'expériences sont relatées dans la littérature (*cf.* Manning et Schütze [MAN 99], Schmid [SCH 94a]). Deux raisons expliquent ce relatif manque d'intérêt : d'une part, les performances comparables ou supérieures atteintes avec les autres méthodes généralement plus faciles à mettre en œuvre, d'autre part le manque de fondement théorique justifiant le recours à un formalisme neuromimétique pour la réalisation d'une telle tâche.
- Systèmes hybrides combinant
    - l'approche linguistique et l'approche probabiliste (El-Bèze et Spriet [ELB 95]),

- des étiqueteurs différents (Mason et Tufiş [MAR 98]), ou plusieurs versions du même étiqueteur entraîné sur des données différentes (Tufiş [TUF 99]), *etc*.

3.2.5. Évaluation des étiqueteurs morphosyntaxiques

Nous introduisons tout d'abord ici les principes de l'évaluation des systèmes d'étiquetage (section 3.2.5.1), puis présentons la campagne d'évaluation GRACE des systèmes d'étiquetage pour le français comme une bonne illustration de la méthode d'évaluation et de la performance atteinte par différents systèmes (section 3.2.5.2).

3.2.5.1. Mesures d'évaluation

Un étiqueteur morphosyntaxique est un système complexe constitué de plusieurs modules (segmentateur, étiqueteur non contextuel, désambiguïseur) utilisant des ressources variées (lexique, jeu d'étiquettes, *etc*.). L'évaluation d'un tel système peut donc concerner plusieurs aspects différents (*cf*. Paroubek et Rajman [PAR 00]).

Dans cette présentation, nous nous limitons au problème de l'évaluation de la précision de l'étiquetage réalisé par un outil, c'est-à-dire le taux d'étiquetage correct. Cependant, un taux seul ne signifie que peu de chose dans la comparaison entre les systèmes, car la précision de chaque système dépend du mode de segmentation et du jeu d'étiquettes utilisé, ainsi que des données de test utilisées. Une démarche d'évaluation simple peut comparer les systèmes en donnant aux systèmes évalués un corpus d'évaluation avec un étiquetage manuel de référence se reposant sur une segmentation et un jeu d'étiquettes donnés.

Pour l'évaluation relative d'un système d'étiquetage, on peut considérer les éléments suivants :

- Une évaluation des types d'ambiguïté pour apprécier la difficulté d'étiquetage : le nombre moyen d'étiquettes possibles à assigner à chaque mot, et les types (ou classes) d'ambiguïté, en précisant notamment la fréquence relative dans le texte de test de chacune de ces classes.
- Une évaluation des types d'erreurs : les types d'ambiguïté conduisant le plus fréquemment aux erreurs d'étiquetage.
- Une évaluation de la précision d'étiquetage.

La section suivante présente une approche pour l'évaluation des systèmes d'étiquetage morphosyntaxique du français.

3.2.5.2. Exemple d'évaluation des systèmes d'étiquetage : Projet GRACE

L'action GRACE (Grammaires et Ressources pour les Analyseurs de Corpus et leur Évaluation, 1995 - 1998)[26] vise la mise en place d'un paradigme d'évaluation pour les analyseurs morphosyntaxiques et syntaxiques du langage naturel et la constitution d'un premier noyau de données réutilisables pour l'évaluation de systèmes linguistiques d'analyse du français. La première session d'évaluation GRACE-I a porté sur les assignateurs de catégories grammaticales (*taggers*) pour le français.

[26] http://www.limsi.fr/TLP/grace/

Dans le cadre de GRACE, un système d'étiquetage est évalué par comparaison de l'étiquetage produit par le système avec l'étiquetage de référence produit par un groupe d'experts humains (le corpus de test est constitué d'environ 650 000 formes). La procédure d'évaluation nécessite de définir un référentiel commun permettant de comparer des systèmes qui n'utilisent pas nécessairement les mêmes jeux d'étiquettes et les mêmes unités lexicales. Le projet GRACE a ainsi choisi une méthode de segmentation minimale (la plus fine possible, *cf.* 3.2.2.1) et 312 étiquettes dérivées du modèle MULTEXT (*cf.* 3.2.1 et 3.3.1) et adaptées par itérations successives en interaction avec les participants du projet pour construire le corpus étiqueté de référence (*cf.* Adda *et al.* [ADD 99]). Chaque système participant fournit donc une table de correspondances pour projecter leurs étiquettes vers les étiquettes du corpus de référence. La segmentation du système à évaluer est pour sa part alignée avec celle du corpus de référence.

L'action GRACE permet d'évaluer la performance des systèmes qui attribuent non seulement une étiquette à chaque unité lexicale mais aussi une liste d'étiquettes. Pour cela, on distingue deux types d'étiquetage : l'étiquetage « strict » (correct ou erroné) consiste à associer à l'unité lexicale une étiquette unique (correcte ou erronée), et l'étiquetage « ambigu » (silence) consiste à associer à l'unité lexicale une liste d'étiquettes possibles. Au cas où cette liste ne contient que des étiquettes correctes, on a un « silence ok ». Si elle ne contient que des étiquettes erronées, on a un « silence erroné ». Dans les autres cas on a un « silence vrai ».

L'évaluation d'un système d'étiquetage peut alors être caractérisée par les sept grandeurs suivantes : nbCas – nombre total d'occurrences d'unités lexicales, NONEVAL – nombre d'occurrences d'unités lexicales non retenues pour l'évaluation (en raison des problèmes d'étiquetage, de segmentation ou de réalignement), OK – nombre d'étiquetages corrects, ERR – nombre d'étiquetages erronés, SIL_{ok} – nombre de silences ok, SIL_{err} – nombre de silences erronés, et $SILOK_{moy}$ – nombre moyen de silences qui pourraient être transformés en cas « étiquetage correct » par un choix aléatoire équiprobable effectué, pour chaque silence, parmi les alternatives évaluables proposées par l'étiqueteur évalué.

Pour une évaluation avec moins de dimensions, on a défini deux nouvelles grandeurs : la précision P quantifie la « justesse » des résultats d'un étiqueteur, et la décision D indique dans quelle mesure cet étiqueteur produit des résultats directement exploitables (sans traitement additionnel pour lever les ambiguïtés d'étiquetage).

- $P = OK/(OK + ERR)$,
- $D = (OK + ERR)/(OK + ERR + SIL)$, où SIL est le nombre total de silences : $nbCas = NONEVAL + OK + ERR + SIL$.

3 autres précisions supplémentaires sont définies :

- $P_{min} = (OK + SIL_{ok})/(OK + ERR + SIL)$ – précision minimale que pourrait avoir le système évalué si l'on force sa décision à 1 ;
- $P_{max} = (OK + SIL - SIL_{err})/(OK + ERR + SIL)$ – précision maximale que pourrait avoir le système évalué si l'on force sa décision à 1 ;
- $P_{moy} = (OK + SILOK_{moy})/(OK + ERR + SIL)$ – précision moyenne qu'aurait le système évalué s'il était complété, pour les cas de silence, par une procédure de choix aléatoire (équiprobable) parmi les alternatives proposées.

Les quatre points (P, D), (P_{min}, 1), (P_{max}, 1), (P_{moy}, 1) forment une zone de fonctionnement facile à visualiser d'un système d'étiquetage, ce qui permet de donner un support concret à l'étude comparative des performances obtenues par plusieurs systèmes. En effet, on peut considérer P_{moy} comme la mesure de la qualité d'un étiqueteur avec son intervalle de variation [P_{min}, P_{max}].

Le meilleure précision observée dans l'évaluation du projet GRACE était de 97.8%, alors qu'un étiqueteur qui retourne toutes les étiquettes associées à une forme donne 88% de précision. Ce dernier score tombe à 59% quand on applique quelques règles non contextuelles pour ramener arbitrairement le nombre d'étiquettes proposées à une par mot (Paroubek et Rajman [PAR 00]).

3.2.6. Bilan et plan de la présentation

Nous avons rapidement présenté un état de l'art de l'étiquetage morphosyntaxique, en détaillant les problèmes et les méthodes utilisées étape par étape : segmentation, étiquetage *a priori*, désambiguïsation et évaluation. Comme nous l'avons vu, un système d'étiquetage morphosyntaxique peut éventuellement faire l'usage d'un lexique fournissant les étiquettes possibles (avec ou sans fréquence correspondante) de chaque mot ou un corpus d'apprentissage étiqueté et validé manuellement. Avant de construire toutes ces ressources, la première question fondamentale est la définition des unités lexicales et du jeu d'étiquettes utilisé.

Pour l'analyse morphosyntaxique du vietnamien, il nous faut développer tous les outils et ressources à partir de zéro. Or le coût de création des ressources linguistiques est très élevé. Nous nous donnons donc pour objectif de construire une base de lexique et corpus étiquetés ouverte dans un cadre normalisé afin d'augmenter l'extensibilité et la réutilisabilité de ces données.

Dans le but de construire une base lexicale hautement réutilisable, nous faisons le même choix que l'action GRACE, c'est-à-dire que nous considérons l'unité lexicale comme l'unité la plus fine ayant une définition stable dans la littérature linguistique vietnamienne (*cf*. section 2.3.2).

En ce qui concerne la définition des jeux d'étiquettes, nous partons des recommandations du projet MULTEXT (*cf*. sections 3.2.1 et 3.3.1), chaque jeu d'étiquettes doit pouvoir être mis en correspondance avec des descriptions lexicales stables sur des bases essentiellement linguistiques. Notre premier travail est donc de construire un lexique qui fournisse de telles descriptions pour chaque unité lexicale. Pour cela, nous nous sommes volontairement basée sur le modèle de descriptions morphosyntaxiques du projet MULTEXT, intrinsèquement dédié aux applications multilingues. Si le modèle MULTEXT est orienté morphologie, notre modèle se base plutôt sur la sémantique, car c'est comme nous l'avons déjà précisé (*cf*. section 2.4) le seul critère qui permette en vietnamien d'exprimer les restrictions de combinaison des mots dans un groupe syntagmatique. Nous visons également à proposer un ensemble de références de descripteurs lexicaux pour le vietnamien à intégrer dans le répertoire de catégories de données (DCR) du comité TC 37/SC 4 de l'ISO (*cf*. section 1.2.2). La construction et la gestion des ressources lexicales sont présentées à la section 3.3.

Nous étudions ensuite, à la section 3.4, le problème d'automatisation de l'étiquetage du vietnamien. Notre objectif n'est pas d'inventer une nouvelle méthodologie, mais d'identifier les problèmes à résoudre pour le vietnamien, d'expérimenter l'application des méthodes « classiques » développées pour les langues occidentales au vietnamien, ainsi que de construire une base de données (corpus brut, segmenté et étiqueté) exploitable et extensible pour les études ultérieures. Nous construisons donc un système d'annotation morphosyntaxique pour les textes vietnamiens, en discutant des sujets suivants :

- Jeux d'étiquettes
- Segmentation
- Étiquetage morphosyntaxique et évaluation
- Codage normalisé de corpus annotés.

Un plan du travail restant à accomplir est présenté en guise de conclusion, à la section 3.5 de ce chapitre.

3.3. Construction de ressources lexicales

S'il existe de nombreux dictionnaires papiers du vietnamien, aucun lexique de TAL n'est à l'heure actuelle publiquement disponible pour la recherche. La construction d'une telle ressource se heurte d'ailleurs au problème de manque de consensus linguistique concernant la classification détaillée des mots vietnamiens (*cf*. 2.4.1.8).

Notre objectif est de construire un lexique morphosyntaxique extensible et réutilisable visant l'annotation morphosyntaxique des textes. À chaque entrée lexicale doivent être associées des descriptions lexicales qui reflètent ses propriétés morphosyntaxiques, et qui permettent de définir ou comparer les jeux d'étiquettes lexicales. Ce lexique peut également constituer une base pour le développement ultérieur d'un lexique syntaxique. Il a vocation à constituer une ressource « ouverte » dans la communauté du TAL.

Dans le cadre de cette thèse, ainsi que du projet national vietnamien KC01-03 (*cf*. section 1.3), avec la coopération du Centre de Lexicographie du Vietnam, nous avons réalisé la construction d'un tel lexique couvrant toutes les entrées du dictionnaire vietnamien Hoàng Phê [HOA 02][27]. La structure de ce dictionnaire est présentée à la section 3.3.3. En outre, nous réalisons également la conversion du format MS Word du dictionnaire vietnamien en format XML d'après les recommandations de la TEI (*cf*. 1.2.1.3) pour que ses informations lexicales soient facilement exploitables dans d'autres applications. Nous avons donc deux tâches principales :

- Étudier le modèle formel ainsi que le codage pour représenter le dictionnaire papier et le lexique morphosyntaxique ;
- Définir les descriptions lexicales pour les entrées du lexique morphosyntaxique, et ceci dans le contexte de différentes opinions sur la classification des unités lexicales vietnamiennes.

Cette section commence par la présentation du modèle sur lequel nous nous appuyons pour définir les descripteurs lexicaux du vietnamien (section 3.3.1). Puis nous détaillons les descriptions grammaticales du lexique vietnamien (section 3.3.2). Ensuite, nous présentons le processus de construction de notre lexique (section 3.3.3), et concluons enfin en abordant le codage adopté pour la gestion des ressources lexicales (section 3.3.4).

3.3.1. Modèle de description lexicale

Des efforts importants ont porté sur la normalisation des données, des outils et des ressources linguistiques pour favoriser leur réutilisabilité pour les recherches et les applications en traitement des langues à base de corpus. MULTEXT (*Multilingual Text Tools and Corpora* - http://www.lpl.univ-aix.fr/projects/multext/) en est un exemple significatif. Dans le cadre de ce projet, un modèle morphosyntaxique a été développé en vue de l'harmonisation de l'étiquetage de corpus multilingues ainsi que de la comparabilité des corpus étiquetés. MULTEXT défend l'idée que dans un contexte multilingue, des phénomènes identiques devraient être encodés de manière similaire dans chaque langue pour faciliter les traitements dans des applications diverses (alignement automatique, extraction de terminologie multilingue, *etc*.).

[27] Ce dictionnaire est le produit constamment mis à jour de l'Institut national de Linguistique et le Centre de Lexicographie du Vietnam, donc nous l'avons choisi pour construire notre lexique.

Un principe du modèle est de *séparer les descriptions lexicales, qui sont en générale stables, et les étiquettes d'un corpus donné*, en général définies pour répondre à un besoin particulier. En ce qui concerne les descriptions lexicales, le modèle utilise deux couches : un noyau commun pour des catégories communes et une couche privée contenant des informations additionnelles qui sont propres à une langue ou aux applications particulières.

Une solution de compromis pour les jeux d'étiquettes morphosyntaxiques dans le noyau commun est un jeu de 11 étiquettes reprises des observations du projet EAGLES : Nom (N), Verbe (V), Adjectif (A), Pronom (P), Déterminant (D), Adverbe (R), Adposition (S), Conjonction (C), Numéral (M), Interjection (I), Résidu (X). L'information optionnelle de la deuxième couche est représentée par les paires attribut-valeur (structure de traits, *cf*. la représentation des dictionnaires à la section 3.3.4.1). Par exemple, un nom commun singulier est représenté par N[type = common gender = masculine number=singular case=n/a] (forme contracté Ncms-). Le passage des descriptions lexicales aux étiquettes de corpus se fait par une application mathématique, qui permet de déterminer clairement la relation entre les jeux d'étiquettes (*cf*. la Figure 3-1, extraite de l'Introduction & Problématique, projet MULTEXT).

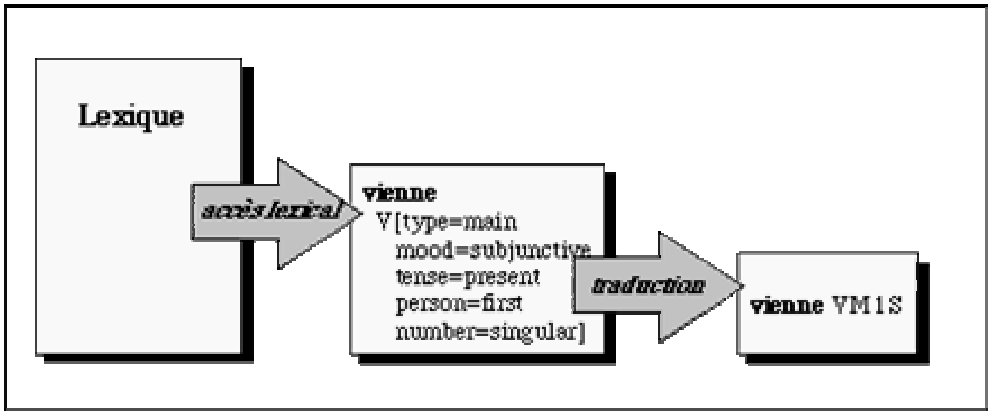


Figure 3-1 Descriptions lexicales et étiquettes de corpus dans le système Multext

Or, il est évident que pour couvrir une plus grande variété de langues, il est nécessaire de présenter plus de flexibilité dans ce cadre fondamental. L'étude que nous présentons sur les descriptions grammaticales du vietnamien prouve qu'en effet quelques catégories peuvent ne pas convenir aux objets linguistiques réels de cette langue. Du point de vue de la normalisation, ceci signifie qu'une étape ultérieure serait soit de décrire une ontologie entière des catégories (comme suggéré par Farrar *et al.* [FAR 02]), soit d'enregistrer la variété des descripteurs possibles à travers les langues en construisant un enregistrement de méta-données (*cf*. Ide et Romary [IDE 01b]). Ces deux options ne sont pas nécessairement contradictoires puisque les catégories de données élémentaires peuvent être mis en parallèle avec les nœuds de l'ontologie, permettant de comparer des jeux d'étiquettes d'une langue à l'autre, ainsi qu'au sein d'une langue donnée. De ce fait, il est important de considérer que pour une langue comme le vietnamien, un schéma d'annotation peut se fonder sur plusieurs couches de granularité d'étiquettes (*cf*. section 3.4.1), et ceci devrait être pris en compte. Notre projet présente une telle stratégie, qui pourrait mener en particulier à la proposition d'un ensemble de descripteurs de référence pour le vietnamien, dans le contexte de construction du répertoire de catégories de données (DCR) du sous-comité TC 37/SC 4 de l'ISO (*cf*. section 1.2.2).

En nous inspirant des principes de construction du modèle MULTEXT, nous avons élaboré des descriptions grammaticales pour le vietnamien dans un schéma comparable à ce modèle. Les descripteurs sont donc organisés en deux couches : la couche noyau, qui contient les grandes catégories grammaticales de base, assez universelles, et la couche privée, qui contient les traits spécifiques à la langue vietnamienne. Comme nous l'avons déjà mentionné (*cf*. 3.2.6), la couche privée de notre modèle se base plutôt sur la sémantique, car c'est le seul moyen de refléter en vietnamien la combinaison des mots dans un groupe syntagmatique. La section suivante présente concrètement ces descriptions, qui sont mises à jour par rapport aux celles introduites dans Nguyen *et al.* [NGU 04a]).

3.3.2. Descriptions lexicales du vietnamien

Le problème de classification des catégories grammaticales en vietnamien est toujours en débat au sein de la communauté linguistique (Hữu Đạt *et al.* [HUU 98], Diệp Q. Ban et Hoàng V. Thung [DIE 99], Cao X. Hạo [CAO 00]). La difficulté vient de l'ambiguïté des rôles grammaticaux de nombreux mots, combinée au phénomène fréquent de mutation catégorielle sans aucune variation morphologique (*cf*. section 2.4.1).

Nous commençons par une classification généralement admise dans la littérature (Comité des Sciences Sociales [UYB 83]), et figurant dans différents dictionnaires vietnamiens dont le dictionnaire Hoàng Phê [HOA 02], sur lequel nous fondons notre lexique. Cette classification, présentée à la section 2.4.1, distingue 8 catégories: Nom, Verbe, Adjectif, Pronom, Mot complément, Conjonction, Interjection, Mot modal.

Les catégories dans la couche noyau sont définies en réorganisant les 8 catégories ci-dessus en 11 catégories, fondées sur leurs sous-catégories. Une catégorie « Mot Modal » est ajoutée par rapport au modèle EAGLES/MULTEXT (*cf*. Tableau 3-1, les catégories vietnamiennes « pures » se trouvent associées à une ou plusieurs catégories VN-MULTEXT).

EAGLES/MULTEXT	**Vietnamien [UYB 83]**	**VN-MULTEXT**
Nom	Nom	Nom
Verbe	Verbe	Verbe
Adjectif	Adjectif	Adjectif
Pronom	Pronom	Pronom
Déterminant	Mot complément	Déterminant
Adverbe	Conjonction	Adverbe
Adposition	Interjection	Préposition
Conjonction	Mot modal	Conjonction
Numéral		Numéral
Interjection		Interjection
		Mot modal
Unique		Non-autonome
Résidu		Résidu

Tableau 3-1 Définition des catégories de la couche noyau du modèle de descriptions lexicales

Notons qu'outre les unités lexicales correspondant aux « mots », le dictionnaire contient également deux autres types d'éléments lexicaux :

- Des locutions, auxquelles peuvent éventuellement être associées des parties du discours selon leur usage. Si aucune description détaillée de l'usage d'une locution n'est fournie, nous la plaçons dans la catégorie résiduelle, d'où la nécessité d'avoir la classe « Résidu » ;

- Éléments de dérivation (*cf.* 2.3.3.1) : il s'agit des syllabes (normalement d'origine chinoise) qui sont employées pour créer les mots dérivés. Par exemple : vô = *{négation}*, thời hạn = *terme/délai* => vô thời hạn = sine die, « *jusqu'à nouvel ordre* ». Nous plaçons ces éléments dans une classe, dite « Non-autonome ».

Les sous-sections suivantes décrivent en détail chacune des catégories de la couche noyau et leurs attributs dans la couche privée. Les attributs proposés sont collectés et synchronisés de différentes sources de descriptions grammaticales du vietnamien ([UYB 83, NGU 98b, HUU 98, DIE 99, CAO 00]).

3.3.2.1. Noms

Les noms décrits à la section **2.4.1.1** sont séparés en trois classes : Nom, Numéral (*cf.* 3.3.2.3) et Déterminant (*cf.* 3.3.2.2). Les numéraux et les déterminants correspondent aux mots de la sous-catégorie « noms de quantité ».

Le nom est un mot qui sert à désigner les êtres, les choses, et les idées. Du point de vue des rôles grammaticaux, un nom doit pouvoir :

- être précédé par un mot désignant une quantité et éventuellement une unité de mesure ou une unité naturelle pour former un NP, par ex. nhà = *maison* ⇒ hai cái **nhà** =$_{litt}$ *deux <unité naturelle des objets> maison* = *deux maisons*,
- être suivi par un pronom démonstratif (par ex. *ce*) pour former un NP, par ex. [cái] **nhà** này =$_{litt}$ *maison ce* = *cette maison*,
- et être sujet ou complément d'objet d'une phrase, par ex. [cái] **nhà** này đẹp = $_{litt}$ *maison ce beau* = *cette maison est belle* ; Tôi thích [cái] nhà này =$_{litt}$ je aimer maison ce = *J'aime cette maison*.

Les attributs proposés ci-dessous reflètent les différentes possibilités de combinaison des noms :

- **Type** : distingue les noms communs (*common)* des noms propres (*proper*).
- **Countability** :
  - *non countable* : noms qui ne peuvent pas être précédés directement par un nombre, par ex. cây cối = *plantes*, nhân dân = *peuple*, nước = *eau*.
  - *direct* : noms qui peuvent toujours être précédés directement par un nombre, par ex. cái = <*classificateur d'objet*>, mét = *mètre*, ý kiến = *opinion*.
  - *indirect* : noms qui peuvent être précédés directement par un nombre dans certains contextes particuliers comme l'énumération. Par exemple : trâu = *buffle*, nhà = *maison*.
- **Unit** : réservé aux noms désignant une unité de mesure ou une entité.
  - *natural* : Les classificateurs d'individus portent cette valeur, par ex. con = <*classificateur d'animal*>, bức = <*classificateur d'un objet aplati*>, …
  - *conventional* : unités de mesure conventionnelles ou scientifiques, par ex. mét = *mètre*, giờ = *heure*, nhúm = *pincée*, hào = *centime*, …
  - *collective* : Les classificateurs des collectifs portent cette valeur, par ex. nhóm = *groupe*, toán = *troupeau*, đôi = *couple*, chục = *dizaine*.
  - *administrative* : unités administratives, par ex. tỉnh = *province*, xã = *commune*, ngành = *division*, môn = *discipline*.

- **Meaning** : Il s'agit de traits sémantiques ayant une influence sur la construction de la phrase. Les valeurs sont les suivantes : *object, plant, animal, social-family relation, human, body part, material, food, disease, sense (*par ex. *màu, sắc, mùi, vị - color, flavour, taste), location, time, turn, fact specificator, abstract, other*.

3.3.2.2. Déterminants

Cette classe contient des mots qui sont généralement placés devant un nom pour déterminer une quantité singulière ou plurielle de celui-ci. Les déterminants se distinguent des numéraux par le fait qu'ils ne peuvent pas être employés seuls.

Deux attributs sont définis :

- **Type** : représente le caractère défini ou indéfini d'un déterminant par les deux valeurs suivantes :
  - *definite* : par ex. các = <*marqueur du pluriel*>[28], mọi = *tous* (dans « *tous les …* ») ;
  - *indefinite* : par ex. những = <*marqueur du pluriel*>, một = *un*, mỗi = *chaque*, từng = *chaque*.
- **Number** : représente le nombre d'un déterminant, avec les deux valeurs suivantes :
  - *singular* : par ex. mỗi, một, từng ;
  - *plural* : par ex. các, những, mọi.

3.3.2.3. Numéraux

Le numéral est un mot qui représente un nombre ou un rang.

Pour cette classe, nous définissons le seul trait **Type** recevant les valeurs suivantes :

- *cardinal* : pour les cardinaux, par ex. một = *un* ; mười = *dix* ; mười một = *onze*.
- *approximate* : pour les quantités d'approximation, par ex. dăm = *environ cinq* ; mươi = *une dizaine*.
- *fractional* : pour les valeurs fractionnelles, par ex. nửa = *moitié, demi, mi* ; rưỡi = *(et) demie*, 50% ; rưỡi = *(et) demi* (*de 100 → 150, 1000 → 1500, etc.*).
- *ordinal* : pour les ordinaux, par ex. (thứ) nhất = *(le) premier* ; (thứ) nhì = *(le) deuxième* ; (thứ) ba = (*le) troisième*.

3.3.2.4. Verbes

Le verbe est un mot exprimant un procès, un état ou un devenir. Les rôles grammaticaux des verbes sont les suivants :

- Prédicat des phrases, par exemple Tôi ngủ =$_{litt}$ *Je dormir = Je dors*.
- Sujet de certains types de phrase, par exemple : Học tập là nhiệm vụ chính của học sinh =$_{litt.}$ *Étudier être tâche principale de élève = Étudier est la tâche principale de l'élève*.
- Modifieur d'un nom, par exemple :
  - sách học =$_{litt}$ *livre apprendre = livre scolaire,* bàn ăn =$_{litt}$ *table manger = table à manger,*
  - thuốc uống =$_{litt}$ *médicament boire = médicament buvable,*

[28] En vietnamien, le marqueur déterminant défini singulier est Ø.

- sách tập đọc =$_{litt}$ *livre <faire des exercices> lire = méthode de lecture*…

- Complément d'un verbe, par exemple : dạy hát =$_{litt}$ *enseigner chanter = enseigner le chant*, bắt làm =$_{litt}$ *forcer travailler = forcer à travailler*, xin nghỉ =$_{litt}$ *demander se reposer = demander la permission d'absence*…

- Modifieur d'un verbe, par exemple : chạy ra =$_{litt}$ *courir sortir* = « *courir en sortant* » = *sortir en courant*, teo lại =$_{litt}$ *dépérir <verbe marquant d'une tendance de réduction> = se ratatiner*, bám lấy =$_{litt}$ *se retenir prendre = se retenir à / s'accrocher à*, vứt đi =$_{litt}$ *jeter aller = jeter*…

Les attributs définis pour des verbes sont :

- **Gradability** : distingue les verbes qui peuvent être modifiés par un adverbe de degré comme « rất » = « *très* ».
  - *gradable* : par ex. thích = *aimer*, yêu = *aimer*, ghét = *détester*, giống = *ressembler*, muốn = *vouloir* ;
  - *non-gradable* : par ex. làm = *faire*, đi = *aller*, phải = *devoir*, bắt đầu = *commencer*, kết thúc = *finir*.

- **Meaning** :
  - *copula* : Par ex. là = *être*, làm = *exercer*, *jouer le rôle de* ;
  - *existence* : verbes pouvant être employés au début de la proposition, correspondant à une formulation impersonnelle en français. Par ex. còn = (il) *rester*, có = *avoir*, *il y a*, mất = *perdre*, hết = *n'avoir plus*, hiện = *paraître*, xuất hiện = *apparaître* ;
  - *transformation* : verbes ayant comme complément un groupe nominal ou adjectival. Le complément doit être toujours présent. Par ex. les verbes hoá, biến, trở thành, nên, thành portent le sens de *devenir* ou *changer en*, … ;
  - *aspect* : verbes concernant le déroulement d'un procès. Ils peuvent avoir un complément (groupe nominal ou un groupe verbal) ou non. Par ex. bắt đầu = *commencer*, tiếp tục = *continuer* ;
  - *comparison* : verbes ayant comme complément un groupe nominal ou un groupe prépositionnel (des prépositions như = *comme*, với = *avec*, *etc*.). Par ex. giống = *ressembler*, khác = *différer*, ăn đứt = *surpasser de beaucoup*, bằng = *égaler* ;
  - *modal* : verbes qui sont obligatoirement associé à un verbe principal dont il complémente le sens dans le but d'exprimer un procès non perfectif. Par ex. toan = *tenter*, định = *compter* (*faire*), có thể = *pouvoir*, nên = *devoir* ;
  - *passivity* : verbes ayant comme complément soit un groupe verbal/adjectival, soit un groupe nominal, soit une proposition. Par ex. bị = *subir*, được = *obtenir*, *gagner*, phải = *subir*, *contracter* ;
  - *feeling* : verbes ne pouvant pas être au perfectif, et ayant comme complément soit un groupe nominal, soit un groupe verbal, soit une proposition avec la conjonction « rằng/là » = « *que* ». Par ex. mong = *espérer*, muốn = *vouloir*, ước = *rêver*, yêu = *aimer*, ghét = *détester* ;
  - *utterance* : verbes ayant comme complément soit un groupe nominal, soit une proposition suivant une conjonction. Le complément peut être absent. Par ex. nghĩ = *réfléchir*, nói = *parler* ;
  - *imperative* : verbes ayant deux compléments : un groupe nominal suivi par un groupe verbal. Par ex. khiến, sai, bảo, bắt = *donner l'ordre (plus ou moins fort)* ;

- *causative* : verbes ayant deux compléments : un groupe nominal suivi par un groupe verbal ou adjectival. Par ex. làm = *rendre* ;
- *dative* : verbes ayant deux compléments : un groupe nominal qui est l'objet et un autre groupe nominal qui est la destination. Par ex. cho = *donner*, tặng = *offrir*, gửi = *envoyer*, lấy = *recevoir* ;
- *directive movement* : verbes représentant un mouvement avec direction, ayant un groupe nominal comme complément de destination. Ces verbes peuvent être employés comme complément de certains verbes ou des adjectifs pour indiquer la direction de l'action ou l'évolution d'un état. Par ex. ra = *sortir*, vào = *entrer*, lên = *monter*, xuống = *descendre* ;
- *non directive movement* : verbes représentant un mouvement sans direction particulière. Ils peuvent avoir un groupe prépositionnel comme complément de direction. Par ex. đi = *aller*, chạy = *courir*, bò = *ramper*, lăn = *rouler* ;
- *moving* : verbes agissant sur un objet (correspondant à un groupe nominal comme complément) et dans une certaine direction (correspondant à un groupe prépositionnel comme complément). Par ex. kéo = *tirer*, đẩy = *pousser*, xô = *bousculer*, buộc = *lacer*, cởi = *enlever* ;
- *direct transitive action* : des verbes transitifs comme : đẽo = *tailler*, gọt = *éplucher*, vẽ = *dessiner*, viết = *écrire* ;
- *intransitve action* : des verbes intransitifs comme : ngồi = *s'assoir*, đứng = *se tenir debout*, nằm = *s'allonger*, cười = *rire*.

3.3.2.5. Adjectifs

L'adjectif est un mot permettant la qualification d'un substantif ou d'un procès. Les rôles grammaticaux qu'un adjectif peut jouer sont :

- Prédicat de phrase, par exemple : cô bé sẽ rất ngoan =$_{litt}$ *<petite fille> <marqueur du futur> très sage = La petite fille va être très sage*.
- Modifieur d'un nom ou d'un verbe, par exemple : nhà đẹp =$_{litt}$ *maison beau = la belle maison*, múa đẹp =$_{litt}$ *danser beau = danser joliment*, cô gái vui vẻ =$_{litt}$ *fille joyeux = la jeune fille joyeuse*, sống vui vẻ =$_{litt}$ *vivre joyeux = vivre joyeusement*.
- Parfois sujet de phrase : Xinh đẹp là lợi thế của cô = $_{litt}$ *joli être avantage de elle = Son avantage est d'être jolie*.

Les deux attributs suivants distinguent la distribution syntaxique des adjectifs :

- **Attribute :**
  - *qualitative* : adjectifs qui répondent à la question « comment ? ». Par exemple : tốt = *bon*, xấu = *mauvais*, đẹp = *beau* ;
  - *quantitative* : adjectifs qui répondent à la question « combien ? ». Par exemple : cao = *haut/grand*, thấp = *petit*, rộng = *large*.
- **Gradability** :
  - *gradable* : adjectifs qui peuvent être précédés par un adverbe de degré comme rất = *très*. Par exemple : tốt, xấu, cao, thấp.
  - *non-gradable* : les autres adjectifs, par exemple : đen sì = *trop noir*.

3.3.2.6. Pronoms

Le pronom est un mot qui remplace généralement un objet ou un fait qui a d'une façon ou d'une autre été mentionné auparavant. Il remplit le rôle grammatical de ce qu'il remplace.

Nous retenons les attributs suivants :

- **Type** : distingue les différents types de pronoms :
  - *personal* : par ex. tôi = *je*, chúng tôi = *nous* ;
  - *pronominal* : par ex. mình = *soi-même*, tự = *de soi-même / se (forme pronominale des verbes)* ;
  - *indefinite* : par ex. người ta = *on*, ai = *quiconque* ;
  - *time* : par ex. bây giờ = *maintenant*, bao giờ = *quand* ;
  - *amount* : par ex. cả = *tout*, tất thảy = *tout*, bao nhiêu = *autant que* ;
  - *demonstrative* : par ex. này = *ce*, kia = *ce..là*, nọ = *ce..là*, đấy = *là* ;
  - *interrogative* : par ex. ai = *qui / quel*, gì = *quoi / que / quel*, nào = *quel*, thế nào = *comment* ;
  - *predicative* : par ex. thế = *comme cela* / *ainsi / tel*, vậy = *comme cela / ainsi* ;
  - *reflexive* : nhau = *l'un l'autre*.
- **Person** : seulement pour les pronoms personnels, les valeurs sont comme dans les autres langues : *first, second, third*.
- **Number** : seulement pour les pronoms personnels, les valeurs sont *singular* et *plural*.

3.3.2.7. Adverbes

L'adverbe est un mot utilisé pour modifier le sens d'un verbe, d'un adjectif ou d'un autre adverbe.

Deux attributs sont définis :

- **Type** :
  - *time* : adverbes désignant le temps du verbe d'une phrase, par ex. đã = <*passif*>, vừa = <*passé proche*>, đang = <*présent*>, sẽ = <*futur*>, rồi = <*passé perfectif*>…
  - *degree* : adverbes de degré, utilisés avec la plupart des adjectifs, et des verbes d'impression. rất = *très*, hơi = *un peu*, quá = *trop*…
  - *regularity* : adverbes désignant la régularité ou la continuation d'une action ou d'un état, par ex. đều = <*uniformité*>, cũng = *également*, vẫn = <*toujours / encore*>, cứ = *tout de même* …
  - *imperative* : adverbes pour former l'impératif, par ex. hãy (*positif*), đừng (*négatif*), chớ (*négatif*)…
  - *polarity* : adverbes portant un sens d'affirmation ou négation, par ex. có = <*affirmation*>, không = <*négation*>, chưa = *pas encore*, chẳng = <*négation*>…
  - *modal* : adverbes de phrase, par ex. bỗng = *brusquement*, hình như = *<il paraît que>*
  - *verbal* : d'autres adverbes, ayant souvent des verbes pour origine, par ex. được dans Tôi sửa được cái máy =$_{litt}$ *Je réparer <état d'obtention> <classificateur d'objet> machine = J'ai réussi à réparer la machine*.
- **Position** :

- *pre* : adverbe se trouvant avant le mot qu'il modifie, par ex. đã, đang, sẽ ;
- *post* : adverbe se trouvant après le mot qu'il modifie, par ex. rồi, được ;
- *both* : adverbe pouvant se trouver avant ou après le mot qu'il modifie, par ex. quá.

3.3.2.8. Prépositions

La préposition est un mot placé devant un nom ou un pronom indiquant la position, la direction, le temps ou d'autres relations abstraites.

Un seul attribut est défini pour cette catégorie :

- **Type** :
  - *position* : par ex. trên = *sur*, dưới = *au dessous de*, trong = *à l'intérieur de*, ngoài = *à l'extérieur de* ;
  - *direction* : par ex. từ = *de (dans « Je viens de Hanoi »)*, qua = *via* ;
  - *time* : par ex. từ = *depuis*, vào = *à (dans « à huit heures »)* ;
  - *objective* : par ex. vì = *pour*, cho = *en vue de* ;
  - *target* : par ex. cho = *pour (dans « livres pour enfant »)*, đến = *à (dans « penser à quelqu'un »)* ;
  - *relation* : par ex. của = *de (dans « de quelqu'un »)*, trừ = *à l'exception de* ;
  - *means* : par ex. bằng = *en* (dans « table en bois »), bởi = *par*.

3.3.2.9. Conjonctions

La conjonction est un mot servant à joindre soit deux propositions, soit deux mots ou groupes de mots de même fonction dans une proposition.

Nous définissons un seul attribut pour cette catégorie :

- **Type** : Deux types de conjonctions sont distingués :
  - *coordination* : par ex. và = *et*, cùng = *avec*, ngược lại = *par contre* ;
  - *subordination* : par ex. nếu … thì = *si … alors*.

3.3.2.10. Interjections

L'interjection est un mot ou un son qui exprime une émotion. Ces mots sont généralement autonomes, sans relation syntaxique avec les autres mots de la même phrase. Nous définissons un attribut :

- **Type** :
  - *exclamation* : mots d'exclamation ; ái chà! = *fichtre !*, ôi chao! = *oh !*
  - *onomatopoeia* : transcriptions de sons d'exclamation. Par ex. ê! = *eh !*, a! = *ah !*, á! = *aie !*

3.3.2.11. Mots modaux

Le mot modal est un mot ajouté soit devant un mot, soit à la fin d'une phrase afin d'exprimer le sentiment (par ex. intensification, surprise, doute, joie, *etc*.) du locuteur. Les mots modaux peuvent être employés pour créer différents types de phrases (interrogatifs, impératifs, *etc*.).

Deux attributs sont définis :

- **Type** :

- *global* : mot modal dont l'influence porte globalement sur toute la phrase. Par exemple : à = *<marqueur d'interrogation> / <marqueur d'exclamation> (suivant le ton du locuteur)*.
- *local* : mot modal dont l'influence porte sur un seul mot. Par exemple : chính dans chính anh ta … =$_{litt}$ *<particule modale> il = c'est lui-même qui …*

- **Meaning** :
  - *opinion* : Par exemple : nhũng = *jusqu'à (dans « il mange jusqu'à cinq bols de riz »)*, mỗi = *un seul*.
  - *strengthening* : Par exemple : thì dans Cô ta thì thông minh =$_{litt}$ *Elle ϕ intelligent = Ça oui, elle est intelligente ! (négatif, ironique)*.
  - *exclamation, interrogation, call* et *imperative* : ces quatre valeurs reflètent quatre types de phrases créées par les mots modaux.

3.3.2.12. Descriptions additionnelles

Locutions

Cette catégorie se compose des locutions et des idiomes. Le dictionnaire papier associe parfois une locution à une partie du discours si l'usage de cette locution lui correspond. Pour les caractériser, nous définissons un attribut :

- **Category** : *noun*, *verb*, *adjective*, *other.*

Éléments de dérivation (Non-Autonome)

Certaines syllabes d'origine chinoise qui servent uniquement à créer des mots composés (*cf.* section 2.3.2), mais ne peuvent pas être autonomes comme une unité lexicale. Pourtant, le fait que les vietnamiens continuent toujours à utiliser ces syllabes pour créer de nouveaux mots complexes nous suggère d'enregistrer telles syllabes avec des descriptions morphosyntaxiques.

Pour l'instant nous ne considérons qu'un seul attribut :

- **Position** : décrit la position de cet élément dans le mot dérivé : devant ou après (*pre, post*).

Résidus

Cette catégorie est utilisée pour marquer les unités qui, soit ne correspondent pas aux catégories ci-dessus, soit ne sont pas assez étudiées pour déterminer leur usage. C'est le cas, par exemple, des unités lexicales auxquelles aucune catégorie n'est attribuée dans le dictionnaire papier étudié.

3.3.3. Processus de la construction du lexique

En nous appuyant sur le Dictionnaire Vietnamien (DV, Hoàng Phê [HOA 02]), nous construisons un lexique de 37 454 entrées décrites grâce au vocabulaire lexical que nous avons défini.

Le DV inclut des termes usuels de la vie quotidienne et des journaux, des termes fréquents en littérature, des termes dialectaux fréquemment utilisés, des termes scientifiques ou techniques dans les documents scientifiques populaires (vulgarisation), des expressions usuelles, des syllabes spéciales seulement utilisées pour la composition des mots (éléments de dérivation), et des abréviations d'usage courant.

Le Centre de Lexicographie du Vietnam (Vietlex) dispose de la version électronique de ce dictionnaire, sous forme des documents Microsoft Word. Le DV contient 39 924 mots d'entrée[29] dont chacun peut avoir plusieurs sens proches ; à chacun de ces sens numérotés sont associés une partie du discours, une note d'usage ou de domaine éventuel, une définition, et des exemples.

Par exemple, le morphème « yêu » correspond à deux entrées dans le dictionnaire, sous le format suivant.

Exemple :

> **yêu**$_1$ d. *(id.)*. Vật tưởng tượng trong cổ tích, thần thoại, hình thù kì dị, chuyên làm hại người.
>
> **yêu**$_2$ **1** đg. Có tình cảm dễ chịu khi tiếp xúc với một đối tượng nào đó, muốn gần gũi và thường sẵn sàng vì đối tượng đó mà hết lòng. *Mẹ yêu con. Yêu nghề. Yêu đời. Trông thật đáng yêu. Yêu nên tốt, ghét nên xấu (tng.).* **2** đg. Có tình cảm thắm thiết dành riêng cho một người khác giới nào đó, muốn chung sống và cùng nhau gắn bó cuộc đời. *Yêu nhau. Người yêu.* **3** đg. Từ dùng sau một động từ trong những tổ hợp tả một hành vi về hình thức là chê trách, đánh mắng một cách nhẹ nhàng, nhưng thật ra là biểu thị tình cảm thương yêu. *Mẹ mắng yêu con. Nguýt yêu. Tát yêu.*

La première entrée correspond au mot portant le sens « *diable* », alors que la deuxième entrée correspond aux différents sens du mot « *aimer* ». Les phrases en italique sont des exemples. Les notes d'usage se trouvent entre parenthèses. Les parties du discours enregistrées appartiennent aux huit catégories présentées à la section 2.4.1.

À partir de ce dictionnaire, nous construisons le lexique en deux étapes :

- Conversion des données en format XML, avec le codage présenté à la section 3.3.4.4. La conversion automatique est réalisée par le Vietlex.
- Entrée manuelle des descriptions lexicales avec l'aide d'une interface qui projette les 8 catégories d'origine sur les 11 catégories VN-Multext (*cf.* section 3.3.2), et qui affiche le contenu des entrées pour faciliter le choix des valeurs des attributs du schéma des descriptions. Les entrées représentées par une même forme (les homonymes) sont regroupées en une seule entrée.

Nous obtenons finalement un lexique de 37 454 unités lexicales enrichies par des descriptions lexicales présentées à la section précédente. Ce lexique couvre 18 732 noms, 975 numéraux et déterminants, 12 209 verbes, 8 442 adjectifs, 543 adverbes, 149 conjonctions, 120 pronoms, 106 interjections, 90 particules modaux, 38 préposition, et environ 1 000 locutions et éléments non autonomes. Le lexique devrait être graduellement enrichi par de nouveaux mots apparus dans les corpus traités.

La section suivante précise le codage adopté pour la maintenance de ces ressources.

3.3.4. Codage de ressources lexicales

Toujours dans une perspective de normalisation, nous choisissons de suivre les recommandations internationales pour coder nos ressources lexicales. Dans cette section, nous présentons d'abord les schémas normalisés de balisage des ressources lexicales TEI (*cf.* 1.2.1.2) et LMF (ISO/TC 37/SC 4, *cf.* section 1.2.2), puis le format de nos données lexicales.

[29] La différence avec le nombre d'entrées du lexique est due au fait que celui-ci rassemble comme une même entrée les homographes distingués par le dictionnaire papier.

3.3.4.1. Modèle pour les bases lexicales

Ide *et al.* [VER 92, IDE 93] font l'analyse du contenu des entrées lexicales dans les dictionnaires papiers. À partir de ces analyses, ils avancent que les modèles proposés précédemment pour représenter les bases lexicales (modèles textuels, modèles relationnels) ne sont pas assez puissants. Ils proposent donc un nouveau modèle basé sur les structures de traits[30] qui sont largement utilisées pour la représentation des informations linguistiques dans les grammaires d'unification (*cf.* 4.2.1.2). Précisément, le modèle mis au point fait usage de structures de traits typées. Par exemple, chaque entrée du dictionnaire est décrite par une structure de traits de type ENTRY dont les traits admissibles sont *form* (les informations concernant la forme de l'entrée lexicale), *gram* (les informations grammaticales), *usage* (les informations d'usage), *def* (les définitions de cette entrée), *etc*. Le domaine des valeurs du trait *form* est constitué de structures de type FORM, acceptant les traits *orth* (pour l'orthographe), *hyph* (pour la césure – *hyphenation* en anglais), *pron* (pour la prononciation), *etc*. Ces traits, à leur tour, acceptent les valeurs atomiques, de type STRING (chaîne de caractères) par exemple.

Dans ce modèle, un type unique de structure de traits est utilisé pour les entrées lexicales, les homographes et les sens. Cela reflète le fait que les données lexicales à ces différents niveaux contiennent les mêmes informations.

Pour représenter des variants dans les entrées lexicales, le modèle fait l'usage de l'opérateur de disjonction de valeurs de trait en ajoutant la spécification « liste » (notée $(x_1, ..., x_n)$) ou « ensemble » (notée $\{x_1, ..., x_n\}$) des valeurs disjonctives (la notation « liste » permet de préserver l'ordre dans les cas où il a un sens, par exemple la première valeur de forme orthographique dans une liste alternative est la forme la plus courante). La notion de disjonction générale (Kay [KAY 85]) spécifiant les sous-parties alternative d'une structure de traits est également étendue avec ces notions de « liste » et « ensemble ». Par ailleurs, la forme des structures de traits est restreinte à la forme normale hiérarchique (dans une structure de traits quelconque, un seul trait peut être une disjonction).

Ce modèle, doté de plusieurs mécanismes de gestion des informations disjonctives et enchâssées (opérateurs *factor* et *unfactor*, mécanisme de surcharge), résout presque tous les problèmes inhérents aux modèles classiques d'une part, et d'autre part permet d'accéder, de manipuler et de fusionner des informations lexicales structurées différemment d'un dictionnaire à l'autre.

Cette représentation de dictionnaires papiers a été partiellement instanciée dans le cadre du projet de normalisation TEI (*cf.* 1.2.1.3).

Ide *et al.* [IDE 00a] ont fait évoluer le modèle ci-dessus à un niveau plus abstrait. Ils introduisent un modèle formel reflétant la hiérarchie d'information d'une entrée lexicale dans un dictionnaire traditionnel ou opérationnel. Divers mécanismes sont définis pour la propagation et l'expression de dépendances entre les traits attachés aux nœuds dans l'hiérarchie d'information. Les auteurs exposent également le codage de ces informations en XML et la possibilité d'extraire et de manipuler ces informations dans un format quelconque grâce au langage de transformation XSL.

Dans la section suivante, nous présentons le codage défini par la TEI pour représenter les dictionnaires papiers.

3.3.4.2. TEI (*Text Encoding Initiative*) et codage de dictionnaires papiers

Un groupe de travail de la TEI est spécialisé dans le codage des dictionnaires papiers (*cf.* Ide et Véronis [IDE 95b], Sperberg-McQueen et Burnard [SPE 94 – chapitre 12]), c'est-à-dire les dictionnaires orientés vers un usage humain, qui peuvent être structurellement très complexes selon le but de leur utilisation. L'objectif est de définir une DTD permettant de valider le codage du contenu des dictionnaires, en SGML pour la première version TEI P3, XML pour la version TEI P4. Deux principes sont mis en avant :

---

[30] Les notions et les opérateurs concernant les structures de traits sont présentés au Chapitre 4, grammaires d'unification

- En premier lieu, puisque la structure des entrées de dictionnaire change considérablement dans un même dictionnaire, et plus encore entre différents dictionnaires, la manière la plus simple pour qu'un schéma de codage soit adapté à la gamme entière des structures rencontrées en pratique est de permettre virtuellement une position libre pour tout élément dans chaque entrée de dictionnaire. Cependant, il existe clairement des principes structuraux assez consistants qui régissent la grande majorité des dictionnaires conventionnels et même la plupart des entrées dans les dictionnaires « exotiques ». Ces principes sont capturés par les directives de codage de la TEI avec la définition de l'élément **<entry>** pour les entrées de dictionnaire. Un deuxième élément **<entryFree>** est défini pour la même structure, mais cette définition permet un ordre beaucoup plus libre de ses composants.
- En second lieu, puisque une grande partie de l'information contenue dans les dictionnaires papiers est implicite ou fortement contractée, la question se pose de savoir si l'encodage doit capturer la forme typographique précise du texte source ou la structure fondamentale de l'information que le texte présente. Les utilisateurs intéressés principalement dans le format imprimé du dictionnaire exigeront d'un codage d'être fidèle à une version imprimée originale, alors que d'autres porteront leur attention sur les possibilités d'extraction du dictionnaire des informations lexicales sous une forme appropriée à un traitement ultérieur, ce qui peut exiger l'expansion ou la remise en ordre des informations contenues dans la forme imprimée. De plus, quelques utilisateurs souhaitent coder les deux types de données, et maintenir les liens entre les éléments relatifs des deux codages. La TEI développe donc des méthodes permettant d'enregistrer ces deux types de données, ainsi que la corrélation entre eux.

L'ensemble d'éléments définis pour l'encodage des dictionnaires de la TEI est très riche et dynamique, et prend en compte les deux principes ci-dessus. Les détails de ces éléments sont documentés sur le site d'Internet de la TEI[31]. Les utilisateurs peuvent ainsi personnaliser la DTD pour obtenir une DTD plus simple et adéquate à la représentation des informations enregistrées dans leur dictionnaire.

Nous appliquons ce schéma pour le DV disponible au Vietlex (voir la section 3.3.4.4). Quant au codage du lexique morphosyntaxique construit, nous prenons en compte les activités de normalisation de représentation des lexiques opérationnels dans le cadre de l'ISO/TC 37/SC 4 (*cf.* 1.2.2). Il s'agit du modèle LMF introduit ci-dessous.

3.3.4.3. LMF (*Lexical Mark-up Framework*)

LMF [ISO 05b] est un méta-modèle abstrait qui fournit une plate-forme pour la construction des lexiques opérationnels pour le TAL. L'objectif est de définir une norme de représentation générique des données lexicales dans les contextes de gestion et d'échange de lexiques.

L'approche du LMF pour la description de la micro-structure des lexiques est d'attacher systématiquement le comportement syntaxique à la description sémantique du mot (*cf.* Romary *et al.* [ROM 04]). Cela est en particulier cohérent avec les principes linguistiques exposés par Saussure, qui considère qu'un mot est décrit par une paire signifiant/signifié, correspondant à une description morphologique/sémantique.

Le modèle LMF se compose d'une partie noyau et des extensions lexicales correspondant aux informations relevant de la morphologie, de la syntaxe, de la sémantique et de l'interlinguistique (*cf.* Figure 3-2). Conformément aux principes généraux de l'ISO/TC 37/SC 4 (*cf.* 1.2.2), ces informations sont décrites par le biais de descripteurs élémentaires, c'est-à-dire des catégories de données, qui sont pour leur part définies dans le DCR central du TC 37. Le processus de composition d'un lexique conforme au LMF est montré à la Figure 3-3.

[31] Pour la version P4 : http://www.tei-c.org/P4X/DI.html

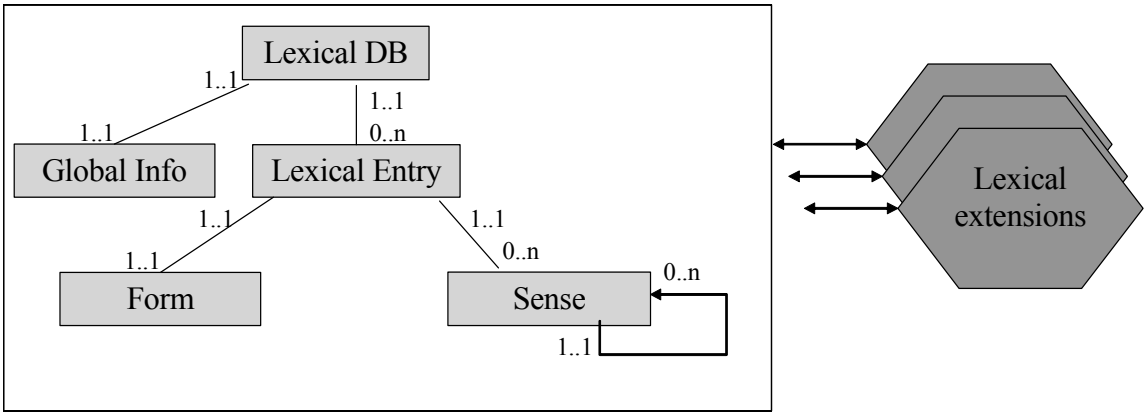


Figure 3-2 LMF – principe du modèle [ROM 04]

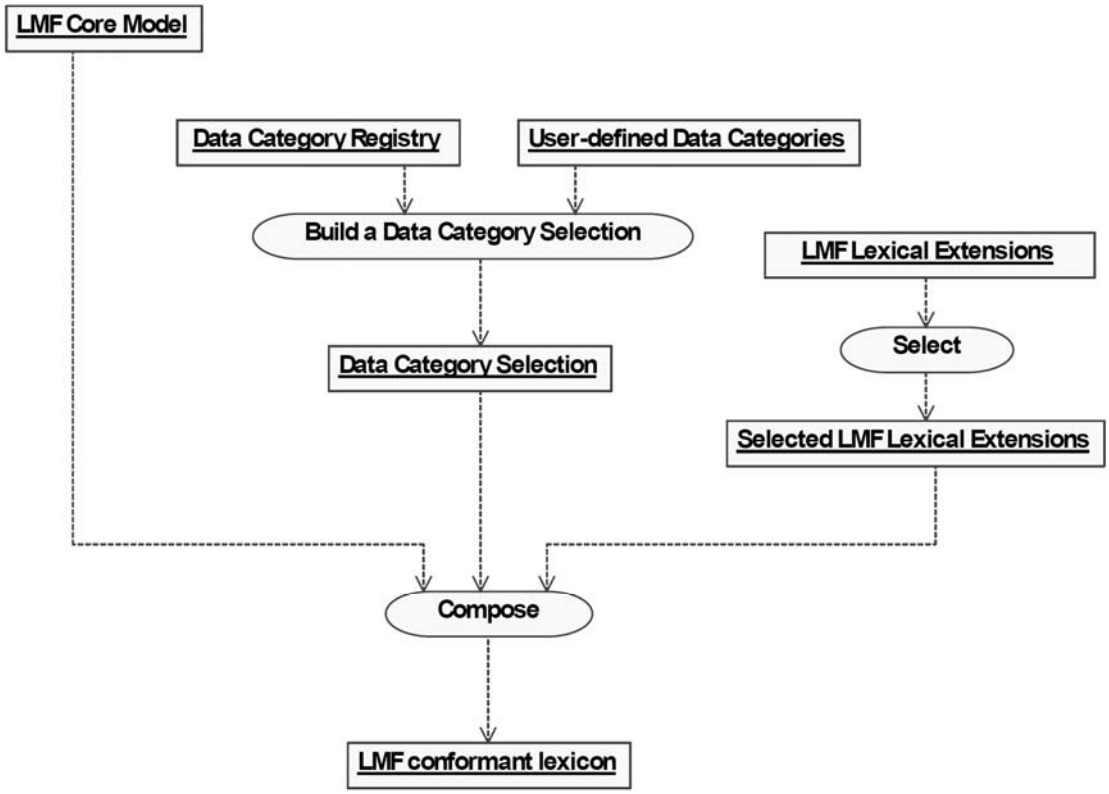


Figure 3-3 Processus d'utilisation de LMF ([ISO 05b])

Considérons par exemple une extension lexicale : la morphologie. Les Figure 3-4 et Figure 3-5 montrent un modèle de lexiques dont les informations associées à chaque entrée lexicale comprennent :

- des informations noyau : forme (descriptions graphique et phonétique), sens (qui peut être à la fois répété ou divisé en plusieurs sens) ;
- des informations étendues de morphologie : paradigme, inflexions.

La Figure 3-6 exemplifie une instanciation concrète (dans le cadre du projet Morphalou [ROM 04]) de ce modèle : l'entrée du mot « *chat* » en français, codée sous format GMT[32] (*Generic Mapping Tool*) avec un schéma compatible au modèle LMF.

[32] GMT est un format intermédiaire défini dans le cadre de la la norme ISO 16642:2003 (TMF – *Terminological Markup Framework*) dans le but de mettre en correspondance les différents codages terminologiques

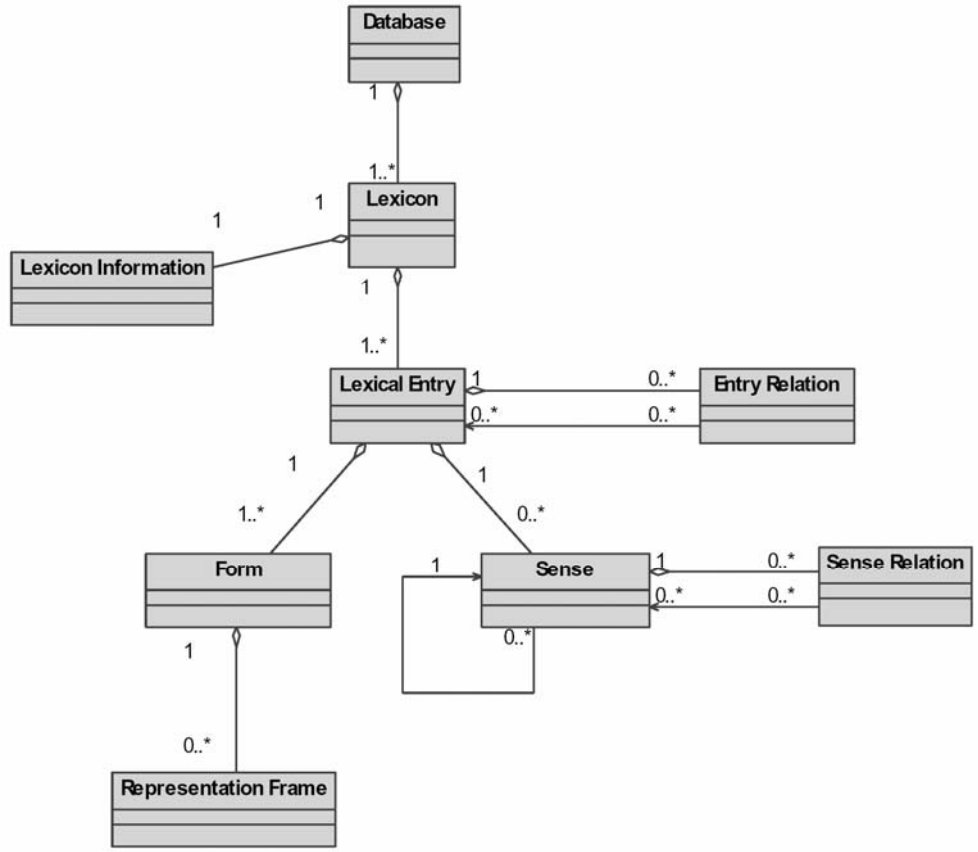


Figure 3-4 LMF – Modèle noyau [ISO 05b]

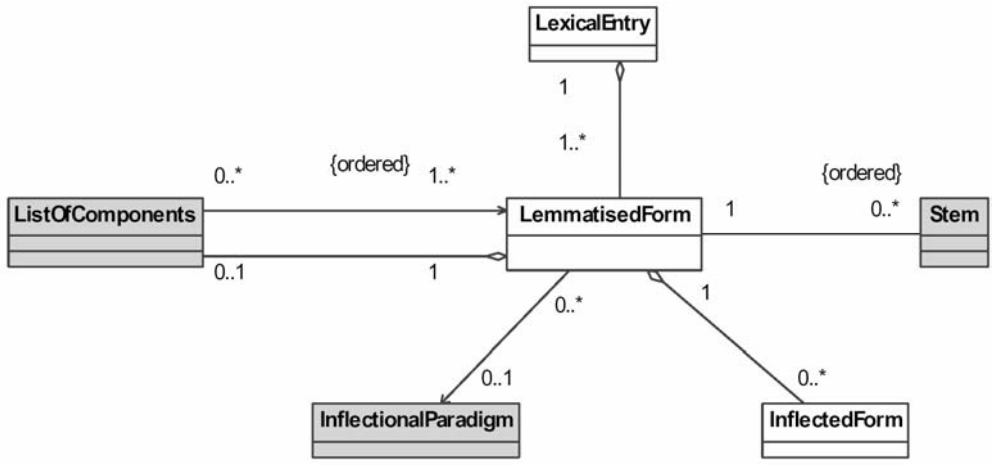


Figure 3-5 LMF - Extensions lexicales pour la morphologie [ISO 05b]

```
<struct type='lexical entry'>
       <feat type='lemma'>chat</feat>
       <feat type='grammatical category'>noun</feat>
       <feat type='gender'>masculine</feat>
       …
       <struct type='morphology'>
              <struct type='paradigm'>
                     <feat type='paradigm identifier'>fr-s-plural</feat>
              </struct>
              <struct type='inflection'>
                     <feat type='orthography'>chat</feat>
                     <feat type='number'>singular</feat>
              </struct>
              <struct type='inflection'>
                     <feat type='orthography'>chats</feat>
                     <feat type='number'>plural</feat>
              </struct>
              …
       </struct>
</struct>
```

Figure 3-6 Codage (GMT) de l'entrée « chat » avec un schéma compatible au LMF [ROM 04]

Nous passons maintenant à l'application des schémas normalisés de représentation des bases lexicales (la TEI pour les dictionnaires papiers, LMF pour les lexiques opérationnels) sur nos ressources lexicales. La section 3.3.4.4 porte sur le codage du DV du Vietlex, et la section 3.3.4.5 discute de la représentation du lexique morphosyntaxique créé.

3.3.4.4. Codage du dictionnaire papier vietnamien du Vietlex

En reprenant les éléments proposés dans le schéma de codage des dictionnaires de la TEI, nous définissons une DTD personnalisée pour encoder les informations enregistrées dans le dictionnaire vietnamien du centre Vietlex. Cette DTD se trouve à l'annexe B. Les informations de chaque entrée sont extraites automatiquement à partir de la forme typographique du dictionnaire. Comme nous nous intéressons actuellement surtout à l'orthographe des mots et à la catégorie grammaticale correspondant à chaque sens d'un mot, notre schéma de marquage reste très simple. Le codage des éléments comme par exemple des exemples d'usage mérite d'être beaucoup plus détaillé ultérieurement.

Voici un exemple d'illustration, recourant à l'exemple à la section 3.3.3.

```
<superEntry n="…">
        <entry n="1">
                <form><orth>yêu</orth></form>
                <sense n="…"> <!-- diable -->
                        <gramGrp><pos>d.</pos></gramGrp>
                        <usg type="style">(id.).</usg>
                        <def>Vật tưởng tượng trong cổ tích, thần thoại, hình thù kì dị, chuyên làm hại người.</def>
                </sense>
        </entry>
        <entry n="2">
                <form><orth>yêu</orth></form>
                <sense n="1"> <!-- aimer (amour général)-->
                        <gramGrp><pos>đg.</pos></gramGrp>
                        <def>Có tình cảm dễ chịu khi tiếp xúc với một đối tượng nào đó, muốn gần gũi và thường sẵn sàng vì đối tượng đó mà hết lòng.</def>
                        <eg>Mẹ yêu con. Yêu nghề. Yêu đời. Trông thật đáng yêu. Yêu nên tốt, ghét nên xấu (tng.).</eg>
                </sense>
                <sense n="2"> <!-- aimer (amour romantique) -->
                        <gramGrp><pos>đg.</pos></gramGrp>
                        <def>Có tình cảm thắm thiết dành riêng cho một người khác giới nào đó, muốn chung sống và cùng nhau gắn bó cuộc đời.</def>
                        <eg>Yêu nhau. Người yêu.</eg>
                </sense>
                <sense n="3"> <!-- aimer – modifieur d'un autre verbe pour exprimer une action tendre, "pas sérieuse" -->
                        <gramGrp><pos>đg.</pos></gramGrp>
                        <def>Từ dùng sau một động từ trong những tổ hợp tả một hành vi về hình thức là chê trách, đánh mắng một cách nhẹ nhàng, nhưng thật ra là biểu thị tình cảm thương yêu.</def>
                        <eg>Mẹ mắng yêu con. Nguýt yêu. Tát yêu.</eg>
                </sense>
        </entry>
  </superEntry>
```

Une fois le DV disponible sous format XML, nous avons développé à l'attention des linguistes du Vietlex une interface permettant d'éditer les descriptions lexicales (comme présentées à 3.3.2) de chacune de ses entrées, en récupérant les parties du discours de base enregistrés dans le DV. Les autres informations de chaque entrée sont visualisées pour aider les linguistes à choisir la valeur de chaque attribut des descripteurs lexicaux. Le lexique morphosyntaxique ainsi construit est soumis au codage présenté dans la sous-section suivante.

3.3.4.5. Codage du lexique morphosyntaxique vietnamien

Le dictionnaire XML obtenu ci-dessus est une base pour la création de notre lexique morphosyntaxique. Nous avons développé une interface simple pour que les lexicographes puissent visualiser les informations concernant chaque entrée lexicale et saisir les valeurs des descripteurs lexicaux définis à la section 3.3.2.

Outre le format compact texte (*cf.* MULTEXT à la section 3.3.1), notre système (*cf.* annexe D) gère le lexique sous un format XML, qui explicite les descriptions lexicales, pour que les ressources lexicales soient d'une part faciles à manipuler et modifier, et d'autre part accessibles à tous.

Revenons à l'exemple de la section précédente : pour un morphème « yêu » il existe deux descripteurs correspondant à deux entrées dans le dictionnaire papier (tous les trois sens de la deuxième entrée ont le même descripteur). Son codage XML est reproduit à la Figure 3-7.

```
<struct type='lexical entry'>
    <feat type='form'>yêu</feat>
    <struct type='grammatical description group'>
          <struct type='grammatical description'> <!--diable-->
                <feat type='pos'>Noun</feat>
                <struct type='subcategory description'>
                      <feat type='type'>common</feat>
                      <feat type='countability'>partial</feat>
                      <feat type='meaning'>abstract</feat>
                </struct>
          </struct>
          <struct type='grammatical description'> <!--aimer-->
                <feat type='pos'>Verb</feat>
                <struct type='subcategory description'>
                      <feat type='grade'>gradable</feat>
                      <feat type='meaning'>feelings</feat>
                </struct>
          </struct>
    </struct>
</struct>
```

Figure 3-7 Codage explicite en XML d'une entrée du lexique morphosyntaxique vietnamien

Nous discutons maintenant de la représentation de notre lexique par un modèle conforme au LMF (*cf.* 3.3.4.3), qui vise à l'interopérabilité des lexiques pour le TAL.

Les entrées lexicales du modèle LMF sont généralement organisées autour d'un lemme et de sa partie du discours. Afin de nous conformer au LMF, nous devons convertir l'entrée « yêu » ci-dessus, qui correspond à une « super-entrée » du dictionnaire papier, en deux entrées lexicales.

Du fait de la nature isolante de la langue vietnamienne, il n'est pas pertinent pour nous d'adopter l'extension de morphologie. Les autres informations que nous avons enregistrées dans les descripteurs lexicaux portent non seulement sur les propriétés syntaxiques mais aussi des propriétés sémantiques. Ainsi les informations sémantiques doivent être regroupées dans le composant « Sense ». Quant aux informations syntaxiques, elles doivent être regroupées dans le composant d'extension syntaxique qui appartient également au composant « Sense ». Autrement dit, toutes les informations du groupe « subcategory description » sont attachées au composant « Sense ».

En résumé, il est tout à fait possible de convertir notre lexique vers une représentation se conformant au modèle LMF.

Nous revenons à la représentation des informations syntaxiques du lexique à la section 4.5.1, au cours de l'étude de la construction du lexique syntaxique du vietnamien.

3.4. Annotation morphosyntaxique de textes vietnamiens

Nous commençons cette section par présenter la phase préliminaire au travail d'étiquetage à proprement parler : définition des unités lexicales et des étiquettes (section 3.4.1), puis sélection et préparation du corpus (section 3.4.2). Nous décrivons ensuite le travail réalisé pour chaque étape de l'étiquetage morphosyntaxique (*cf.* 3.2) : segmentation (section 3.4.3), étiquetage *a priori* (section 3.4.4), et désambiguïsation (section 3.4.5). Nous finissons par une évaluation de notre système (section 3.4.5.2).

3.4.1. Définition des jeux d'étiquettes

Les catégories grammaticales reflètent des oppositions diverses dans le système syntaxique. Le principal critère pour la définition du jeu d'étiquettes est donc la distribution syntaxique. Il est *a priori* nécessaire de disposer d'un important jeu d'étiquettes pour refléter exactement toutes les relations syntaxiques. Cependant, plus le jeu d'étiquettes est important, plus la tâche d'annotation est difficile. Ainsi, nous avons besoin d'un compromis pour parvenir à un jeu d'étiquettes assez précis et de taille acceptable.

Nous choisissons de considérer des jeux d'étiquettes à deux niveaux de granularité : l'un correspond aux catégories décrites à la couche noyau du modèle de descriptions lexicales (*cf.* 3.3.2) qui sont assez claires et universelles, et l'autre correspond à une projection des descriptions de la couche privée.

3.4.2. Gestion des corpus annotés

L'un des objectifs de notre travail est de construire un corpus annoté de référence pour les tâches d'analyse automatique des textes vietnamiens. La gestion des corpus sous un format normalisé est donc souhaitable. Nous suivons donc la discussion menée sur la normalisation de la représentation de l'annotation morphosyntaxique dans le cadre du projet MAF de l'ISO / TC 37 / SC 4 (*cf.* 1.2.2).

Dans cette section, nous traitons tout d'abord de la collection et du codage structuré des corpus (sous-section 3.4.2.1), puis du processus de construction des corpus étiquetés ainsi que de leur codage (sous-section 3.4.2.2).

3.4.2.1. Collection et codage structurel des corpus

Collection des textes

Visant un travail ultérieur sur l'alignement multilingue de textes français-vietnamiens et anglais-vietnamiens, nous avons choisi des corpus pour lesquels une version correspondante en français et/ou en anglais est disponible.

Nous avons ainsi pu obtenir un corpus bilingue français-vietnamien contenant des textes suivants :

- Nouvelles françaises traduites en vietnamien et réciproquement, récupérées sur Internet ou acquises à partir de livres par numérisation et reconnaissance de caractères, contenant environ 140 000 syllabes vietnamiennes et 130 000 mots français.
- Textes de droit vietnamien et traduction en français, offerts par la Maison du Droit Vietnamo-Français[33], contenant environ 384 000 syllabes vietnamiennes et 296 000 mots français.

[33] http://www.maisondudroit.org

- Textes d'économie en français et traduction en vietnamien, et réciproquement, dans le cadre du Forum Franco-Vietnamien Économique et Financier, offert par l'ADETEF[34]-Vietnam, contenant environ 150 000 syllabes vietnamiennes et 100 000 mots français.
- Nouveau Testament en vietnamien (d'environ 198 000 syllabes) et en français (d'environ 176 000 mots), distribués par Philip Resnik[35].

Codage structuré des textes

Les textes collectés sont codés en XML, en adoptant les directives de la TEI (http://www.tei-c.org/P4X/index.html). Nous balisons principalement trois éléments :

- : les divisions, correspondant aux parties de texte comme chapitre, section, *etc*.
- <p> : les paragraphes de texte
- <seg> : les segments, correspondant généralement aux phrases dans un texte.

Le codage des caractères est l'Unicode (UTF-8). Cependant, la gestion du codage de caractère est externe, car dans les fichiers XML tous les caractères « spéciaux » vietnamiens sont représentés par des entités nommées.

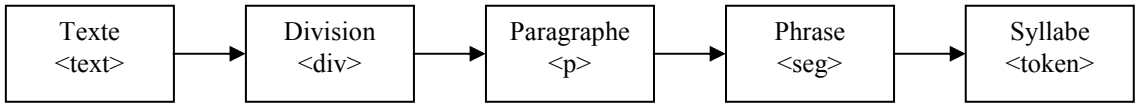
3.4.2.2. Codage de corpus étiquetés

Partant d'un principe de séparation entre les annotations assez stables (comme par exemple de la structure logique des documents) et les annotations dépendantes des applications (comme par exemple des mots et des étiquettes morphosyntaxiques), nous choisissons une annotation externe (« *stand-off* », *cf*. Bonhomme [BON 00a]) pour l'annotation morphosyntaxique des corpus. Le principe de l'annotation externe consiste à placer les informations ajoutées dans un fichier séparé faisant référence au fichier d'origine sans reproduire la totalité des informations (structure, contenu, *etc.*) de celui-ci. Le fichier d'origine ne subit pour sa part aucune modification.

Les informations fournies par une annotation morphosyntaxique sont généralement : la composition des mots annotés, leur partie du discours, et d'autres caractéristiques morphosyntaxiques.

Une annotation externe a besoin de points de référence pour identifier les mots dans le texte primaire. On peut, par exemple, prendre la position du premier caractère d'un mot dans la chaîne pour repérer ce mot. Pour le vietnamien, qui est une langue monosyllabique (*cf*. 2.3.1), le découpage automatique d'un texte en syllabes en se basant sur les espaces est fiable, et les occurrences de mots peuvent être identifiées grâce à un indice de position en nombre de syllabes. Nous choisissons donc le découpage en syllabes comme une segmentation de référence pour l'annotation morphosyntaxique externe.

En résumé, la base de corpus annoté se compose d'un corpus primaire et d'un corpus d'annotation externe. Chaque texte de notre corpus primaire est segmenté selon la hiérarchie suivante :



Le corpus d'annotation externe contient des annotations linguistiques qui font référence aux identificateurs des <*token*> dans le corpus primaire.

[34] Association française pour le Développement des Échanges en Technologies Économiques et Financières

[35] http://www.umiacs.umd.edu/~resnik/parallel/bible.html

En ce qui concerne le schéma d'annotation, le projet MAF (*Morphosyntactic Annotation Framework*) de l'ISO/TC 37/SC 4 (*cf.* 1.2.2) a pour but de définir un modèle générique dédié à l'annotation morphosyntaxique (future norme ISO 24611, *cf.* [ISO 05a]). Ce modèle combine, d'une part, deux niveaux de segmentation et de catégorisation linguistique (étiquetage morphosyntaxique), et d'autre part, un ensemble de catégories de données linguistiques (*cf.* DCR à la section 1.2.2) permettant l'échange et l'interaction de données. Selon ce principe, les informations linguistiques (comme les parties du discours, les traits morphologiques, *etc*.) de chaque annotation conforme au MAF doivent pouvoir être mises en correspondance avec les catégories de données définies dans le DCR.

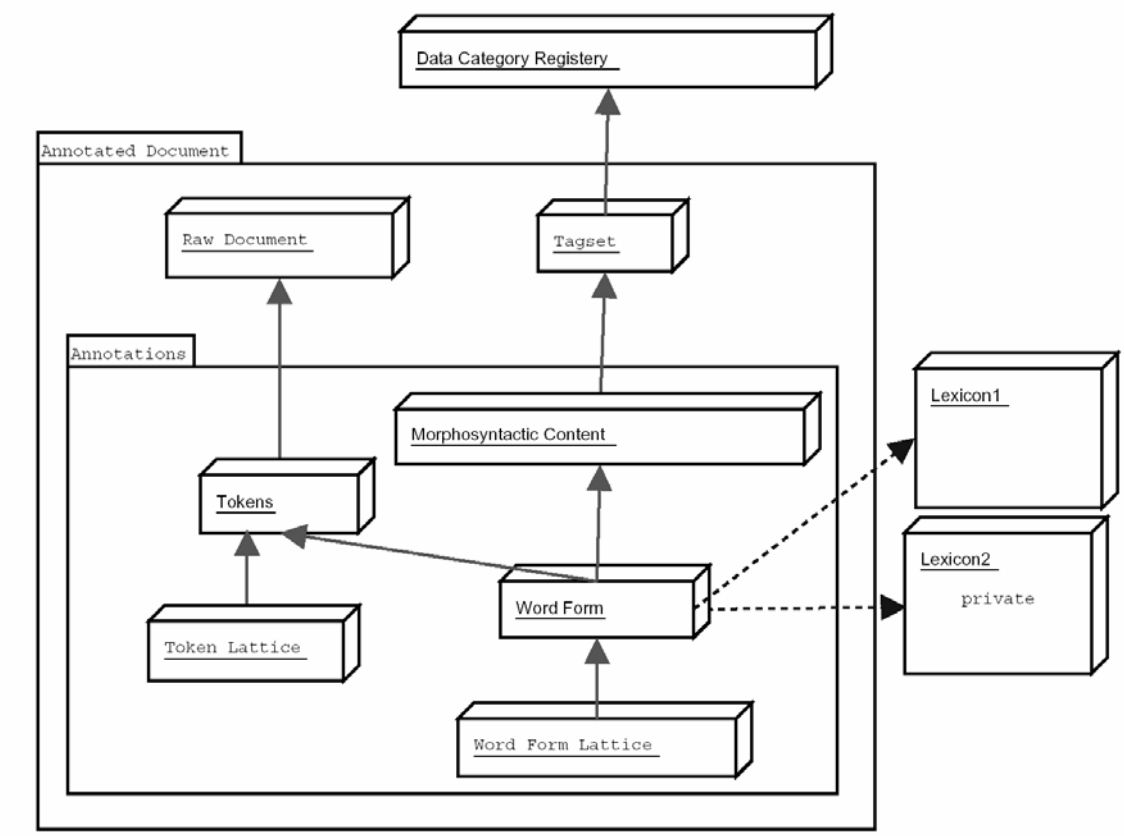


Figure 3-8 Vue simplifiée du méta-modèle MAF [ISO 05a]

La Figure 3-8 présente une vue simplifiée du méta-modèle MAF. Chaque objet *wordForm* contient les informations linguistiques sur le groupe de tokens reconnus comme une unité lexicale du texte annoté. Outre l'étiquette morphosyntaxique attachée à cette unité, il est également possible de la mettre en lien avec son entrée dans un lexique de référence.

Nous appliquons un schéma de codage correspondant effectivement à ce modèle du MAF. La représentation des textes annotés est rendue relativement simple par l'absence de variation morphologique du vietnamien. Voici un exemple simplifié du codage d'une phrase segmentée et étiquetée avec le jeu d'étiquettes de granularité grossière.

Texte d'entrée :

Ông già đi rất nhanh.

Le texte est d'abord segmenté en unités élémentaires de référence. Ces unités correspondent à des syllabes du vietnamien ou des ponctuations. Cela est représenté comme suit :

```
<seg id = "n">
    <token id = "t1">Ông</token>
    <token id = "t2">già</token>
    <token id = "t3">đi</token>
```

```
    <token id = ″t4″>rất</token>
    <token id = ″t5″>nhanh</token>
    <token id=″t6″>.</token>
</seg>
```

Une fois segmenté en mots et étiqueté, nous pouvons obtenir la représentation suivante correspondant à deux solutions de segmentation possibles :

```
<seg id = ″texte_src@n″>
    <alt> <!-- Solution 1 -->
        <wordForm entry = ″Ông già″ tokens = ″t1 t2″ tag = ″pos@N″ />
                                                <!– homme vieux -->
        <wordForm entry = “đi” tokens = “t3” tag = “pos@V” />
                                                <!– marcher / mourrir -->
        <wordForm entry = “rất” tokens = “t4” tag = “pos@J” />
                                                <!-- très -->
        <wordForm entry = “nhanh” tokens = “t5” tag = “pos@A” />
                                                <!-- vite -->
        <wordForm entry = “.” tokens = “t6” tag = “pos@dot” />
        <!–Le vieux marche/meurt très vite -->
    </alt>
    <alt> <!-- Solution 2 -->
        <wordForm entry = ″Ông″ tokens = ″t1″ tag = ″pos@P″ />
                                                <!-- Vous (homme) -->
        <wordForm entry = ″già″ tokens = ″t2″ tag = ″pos@A″ />
                                                <!-- vieux -->
        <wordForm entry = “đi” tokens = “t3” tag = “pos@J” />
                                                <!-- aller -->
        <wordForm entry = “rất” tokens = “t4” tag = “pos@J” />
                                                <!-- très -->
        <wordForm entry = “nhanh” tokens = “t5” tag = “pos@A” />
                                                <!-- vite -->
        <wordForm entry = “.” tokens = “t6” tag = “pos@dot” />
        <!–Vous vieillissez très vite -->
    </alt>
</seg>
```

Les sections suivantes discutent des tâches d'annotation à proprement parler : segmentation du texte en unité lexicale (3.4.3), étiquetage *a priori* (3.4.4), et désambiguïsation des étiquettes (3.4.5).

3.4.3. Segmentation

Nous présentons d'abord les problèmes spécifiques du vietnamien pour la tâche de segmentation, puis les approches possibles pour résoudre ces problèmes.

3.4.3.1. Problèmes et état de l'art

La tâche de segmentation d'un texte consiste à identifier dans le texte les différents segments (unité lexicale) : les ponctuations, les symboles, les nombres, les dates, les noms propres, les mots, *etc*.

Le système de ponctuations du vietnamien est tout à fait similaire à ceux de l'anglais ou du français, on observe donc les mêmes problèmes que pour ces langues à propos des ponctuations (comme par exemple le problème du point après une abréviation). Les numéros et les dates en chiffre suivent le format et l'ordre français. En revanche, en ce qui concerne les noms propres et les mots il existe plusieurs problèmes spécifiques au vietnamien :

1. Le problème principal vient des **mots complexes** dont la définition est donnée à la section 2.3.2. Comme d'autres langues asiatiques isolantes, le vietnamien est monosyllabique, mais les mots complexes de plusieurs syllabes sont fréquents (*cf.* 3.2.2.2). Or, les syllabes vietnamiennes se sont séparées par des espaces, cela ne permet pas l'identification des mots complexes en s'appuyant sur les blancs dans un texte.
2. Les **noms propres** rendent plus ambiguë la segmentation à cause de deux phénomènes suivants :
   2a. Le cas « exceptionnel » en français des noms propres pouvant aussi être interprétés comme noms communs constitue la majorité des cas en vietnamien, où les noms de personnes ou de lieu ont généralement un sens dans la langue. Seule la majuscule initiale permet alors de distinguer les noms propres, ce qui pose naturellement problème en début de phrase.
   2b. Les formes des noms propres des organisations ou des titres ne sont pas unifiées. Par exemple, le « *Département d'Informatique* » peut être écrit en vietnamien « Khoa Tin học », « Khoa tin học », « Khoa Tin Học » ou encore « khoa Tin học »[36].
3. Enfin, les **formes redoublées**[37] complexes (*cf.* 2.4.2.1) constituent une source de difficulté importante : comme elles ne sont pas recensées dans les dictionnaires, leur détection doit se faire « à la volée », mais les règles qui régissent leur construction sont relativement floues et font appel à des critères phonétiques d'accès non trivial à partir du seul texte.

À l'heure actuelle, nous n'avons connaissance que de deux travaux existants sur la segmentation du vietnamien : Dinh *et al.* [DIN 01] et Ha A. L. [HAL 03].

[DIN 01] présente une approche de segmentation en se basant sur le modèle WFST (*Weighted Finite State Transducer*) combiné à un modèle neuronal. Les ressources utilisées sont un corpus segmenté de deux millions de mots, et un lexique d'environ 34 000 entrées extraites du dictionnaire de Hoàng Phê (l'édition 2000). Les auteurs avancent un taux de précision de la segmentation de 97% sur leur corpus d'évaluation. Malheureusement, l'article manque grandement de précision concernant le travail mis en oeuvre, et notamment l'ensemble de parties du discours choisi pour alimenter les ressources : en effet, les auteurs annoncent qu'il s'agit des catégories morphosyntaxiques présentées par le dictionnaire de Hoàng Phê, mais celui-ci n'en mentionne que 8, alors que les auteurs font usage de 9 catégories. La précision de 97% n'a pas vraiment de sens, car cela dépend fortement de la définition des segments considérés (par exemple, [DIN 01] considère comme mot le groupe nominal « sự cố gắng » =$_{litt}$ « (*le*) *fait s'efforcer* » = « *l'effort* », ce qui est pour nous deux unités lexicales). L'outil de segmentation et les ressources ne faisant pas l'objet d'une distribution publique, nous n'avons pu le tester pour le comparer précisément à la technique que nous proposons.

[HAL 03] propose une segmentation en utilisant la méthode d'apprentissage non supervisé avec l'algorithme MLE (*Maximum Likelihood Estimation*). La notion de l'unité « mot » n'est pas définie mais seulement intuitive. Le travail n'est encore dans son état actuel qu'à un niveau très expérimental.

3.4.3.2. Solutions proposées

Nous considérons d'abord le problème simplifié de segmentation, où la phrase à segmenter ne contient que des mots du dictionnaire.

---

[36] khoa = *département*, tin học = *informatique*

[37] On distingue un mot redoublé avec une forme redoublée. Les mots redoublés sont généralement listés dans le lexique, alors que les formes redoublées ne le sont pas.

Nous avons adopté les automates d'états finis (*cf.* Figure 3-9) pour identifier des segmentations possibles pour chaque phrase (délimitée par des ponctuations). Nous retenons comme « meilleure » segmentation le chemin le plus court dans le graphe[38]. En cas d'ambiguïté (plusieurs chemins de même longueur, *cf.* Figure 3-10), il est nécessaire de faire appel au jugement humain ou, si une automatisation totale est souhaitée, de définir une heuristique de choix : il est par exemple possible de retenir la segmentation faisant apparaître les mots composés les plus longs en premier ou, au contraire, en dernier, ou même de choisir aléatoirement parmi les solutions possibles – ces trois approches fournissant des résultats qualitativement très proches. La seule heuristique réellement efficace consisterait à prendre en compte les fréquences d'apparition des mots pour choisir la séquence la plus probable, mais cela nécessiterait de faire appel à un apprentissage supervisé, ce que nous ne souhaitons pas faire à cette étape.

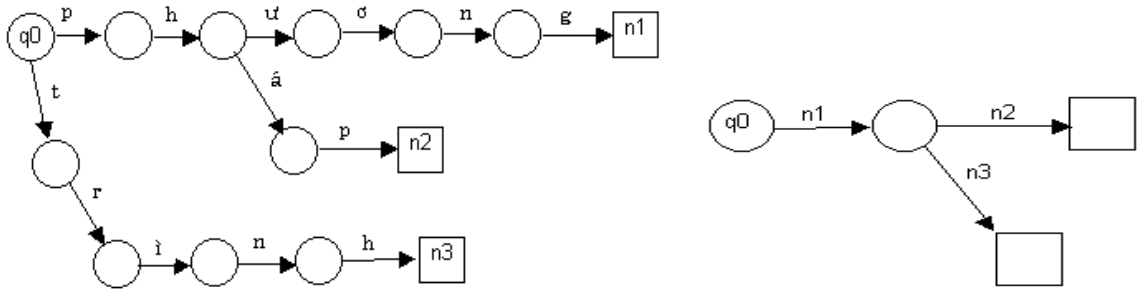


Figure 3-9 Automates acceptant les syllabes et les unités lexicales

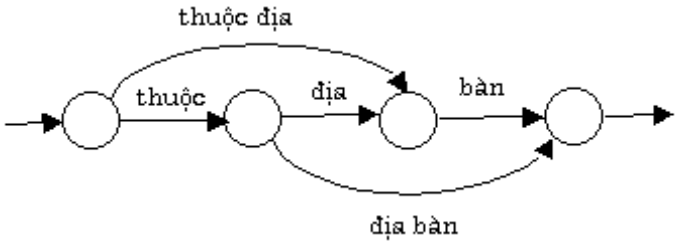


Figure 3-10 Exemple d'ambiguïté de segmentation

Heuristique employée en cas d'ambiguïté	Précision	Rappel
Plus court d'abord	98,85 %	98,28 %
Plus long d'abord	98,78 %	98,21 %
Choix aléatoire	98,83 %	98,26 %

Tableau 3-2 Précision et rappel de l'algorithme de segmentation mis au point, sous diverses hypothèses de résolution des ambiguïtés

[38] Ce principe se rapproche de celui de segmentation des mots chinois basé sur l'algorithme « maximum matching » (*cf.* Wong et Chan [WON 96]), qui atteint une précision d'environ 95% pour les textes chinois.

Le Tableau 3-2 ci-dessus présente la performance du système mis en œuvre sur un corpus de 110 000 unités lexicales (ponctuations comprises), dont 20 % environ de mots composés, en employant les trois méthodes automatiques simples mentionnées pour lever les ambiguïtés. Le système évalué réalise, outre la segmentation selon les principes décrits précédemment, une fusion automatique des noms propres lorsque plusieurs syllabes successives portent des majuscules. La *précision* est la proportion d'occurrences de mots du texte segmenté automatiquement correspondant à une occurrence du texte de référence, et le *rappel* est la proportion d'occurrences de mots du texte de référence qui apparaissent dans le texte segmenté automatiquement.

Les résultats obtenus sont d'assez bonne qualité, en particulier si l'on considère qu'ils sont obtenus sans apprentissage sur un texte pré-segmenté. Quoique les différences entre les différentes heuristiques de résolution des ambiguïtés soient à peine significatives, les résultats suggèrent que l'approche « plus court d'abord » est la plus efficace – notamment si l'on considère les positionnements respectifs de « plus court d'abord » et « plus long d'abord » par rapport à la sélection aléatoire. Le principal type d'erreur observé est le « sur-rassemblement », où deux mots monosyllabiques successifs sont rassemblés par le système pour former un mot composé ; ce phénomène correspond à lui seul à un taux d'erreur de 0,6 % environ. Ce type d'erreur étant inévitable si l'on s'en tient à l'approche adoptée, il est nécessaire si l'on souhaite faire descendre le taux d'erreur sous cette proportion de 0,6 % de fonder la segmentation sur des données acquises sur un corpus d'apprentissage.

La seconde faiblesse de l'approche retenue est de ne pouvoir regrouper que des syllabes formant des mots apparaissant dans le dictionnaire de référence, ce qui empêche que soit pris en compte les « nouveaux » mots, beaucoup de formes redoublées, ainsi que des locutions considérées au cours de la segmentation manuelle comme constituant une entité lexicale. Cette limitation peut dans certains cas admettre une solution autre que l'extension du dictionnaire employé. En effet, lorsqu'il existe un nouveau mot dans le texte, ce nouveau mot peut être :

- une transcription d'un mot étranger, ou
- un nouveau terme dérivé des syllabes existantes dans la base des syllabes vietnamiennes, parmi lesquelles :
    - syllabes sino-vietnamiennes (ne se trouve pas toute seule)
    - syllabes vietnamiennes ou syllabes sino-vietnamiennes déjà vietnamisées[39]

Si leur graphie suit les recommandations officielles (syllabes séparées par des tirets), la gestion des nouveaux mots du premier cas n'est pas difficile ; en revanche, nous ne disposons pas de solution pour les cas où la transcription est faite sans tirets. Dans le deuxième cas, on peut se fonder sur les règles d'utilisation des syllabes sino-vietnamiennes non autonomes. Dans le troisième cas – et cette proposition est également valable pour les autres types de mots composés non connus – il sera sans doute profitable d'adapter au vietnamien les outils de recherche de termes complexes fondés sur des mesures statistiques telles que l'information mutuelle, développés pour les langues occidentales (par exemple Daille [DAI 94] ou Bourigault [BOU 94] pour le français, Lauriston [LAU 94] pour l'anglais).

En ce qui concerne les formes redoublées, il sera nécessaire de construire une base de règles de redoublement, entreprise qui se heurte, comme nous l'avons déjà évoqué, aux difficultés inhérentes au passage écrit-phonétique. Il existe néanmoins dans la littérature linguistique vietnamienne des études de ce phénomène également approfondies d'un point de vue orthographique. Il sera par ailleurs nécessaire de développer des heuristiques permettant de faire la distinction entre rapprochements phonétiques « accidentels » et volontaires.

[39] selon notre convention, il s'agit des syllabes sino-vietnamiennes qui peuvent-être utilisées comme des mots simples en vietnamien moderne

Le cas des noms propres, enfin, peut être abordé sous deux angles. Le premier est la présence de majuscules : la reconnaissance des noms propres de personnes et de lieux en milieu de phrase ne pose pas de problème grâce à cet indice ; il est ainsi possible de fusionner les parties des noms propres composés, et d'éviter que les noms propres simples ne fusionnent avec d'autres syllabes. Il semble en revanche impossible de trouver une solution au problème de la reconnaissance de noms propres en début de phrase sans faire appel à des considérations sur la structure des énoncés relevant (au moins) du niveau de l'analyse syntaxique. Nous décidons donc dans l'immédiat de ne pas considérer ce problème, qui reste d'une fréquence très limitée. La seconde approche possible du problème des noms propres consisterait à construire une base les recensant. Cela est impossible pour les noms de personnes et de lieux qui ont le plus souvent, comme nous l'avons déjà précisé, un sens ; il est en revanche possible de recenser les noms d'organisations, institutions, *etc.*, les plus courants, ce qui doit également permettre de contourner le problème de leur capitalisation irrégulière.

En résumé, nous ne cherchons pas encore à résoudre tous les problèmes de segmentation, mais la méthode simple présentée ci-dessus nous a permis d'automatiser une grande partie du processus de segmentation. D'autres solutions envisageables sont l'utilisation de fréquence de mots acquise d'un grand corpus déjà segmenté, l'intégration du *tokenizer* à l'étiqueteur en laissant les cas ambigus (qui seront désambiguïsés en même temps que les étiquettes), *etc*.

3.4.4. Étiquetage *a priori*

Avec l'aide d'un lexique, c'est l'étape le plus simple du système : les mots sont étiquetés *a priori* par leurs étiquettes trouvées dans le lexique. Au cas où un mot n'appartient pas au lexique, un prédicteur simple permet de déterminer si le mot est un nombre ou un nom propre (première lettre en majuscule), ou encore une abréviation (toutes les lettres sont en majuscule), *etc*. Si oui, l'étiquette correspondante est attribuée, sinon le nouveau mot reçoit toutes les étiquettes existantes dans le lexique.

3.4.5. Désambiguïsation

Nous évaluons dans cette section les résultats de l'application à des textes vietnamiens d'une méthode « classique » probabiliste de désambiguïsation d'étiquetage pour les langues occidentales, décrite dans ce chapitre à la section 3.2.4.2.

3.4.5.1. Algorithmes

Afin de réaliser la désambiguïsation des textes vietnamiens, nous avons adopté un algorithme probabiliste couramment employé pour répondre à ce type de problématique (*cf*. 3.2.4.2), par exemple dans le programme TATOO développé à l'ISSCO[40] ou l'étiqueteur QTAG (*cf*. Mason et Tufiş [MAS 98]).

Le programme TATOO modélise chaque phrase à désambiguïser par une chaîne de Markov cachée, dont les états observables sont les classes d'ambiguïté (groupe d'étiquettes syntaxiques parmi lesquelles un choix doit être effectué) à traiter, et les états sous-jacents les étiquettes réelles à retrouver.

[40] *http://www.issco.unige.ch/staff/robert/tatoo/tatoo.html*

Disposant d'une quantité de données étiquetées manuellement relativement importante, nous avons pour notre part fait appel au modèle de l'étiqueteur QTAG, où les états observables sont les mots eux-mêmes et non pas les classes d'ambiguïté, ce qui permet de faire usage des probabilités des étiquettes envisageables pour chaque mot (prenant ainsi en compte, par exemple, le fait que l'étiquette "nom" pour le mot "ferme" peut être absolument non pertinente pour de nombreux textes), que nous avons la chance de pouvoir calculer. Toutes les probabilités sont calculées à partir des observations effectives ; concernant les transitions non observées dans les textes, nous évaluons leur probabilité à partir des probabilités individuelles d'apparition des mots, dans les limites d'un « poids de probabilité » total pour ces cas résiduels calculé par l'algorithme de Good-Turing (dans sa version présentée par Gale et Sampson dans [GAL 95]).

L'algorithme de calcul des probabilités employées pour l'étiquetage, ainsi que l'implémentation de l'algorithme de Viterbi pour la désambiguïsation ont été implémentés en Perl. Nous disposons également d'une version QTAG (implémentée en Java) adaptée pour le vietnamien (vnQTAG, *cf.* Nguyen *et al.* [NGU 03a, b]) mise en ligne[41], ainsi que des ressources linguistiques associées (lexique, corpus d'apprentissage).

Dans le cadre général des méthodes probabilistes présenté à la section 3.2.4.2, nous avons réalisé des expériences d'étiquetage bayésien (sélection pour chaque mot de son étiquette la plus fréquente), markovien (prise en compte des probabilités d'association mot-étiquette et des probabilités d'apparition de bigrammes d'étiquettes), et par trigramme (*idem* markovien, en prenant en compte des trigrammes au lieu de bigrammes).

Nous avons également introduit une variante par rapport au modèle probabiliste pur : au cours de l'apprentissage des probabilités d'apparition de bigrammes et trigrammes, le système développé recense l'ensemble des séquences de deux mots apparaissant au moins deux fois dans le corpus d'apprentissage et telles que les étiquettes associées aux mots qui composent une séquence sont toujours les mêmes, dans toutes ses occurrences. Au cours de l'étiquetage, la connaissance de cette régularité est ensuite exploitée comme une règle, avant application de la méthode probabiliste de sélection des étiquettes optimales.

La sous-section suivante présente une évaluation des résultats que ces méthodes nous ont permis d'atteindre pour le vietnamien.

3.4.5.2. Évaluation

L'apprentissage et l'étiquetage ont été réalisés pour évaluation en considérant trois jeux d'étiquettes distincts (*cf.* Annexe B) :

- le premier contient 8 catégories principales proposées par le Comité des Sciences Sociales [UYB 83] : Nom, Verbe, Adjectif, Pronom, Adverbe, Conjonction, Particules modales et Interjection ;
- le second est un jeu de 18 étiquettes constituant une « moyenne » entre les deux autres jeux ;
- le troisième jeu est constitué de 55 étiquettes qui détaille les sous-familles de noms, verbes, *etc*.

En outre, pour chacun de ces jeux d'étiquettes, nous avons une étiquette *X* pour étiqueter des mots non catégorisés, et des étiquettes distinctes pour les ponctuations.

[41] http://led.loria.fr/outils.php

Les deux « petits » jeux d'étiquettes étant des sous-ensembles du troisième, il n'a pas été nécessaire de réaliser plusieurs fois l'étiquetage du corpus de référence. Celui-ci a été réalisé manuellement par plusieurs experts du Centre de Lexicographie du Vietnam. Le corpus ainsi constitué contient environ 110 000 unités lexicales (~ 6 300 phrases). La plupart des textes sont des textes littéraires, le reste étant extrait de journaux d'information. Malgré toute l'attention accordée à cette opération d'étiquetage manuel, il s'agit d'une tâche difficile, notamment étant donnée la finesse et la nature parfois sémantique des étiquettes ; quelques erreurs résiduelles sont donc toujours observables dans le corpus de référence, mais la qualité de celui-ci reste très nettement supérieure à ce que peut être le résultat d'un étiquetage automatique.

Les résultats présentés doivent être interprétés en gardant à l'esprit le très fort taux d'ambiguïté de l'étiquetage en vietnamien : pour le jeu de 8 étiquettes, celui-ci est de 1,8 étiquettes/mot en moyenne, et cette valeur passe à 2,7 étiquettes/mot pour le jeu de 55 étiquettes. À titre de comparaison, le taux d'ambiguïté pour le français observé par l'étiqueteur TreeTagger[42] lors de l'étiquetage de « La République » de Platon est de 1,6 étiquettes/mot avec un jeu de 14 étiquettes.

La validation a été effectuées selon le principe classique de la validation croisée répétée 10 fois : le corpus étiqueté disponible est découpé par tirage au sort en 10 parties de taille équivalente (dans notre cas, ces parties contiennent un nombre de phrases identique), et l'on répète dix fois l'opération consistant à réaliser l'apprentissage sur neuf des dix parties et l'évaluation sur la dixième. Les chiffres donnés ci-après correspondent aux valeurs moyennes calculées.

	Aléatoire	Sans séquence			Avec séquences		
		Bayésien	Bi-gramme	Tri-gramme	Bayésien	Bi-gramme	Tri-gramme
8 étiq.	27,44%	6,35%	6,19%	6,07%	4,60%	4,55%	4,49%
18 étiq.	30,17%	7,48%	7,51%	7,39%	5,87%	5,89%	5,82%
55 étiq.	49,94%	12,76%	12,70%	12,65%	11,02%	11,04%	10,98%

Tableau 3-3 Taux d'erreurs de l'étiquetage automatique avec une méthode probabiliste

Le Tableau 3-3 présente les taux d'erreurs obtenus en appliquant les méthodes probabilistes avec trois tailles de voisinage (*k*, *cf.* 3.2.4.2) différentes : $k = 1$ (modèle bayésien), $k = 2$ (modèle bi-gramme), $k = 3$ (modèle tri-gramme) sur les trois jeux d'étiquettes (8, 18 et 55 étiquettes) ; dans un premier temps, sans utiliser l'information des séquences fixes apprises du corpus d'entraînement, et dans un deuxième temps, avec l'utilisation de ces séquences. La première colonne mentionne, pour référence, le taux d'erreur d'un étiquetage réalisant une sélection aléatoire parmi toutes les étiquettes possibles de chaque occurrence de mot.

Comme nous pouvons le constater, la connaissance des séquences fixes (séquences de deux mots apparaissant toujours avec une même étiquette pour chaque mot à toutes ces occurrences) améliore notablement la qualité de l'étiquetage automatique (au moins 1,5% de gain de précision dans tous les cas). En revanche, l'apport du modèle tri-gramme est très faible par rapport au modèle bayésien, et le modèle bi-gramme n'est nullement meilleur que le modèle bayésien.

La première constatation prouve l'importance de la connaissance des séquences fixes. L'examen de la liste de ces séquences fait apparaître les catégories suivantes :

- ponctuation suivie d'un mot généralement fonctionnel (conjonction, déterminant, pronom, adverbe),

[42] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html

- partie d'expression figée,
- groupe classificateur-nom,
- groupe déterminant-classificateur ou déterminant-nom.

La première de ces catégories correspond à une information redondante par rapport à la connaissance apportée par les *n*-grammes ; l'apport des séquences fixes n'est dans ce cas que de « forcer la main » à l'algorithme de Viterbi, ce qui ne présente de réel intérêt que du point de vue du gain en complexité. Les trois catégories suivantes apportent pour leur part une information « lexicalisée » réellement supplémentaire par rapport à celle contenue dans les probabilités de *n*-grammes d'étiquettes. Si les deuxième et quatrième catégories correspondent à des motifs que l'on pourrait observer en appliquant une méthode similaire à un texte occidental, la troisième reflète un phénomène linguistique propre au vietnamien, le motif « classificateur-nom ». Une grande partie des classificateurs pouvant également être employés comme des noms à part entière, la désambiguïsation réalisée grâce à ces motifs est très importante en volume, et également très fiable, un nom donné ne pouvant être introduit que par un petit nombre de classificateurs.

La seconde constatation effectuée, concernant le faible intérêt relatif de l'emploi de *n*-grammes pour la désambiguïsation de l'étiquetage, paraît *a priori* plus surprenante : les méthodes à base de modèles de Markov cachés sont très répandus pour l'étiquetage de textes en langues occidentales, et généralement reconnus comme donnant des résultats d'assez bonne qualité. Ce résultat trouve son explication dans la définition des jeux d'étiquettes utilisés, eux-mêmes conditionnés par les propriétes particulières de la langue vietnamienne. Pour une langue flexionnelle, en effet, la cohésion du flux textuel est en grande partie assurée par les phénomènes d'accord. La définition de jeux d'étiquettes « étendus » par rapport aux types morphosyntaxiques de base, par exemple pour permettre une analyse syntaxique du texte, passe ainsi typiquement par l'intégration d'informations flexionnelles aux étiquettes. Le phénomène d'accord se manifestant souvent dans un contexte de très forte proximité, les *n*-grammes constituent dans de nombreux cas un outil suffisant pour assurer la cohérence de l'étiquetage d'un texte.

Dans le cas du vietnamien, où il n'existe pas de variation flexionnelle, il est nécessaire d'avoir recours à un autre type d'informations afin de pouvoir exprimer les contraintes régissant la cohésion du texte. Les étiquettes que nous avons définies intègrent, afin de permettre par la suite un travail d'analyse syntaxique, des informations de type sémantique : « nom de matériau », « verbes d'émotion », *etc*. Contrairement aux relations flexionnelles, les relations sémantiques entre mots interviennent souvent à plus longue portée dans le texte ; il est également probable qu'il soit nécessaire, afin de décider du sous-groupe sémantique auquel un mot doit être rattaché dans un usage donné, de faire usage de la connaissance de mots qui l'entourent, et non seulement de leurs étiquettes.

Cette dernière constatation, cohérente avec l'importance que nous avons vu devoir accorder au rôle des séquences fixes (qui mêlent également mots et étiquettes), milite en faveur de l'usage d'une méthode d'étiquetage à base de règles (*cf.* 3.2.4.1) comme l'étiqueteur de Brill [BRI 95], capable d'extrapoler à partir de ses observations des règles mêlant étiquettes et mots. On peut également envisager, plus cette fois pour répondre à la problématique de non détermination sémantique par le contexte immédiat, d'employer une méthode toujours probabiliste mais ne se limitant pas aux *n*-grammes immédiats, par exemple en construisant des profils statistiques des mots employés dans les voisinages de *n* mots autour des occurrences d'une étiquette.

3.5. Bilan et perspectives

Du fait de la jeunesse de l'implication en TAL des chercheurs vietnamiens, nous avons dû construire toutes les ressources linguistiques nécessaires à l'étiquetage lexical et définir toutes les structures de données sans aucune base préexistante. Nous avons néanmoins pu pour nos premières expériences tirer bénéfice de quelques avantages : le grand nombre de méthodologies existantes pour l'annotation morphosyntaxique et une forte conscience de la nécessité de normalisation des ressources.

La Figure 3-11 résume le travail que nous avons réalisé pour l'annotation morphosyntaxique du vietnamien. Il s'agit de deux tâches principales : d'une part, la construction des ressources linguistiques participant au processus d'étiquetage lexical, d'autre part, le développement des outils d'étiquetage et d'assistance à la construction de ces ressources.

Les ressources linguistiques construites se composent d'un lexique morphosyntaxique et d'un corpus de référence annoté morpho-syntaxiquement. Pour la construction de la base lexicale, nos contributions concernent notamment les trois points suivants :

- Définition des unités lexicales : nous choisissons une définition d'unité lexicale qui est assez fine. Ce choix nous laisse toujours ouverte la possibilité de traiter les éléments composés en cas de besoin et assure la réutilisabilité des ressources annotées.
- Définitions des descripteurs lexicaux : nous définissions les jeux d'étiquettes en partant des descriptions lexicales stables de la langue vietnamienne. Le jeu d'étiquettes défini pourrait aisément être réajusté et étendu grâce à ces descriptions. Celles-ci sont inspirées du projet MULTEXT, et se fondent sur une structure à deux niveaux de granularité : la couche noyau correspond à une classification « universelle », permettant la comparabilité des langues, la couche privée contient des informations de spécification de chaque catégorie de la couche noyau. Ces spécificités sont représentées grâce à des structures de traits, dont il a été montré qu'elles sont convenables pour la représentation des bases lexicales (*cf.* Ide *et al.* [VER 92, IDE 93]). Nous visons également l'intégration de ces descripteurs au DCR (*Data Category Registry, cf.* 1.2.2) – l'ensemble de catégories de données normalisées pour l'annotation morphosyntaxique.
- Proposition de la structure de représentation de données lexicales conformée aux modèles définis dans le cadre des activités internationales de normalisation (TEI, ISO/TC 37/SC 4).

La mise en application de ces concepts afin de construire un lexique couvrant les entrées d'un dictionnaire populaire du vietnamien a été effectuée par les linguistes du Centre de Lexicographie du Vietnam, dans le cadre du projet KC01-03 (*cf.* 1.3). Les outils d'assistance à la gestion de la base lexicale, ainsi qu'à la gestion et la révision manuelle des corpus annotés ont été développés par un stagiaire de l'Institut de la Francophonie pour l'Informatique à Hanoï (*cf.* Nguyen T. Bon [NGU 04b]).

Les ressources obtenues constituent une base pour d'autres recherches dans le domaine de TAL pour le vietnamien : analyse syntaxique, recherche d'information, alignement multilingue, traduction automatique, *etc*. Elles sont mises à la libre disposition de la communauté académique sur le site Internet de l'équipe Langue et Dialogue du Loria (http://led.loria.fr/outils.php).

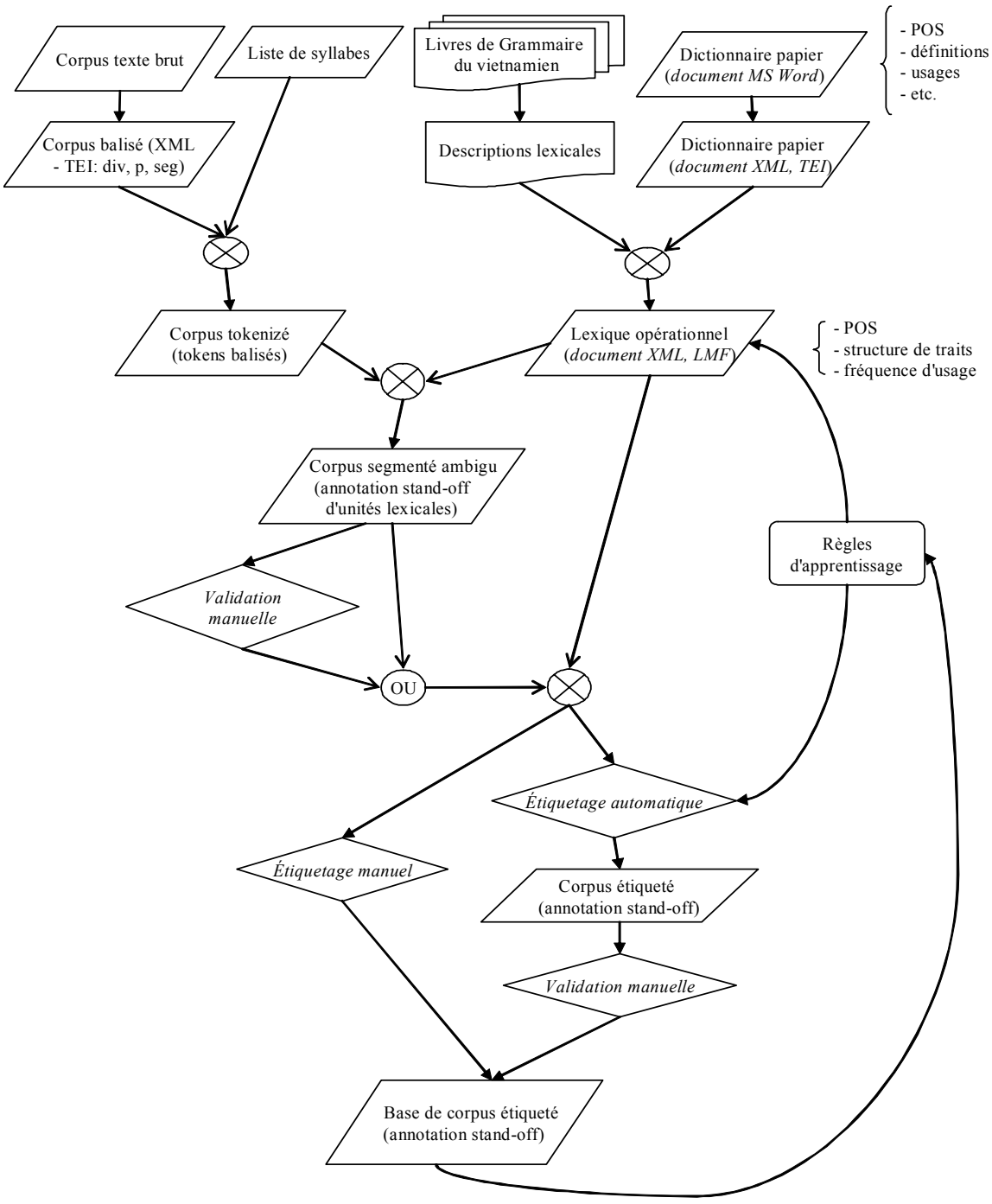


Figure 3-11 Schéma du travail effectué

Afin de favoriser la construction d'un corpus annoté de référence de plus grande taille, nous avons étudié l'application de méthodes « classiques » pour développer les outils de segmentation et d'étiquetage automatique. Possédant un lexique qui couvre la plupart des mots dans les textes, nous avons pu appliquer une méthode très simple (l'utilisation d'un automate d'états finis pour reconnaître les chemins possibles, puis choisis comme solution la segmentation correspondant au chemin le plus court) pour obtenir une segmentation automatique avec une haute précision (près de 99%) sans requérir aucune donnée d'apprentissage. L'étude que nous avons effectué des résultats obtenus par cette méthode suggère qu'il serait nécessaire, pour obtenir des résultats supérieurs, de faire appel à un corpus d'apprentissage pour estimer, par exemple, les probabilités d'apparition des mots ou de cooccurrences entre mots, ou acquérir la connaissance de formes composées non présentes dans le dictionnaire par étude des collocations.

En ce qui concerne l'automatisation de l'étiquetage morphosyntaxique automatique, nous développons et évaluons un outil d'étiquetage se fondant sur une méthode probabiliste, avec une variante exploitant la connaissance, acquise sur un corpus de référence, de séquences de mots telles que leurs étiquettes sont toujours identiques. L'évaluation sur un corpus de référence d'environ 10 000 unités lexicales, avec trois jeux d'étiquettes de granularité différente, a montré le grand impact positif de la connaissance de ces séquences, ainsi que la faiblesse relative de l'apport des modèles *n*-grammes par rapport au modèle bayésien. Ces deux particularités s'expliquent d'une part par la petite taille du corpus d'apprentissage, mais également par des spécificités linguistiques du vietnmaien, que nos expériences ont participé à mettre au jour et qui constituent un guide appréciable pour le développement de nouveaux outils. Nos constatations suggèrent en particulier que l'implémentation d'une méthode d'étiquetage à base des règles devrait être favorisée dans l'avenir.

Nous discutons dans la suite de cette section du travail restant à réaliser afin d'améliorer et étoffer les ressources annotées morpho-syntaxiquement. Deux sujets sont à considérer : l'amélioration du lexique (3.5.1) et celle du système d'étiquetage lexical (3.5.2).

3.5.1. Amélioration des ressources lexicales du vietnamien

L'objectif est de construire un lexique ouvert du vietnamien, mis à disposition de la communauté d'informatique linguistique pour accélérer la recherche en traitement automatique du vietnamien[43]. Aujourd'hui, dans le domaine de TAL, il existe un fort consensus sur la nécessité de partager des ressources linguistiques monolingues ainsi que multilingues. Pour le français, on peut constater plusieurs projets qui visent la construction de ressources ouvertes et coopératives au profit de la communauté de recherche en TAL : le projet Morphalou[44] pour la construction d'un lexique morphosyntaxique, le projet Papillon (*cf.* 1.1.1.2) pour la construction des lexiques multilingues, le projet FReeBank (*cf.* 1.1.4.1) pour la construction d'un corpus annoté à plusieurs niveaux (morphosyntaxique, syntaxique et co-référence), *etc*.

Pour la construction du lexique, deux aspects à considérer sont la gestion informatique et le contenu linguistique.

Au niveau de la gestion de la base de données, nous possédons d'ores et déjà un lexique au format XML qui s'intègre aux travaux de normalisation internationale en cours (projet LMF de l'ISO/TC 37/SC 4) permettant l'interopérabilité ainsi que l'extensibilité des lexiques. La structure de notre lexique est conçue de manière à se prêter à la perspective plus ambitieuse d'en réaliser une extension vers un lexique syntaxique du vietnamien (*cf.* 4.5.1).

Afin d'assurer une distribution libre ainsi qu'une contribution coopérative et régulière du lexique, plusieurs question pratiques restent à régler :

---

[43] Cet objectif pourrait trouver le cadre de sa réalisation dans un prochain projet national de TAL au Vietnam, projet à l'élaboration duquel nous participons.

[44] http://loreley.loria.fr/morphalou

- établir une licence de bonne utilisation (comme par exemple la licence du projet LeFFF[45]) ;
- élaborer une documentation précise du format adopté, comprenant l'ensemble des catégories de données utilisées ;
- développer un service web donnant les descriptions morphosyntaxiques des unités lexicales à la demande de l'utilisateur, outre une archive XML téléchargeable directement à partir du site ;
- définir un mécanisme de mise à jour permettant à des contributeurs de soumettre des *patch* (remplacement, modification, ajout) à la ressource, soit directement sous la forme d'un fichier, soit par le biais d'un formulaire. Les procédures de validation linguistique de ces constributions doivent également être définies.

Pour la gestion des contributions et des versions (au niveau d'entrée lexicale ou globale), il est envisageable d'adopter des descripteurs de méta-données définis dans les initiatives internationales comme OLAC[46] (*Open Language Archives Community*) ou IMDI[47] (*ISLE MetaData Initiative*).

Quant au contenu linguistique, le lexique actuel contient toutes les unités lexicales d'un dictionnaire publié par l'Institut National de Linguistique et le Centre de Lexicographie du Vietnam. Notre lexique comporte environ 37 000 unités lexicales, dont 18 732 noms, 975 numéraux et déterminants, 12 209 verbes, 8 442 adjectifs, et environ 2 000 mots outils. Nous constatons qu'il reste encore environ 1 000 mots du dictionnaire qui ne sont pas catégorisés dans la base (classe Résidu *X*), ce qui demande une étude empirique afin de trouver des descripteurs convenables pour ces unités.

Nous envisageons également de compléter la base avec des éléments non autonomes (*cf.* 3.3.2.12) qui ne sont pas présents dans le dictionnaire papier. Ces éléments participent potentiellement à la composition des nouveaux mots.

Une tâche importante est d'améliorer la qualité des descripteurs lexicaux, et de contrôler la cohérence d'annotation des unités lexicales. En effet, bien qu'ayant profité pour l'étiquetage de la collaboration de plusieurs linguistes, nous n'avons pas pu réaliser d'annotation multiple, seule méthode permettant de faire apparaître les incohérences ou points de désaccord entre experts, et d'obtenir un corpus de référence réellement fiable.

3.5.2. Amélioration du système d'étiquetage lexical

Outil d'étiquetage

Nous avons déjà présenté au cours de ce chapitre plusieurs pistes envisageables pour accroître la qualité des résultats de la segmentation et de l'étiquetage automatique, et ne revenons pas ici plus en détail sur ces considérations techniques.

Deux points de travail restent néanmoins ouverts :

- L'interaction entre segmentation et étiquetage n'a pas été abordée au cours de notre étude, quoiqu'il s'agisse d'un point potentiellement très important : en particulier, la définition de la granularité de segmentation cherchée dépend beaucoup du jeu d'étiquettes employé ; il semble à tout le moins nécessaire que le système d'étiquetage ait la capacité de revenir sur certains choix effectués au cours de la segmentation, et que le système de segmentation puisse choisir en cas d'ambiguïté forte de déléguer la responsabilité de la désambiguïsation au système d'étiquetage. Un travail d'intégration de ces deux outils paraît donc potentiellement intéressant.

[45] Clément *et al.* [CLE 04], http://www.lefff.net/

[46] http://www.language-archives.org/

[47] http://www.mpi.nl/IMDI/

- Le choix du jeu d'étiquettes conditionne largement, comme nous l'avons vu, la qualité de l'étiquetage obtenu. Nos choix ont dans ce domaine principalement été guidés par les exigences de la tâche d'analyse syntaxique (que nous présentons en détail au chapitre suivant), mais il est également possible de considérer qu'il serait plus pertinent de définir pour l'étiquetage morphosyntaxique des étiquettes « faciles », en laissant à l'analyse syntaxique le soin d'attribuer des étiquettes plus fines. Une étude systématique serait dans ce cas nécessaire pour définir le jeu d'étiquettes rencontrant un « succès » optimal.

Corpus annoté de référence

Le but est de concevoir un corpus annoté de grande taille.

Comme nous l'avons déjà mentionné, le Centre de Lexicographie Vietlex possède un large corpus brut libre de droit de 50 millions de syllabes (environ 2 millions de phrases). Ce corpus, collecté par le Vietlex durant plusieurs années, fait à l'heure actuelle l'objet de négociations pour pouvoir être mis à la disposition de la communauté de recherche publique au Vietnam. Selon la main d'œuvre disponible pour réaliser cette tâche, tout ou partie de l'annotation morphosyntaxique pourra être validée manuellement.

Le corpus de référence doit fait également l'objet d'une distribution publique au sein de la communauté de TAL. Des procédures similaires à celles définies pour la distribution du lexique, présentée au paragraphe ci-dessus, pourraient s'appliquer pour que de multiples équipes de recherche puissent profiter de ces ressources et les faire évoluer.

En conclusion, nous avons présenté dans ce chapitre les travaux menés pour construire les ressources linguistiques fondamentales pour l'annotation morphosyntaxique : un lexique avec des descriptions lexicales à large couverture, et un premier corpus annoté morpho-syntaxiquement. Bien que les ressources obtenues soient encore loin d'être parfaites, elles sont prêtes à être améliorées et constituent les premières briques pour l'annotation morphosyntaxique du vietnamien. Nous avons rencontré beaucoup d'obstacles pour leur construction, dus au faible consensus des linguistes sur la catégorisation grammaticale du vietnamien et à la difficulté de désambiguïsation manuelle des étiquettes des mots en contexte (*cf.* 2.4.1.8). Cependant, cela ne doit pas être une source de découragement, car l'annotation des corpus réels est en effet un moyen de consolider les choix linguistiques. Nous avons également mis au point des systèmes pour l'automatisation de la segmentation et de l'étiquetage des textes vietnamiens, atteignant une relativement bonne précision pour la première de ces tâches, et proposant pour la seconde une première solution fonctionnelle, quoique demandant encore à être amendée pour mieux prendre en compte les spécificités de la langue vietnamienne. Ces premiers outils doivent nous permettre de stimuler le développement de ressources linguistiques pour le vietnamien, toujours dans un cadre de normalisation permettant leur partage et réutilisation.

Ce premier niveau de connaissance grammaticale sur les textes vietnamiens étant ainsi atteint, nous pouvons maintenant nous intéresser à l'étape suivante dans l'analyse des textes et la construction de ressources linguistiques : l'analyse syntaxique.

Chapitre 4

Ressources linguistiques pour l'analyse syntaxique du vietnamien

Ce chapitre discute de la modélisation de la grammaire vietnamienne à l'aide du formalisme TAG (Grammaire d'Arbres Adjoints), que nous avons expérimentée grâce au parseur LLP2 développé au Loria. Dans le cadre de cette thèse, il n'est bien sûr pas question d'aboutir à une analyse syntaxique à large couverture, mais nous montrons que l'approche TAG permet de couvrir suffisamment de phénomènes observables sur le vietnamien. Nous finissons ce chapitre par une spécification de ce que pourrait être une TreeBank à la vietnamienne.

- *Introduction : Analyse syntaxique*
- *Formalismes de grammaire et analyseurs syntaxiques*
- *Descriptions syntaxiques du vietnamien*
- *Conclusions et Perspectives*

4.1. Introduction

De nombreuses applications dans le domaine du TAL utilisent directement des composants syntaxiques : la correction d'orthographe, l'indexation automatique, l'interrogation de bases de données, la simplification de textes (pour un traitement ultérieur comme le résumé de texte ou la traduction), l'alignement automatique, l'extraction de connaissances linguistiques à partir de textes, la génération de phrases, *etc.* La tâche d'analyse syntaxique est donc essentielle pour le développement de nombreux outils de TAL. Elle peut être décomposée en plusieurs sous-problématiques distinctes quoique pas tout à fait indépendantes (Abeillé [ABE 00]) :

- parenthésage (identification des frontières syntagmatiques majeures),
- assignation de fonctions aux syntagmes distingués (ou à leur tête),
- désambiguïsation syntaxique des têtes prédicatives (cadres de sous-catégorisation, actif/passif, *etc.*),
- assignation d'une structure syntaxique globale (un arbre) à chaque phrase.

On distingue deux familles principales d'analyseurs syntaxiques : analyseurs « de surface » (*shallow parsers*) et « en profondeur » (*deep parsers*). Les analyseurs de surface se limitent à l'identification des frontières syntagmatiques et à la mise en évidence de certains liens syntaxiques majeurs (typiquement, tête-complément), et se basent uniquement sur les catégories morphosyntaxiques des mots en employant des règles probabilistes. Ils présentent l'avantage d'être moins sensibles aux phénomènes d'agrammaticalité des textes analysés et beaucoup plus rapides que les analyseurs en profondeur, qui associent à chaque phrase une structure arborée complète.

Deux défis majeurs doivent être relevés pour le développement d'analyseurs syntaxiques en profondeur (Villemonte et Rajman [VIL 03]). Le premier est celui de « couverture grammaticale », qui implique de définir des grammaires susceptibles de rendre compte d'une part la plus importante possible des phénomènes grammaticaux observables dans un texte. Néanmoins, le nombre d'analyses possibles pour une phrase donnée tend à croître considérablement avec la sophistication de la grammaire employée, faisant apparaître la seconde difficulté a surmonter : celle de maîtrise de l'ambiguïté.

Avec la stabilisation des formalismes de syntaxe, on constate un effort de construction de grammaires à large couverture des langues, qui vise à la « réutilisabilité » (et la réversibilité) de ces grammaires pour multiples tâches ultérieures. Or, « une grammaire électronique doit être basée sur un formalisme pour être cohérente et extensible, mais aussi sur des données pour être 'couvrante' » (Abeillé [ABE 00]).

Dans le cadre de notre thèse, nous avons pour but de mettre en place un cadre de construction des ressources linguistiques pour l'analyse syntaxique du vietnamien. Etant donné qu'aucune ressource pour le vietnamien n'est disponible jusqu'à présent, et aucune modélisation formelle de la langue vietnamienne n'a été faite, nous tentons donc de développer deux types de ressources :

- Une modélisation de la grammaire vietnamienne suivant un formalisme syntaxique ;
- Un corpus arboré de type TreeBank (*cf.* 1.1.4) du vietnamien.

Ce chapitre donne une brève introduction des formalismes syntaxiques et des systèmes d'analyse syntaxique (section 4.2.1), ainsi que des méthodes d'évaluation (section 4.2.2). Cela constitue le contexte de notre travail, qui nous guide ensuite à choisir un formalisme (section 4.3) pour la représentation de connaissances syntaxiques du vietnamien (section 4.4) et à analyser la possibilité de construction des ressources pour le traitement syntaxique de la langue vietnamienne (section 4.5).

4.2. Formalismes de grammaire et systèmes d'analyse syntaxique

4.2.1. Formalismes de grammaire

Les formalismes grammaticaux sont des langages (métalangages) destinés à décrire l'ensemble des phrases possibles d'une langue, les propriétés structurales de ces phrases (leur syntaxe) et parfois leur signification (leur sémantique). Comme énoncé dans Schieber [SHI 86], la définition d'un métalangage peut se justifier en fonction des buts suivants :

- fournir un outil précis de description des langues naturelles ;
- délimiter la classe des langues naturelles qu'il est possible de décrire grâce à cet outil ;
- donner une caractérisation des langues naturelles qui soit interprétable par ordinateur.

Le choix du métalangage se base sur les critères suivants pour juger de l'adéquation des formalismes grammaticaux : la facilité linguistique, le pouvoir expressif et l'efficacité pour le traitement automatique.

Nous donnons maintenant une brève introduction sur les formalismes de grammaire fondamentaux ou particulièrement actifs dans le domaine du TAL, ce qui nous guide dans notre choix d'un formalisme pour décrire la grammaire du vietnamien.

4.2.1.1. Grammaires de réécriture

Le développement des théories syntaxiques formelles applicables en TAL est marqué par plusieurs contributions importantes de N. Chomsky [CHO 57]. Chomsky a fondé la théorie des langages formels et défini une hiérarchie de classes de grammaires et de langages[48], dont chacune correspond à un nouvel ordre de grandeur de complexité algorithmique d'analyse (*cf.* Tableau 4-1).

Type de langage	Nom de la grammaire	Complexité d'analyse
0	non contrainte	indécidable
1	contextuelle	exponentielle
2	hors contexte	polynomiale $O(n^3)$
3	régulière	linéaire

Tableau 4-1 Complexité d'analyse des grammaires

La notion d'équivalence entre grammaires constitue un outil important pour la comparaison de celles-ci : deux grammaires sont dites ***faiblement équivalentes*** si elles génèrent le même langage. Elles sont ***fortement équivalentes*** si elles génèrent le même langage par les mêmes dérivations (c'est-à-dire en associant les mêmes descriptions syntagmatiques aux mêmes phrases).

[48] Un rappel de ces notions se trouve au Glossaire

Les grammaires hors contexte (*CF*G – *Context Free Grammar*) et régulières, couramment appelées ***grammaires de constituants*** (ou ***syntagmatiques***), ont tout d'abord été le principal centre d'attention des recherches portant sur les grammaires génératives, parce qu'elles reflétaient le principe d'analyse des phrases en constituants immédiats des grammaires « traditionnelles » (Bloomfield [BLO 33], Harris [HAR 51]) et se prêtaient aisément à une implémentation informatique. Pourtant, Chomsky [CHO 57] a soutenu que la capacité générative faible[49] des grammaires régulières ne suffisait pas à couvrir le langage naturel, les *CF*G n'étant pour leur part pas suffisantes pour ce qui est de la capacité générative forte – c'est-à-dire qu'elles permettent de générer les énoncés voulus mais que les schémas de dérivation employés pour cela ne peuvent refléter des phénomènes tels que dépendances croisées, expressions ambiguës, parenté entre phrases, *etc.*

Afin de répondre a ces critiques, Chomsky a proposé un modèle alternatif appelé ***grammaire générative transformationnelle*** constitué de deux composants : un composant génératif (règles de réécriture syntagmatiques) pour produire la structure profonde[50] d'une phrase et un composant transformationnel (règles de transformation) pour obtenir sa structure de surface. Le modèle transformationnel a été développé en plusieurs étapes successives pour aboutir enfin à la formulation de la Théorie Standard Étendue, dominante durant les années soixante-dix. Pourtant, ce type de formalisme se heurtait à des problèmes d'implémentation : ses grammaires étaient formellement équivalentes aux grammaires non contraintes (donc indécidables), difficiles à maintenir et à mettre à jour, non réversibles, *etc.* Il a également reçu des critiques sur les aspects sémantique et psycholinguistique (*cf.* Abeillé [ABE 93]).

Parallèlement aux travaux de Chomsky, le développement des systèmes de programmation logique a permis l'émergence des formalismes grammaticaux fondés sur l'unification, dont nous présentons ci-après le principe. Ces formalismes sont développés dans une tentative de réhabiliter les différents modèles de grammaires formelles critiqués par Chomsky, en visant à enrichir le pouvoir expressif des grammaires de réécriture.

4.2.1.2. Grammaires d'unification

Une très bonne introduction aux théories de grammaire à base de l'unification est présentée dans Schieber [SHI 86]. Dans la littérature française, on peut recourir à l'ouvrage édité par Philip Miller et Thérèse Torris [MIL 90a], ainsi qu'à l'ouvrage d'Anne Abeillé [ABE 93] pour les présentations détaillées des grammaires d'unification. Dans cette section, nous rappelons les concepts principaux de ces formalismes.

Les formalismes de grammaire à base de l'unification sont le résultat de recherches distinctes en linguistique informatique, en linguistique formelle et en TAL. On trouve des techniques apparentées dans des domaines tels que la démonstration automatique de théorèmes, la recherche sur la représentation de connaissances et la théorie des types de données. Plusieurs courants de recherche, à l'origine indépendants, ont donc convergé vers l'idée de l'unification comme moyen de contrôle du flux d'information.

Le premier point commun des grammaires d'unification est le rejet des transformations chomskyennes. Elles émettent en outre plusieurs contraintes spécifiques sur les qualités nécessaires d'un formalisme grammatical :

- *surfaciste* : le formalisme doit donner une description directe de l'ordre réel des chaînes dans la phrase,
- *informatif* : il doit associer aux chaînes linguistiques des informations résultant de la combinaison d'attributs prédéfinis,

---

[49] Le langage généré par une grammaire définit la ***capacité générative faible*** de celle-ci, alors que l'ensemble des séquences de symboles, terminaux ou non, qu'elle engendre, constitue sa ***capacité générative forte***.

[50] voir Glossaire: structure profonde, structure de surface

- *inductif* : il doit définir l'association entre chaînes et éléments d'information de manière récursive, afin que de nouvelles associations puissent être dérivées en combinant selon des opérations préétablies les chaînes partielles et les informations qui leur sont associées,
- *déclaratif* : l'association entre chaînes et éléments d'information doit être définie par des assertions et non par des procédures. Les règles ne sont pas ordonnées, et chaque règle ne peut qu'ajouter de l'information, sans modification destructrice.

Les éléments d'un domaine d'information sont des ***structures de traits*** qui associent aux traits des valeurs tirées d'un ensemble bien défini, et éventuellement structuré. L'opération fondamentale pour l'association des éléments d'information est l'***unification***. Les définitions des notions concernant les structures de traits et leurs opérations se trouvent au Glossaire.

La plupart des formalismes de grammaire contemporains peuvent être classés dans le courant de grammaires d'unification. Shieber [SHI 86] distingue deux classes de formalismes :

- Des formalismes du type *outil* (par ex. FUG, DCG, PATR, *cf.* [MIL 90a, ABE 93]) conçus à des fins d'implémentation, de métalangages de description, mais non de théories syntaxiques, dans le sens que l'on ne voit pas se dégager de principes linguistiques généraux qu'on pourrait reprendre pour passer d'un phénomène à un autre ou d'une langue à une autre.
- Des formalismes du type *théorie syntaxique* (par ex. LFG, GPSG, *cf.* [MIL 90a, ABE 93]) qui ont tendance à incorporer des instruments définis pour des besoins spécifiques, liés aux hypothèses linguistiques auxquelles ils adhèrent pour atteindre l'adéquation explicative.

Ayant pour but de modéliser la grammaire vietnamienne, nous nous intéressons ici au deuxième type de formalismes : les formalismes comme théories syntaxiques. Quatre modèles représentatifs de cette classe sont détaillés dans Abeillé [ABE 93], accompagnés de nombreuses données linguistiques du français : grammaire lexicale fonctionnelle (***LFG*** – *Lexical Functional Grammar*) – Bresnan [BRE 82] –, grammaire syntagmatique généralisée (***GPSG*** – *Generalized Phrase Structure Grammar*) – Gazdar *et al.* [GAZ 85] –, grammaire syntagmatique guidée par la tête (***HPSG*** – *Head-driven Phrase Structure Grammar*) – Pollard et Sag [POL 87, 94] –, grammaires d'arbres adjoints (***TAG*** – *Tree Adjoining Grammar*) – Joshi et Shabes [JOS 87, 91] –, le modèle GPSG servant en fait d'étape pour la mise au point du modèle HPSG. Nous rappelons ci-dessous les principes généraux communs de ces modèles : la réinterprétation et l'enrichissement des règles de réécriture, l'utilisation des structures de traits complexes comme nouveau mode de représentation syntaxique (à la place de la représentation d'arbres syntaxiques classiques), l'ajout des principes de bonne formation linguistique, et l'articulation explicite des descriptions syntaxiques, lexicales et sémantiques.

La réinterprétation et l'enrichissement des règles de réécriture :

Comme nous l'avons mentionné à la section précédente (4.2.1.1), les grammaires de réécriture ont été critiquées par Chomsky pour leur incapacité à capturer des phénomènes linguistiques. Chomsky a donc proposé d'utiliser des règles syntagmatiques (hors contexte ou contextuelles) uniquement pour générer la structure profonde d'une phrase, et ensuite des règles de transformation pour aboutir à sa description de surface.

Les grammaires d'unification, en réaction au modèle transformationnel, ont réhabilité la description directe de surface en enrichissant les règles de réécriture par des contraintes sur les ***traits*** associées aux symboles auxiliaires et en réinterprétant ces règles non comme des procédures de dérivation mais comme des descriptions de structures syntaxiques bien formées.

L'utilisation des traits permet d'éviter la multiplication de règles et de symboles auxiliaires des grammaires hors contexte ou contextuelles. Par exemple, pour décrire l'accord en genre et en nombre entre le déterminant (D) et le nom (N) d'un groupe nominal (GN), dans une grammaire hors contexte « classique » on doit multiplier par quatre les symboles auxiliaires et les règles de réécriture :

GN1 $\rightarrow$ D1 N1 (D1 = déterminant masculin singulier, N1 = nom masculin singulier)

GN2 $\rightarrow$ D2 N2 (D2 = déterminant féminin singulier, N2 = nom féminin singulier)

GN3 → D3 N3 (D3 = déterminant masculin pluriel, N3 = nom masculin pluriel)

GN4 → D4 N4 (D4 = déterminant féminin pluriel, N4 = nom féminin pluriel)

Avec l'utilisation de deux traits (attributs) Genre (pour le genre) et Num (pour le nombre) associés aux catégories, d'une part, une seule règle de réécriture est nécessaire :

GN → D N <D Num> = <N Num>; <D Genre> = <N Genre>

D'autre part, cette règle peut être réinterprétée non comme une procédure de dérivation mais comme une description de structure syntaxique permettant de filtrer les sous-arbres compatibles avec la grammaire considérée (*cf.* Figure 4-1).

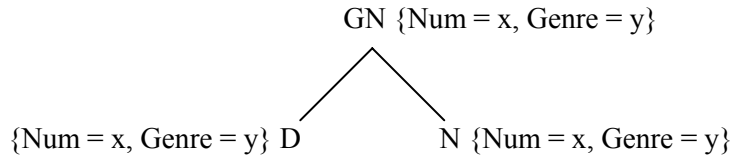


Figure 4-1 Description du groupe nominal avec les structures de traits

Les structures de traits complexes comme nouveau mode de représentation syntaxique :

Dans [KAY 79], Martin Kay a introduit l'idée d'utiliser les structures de traits complexes à la place des représentations syntaxiques arborescentes. Ce nouveau mode de représentation permet d'annoter directement des informations linguistiques qui sont difficilement représentables dans les arbres, comme par exemple les fonctions grammaticales (*cf.* l'exemple dans la Figure 4-2 [51]).

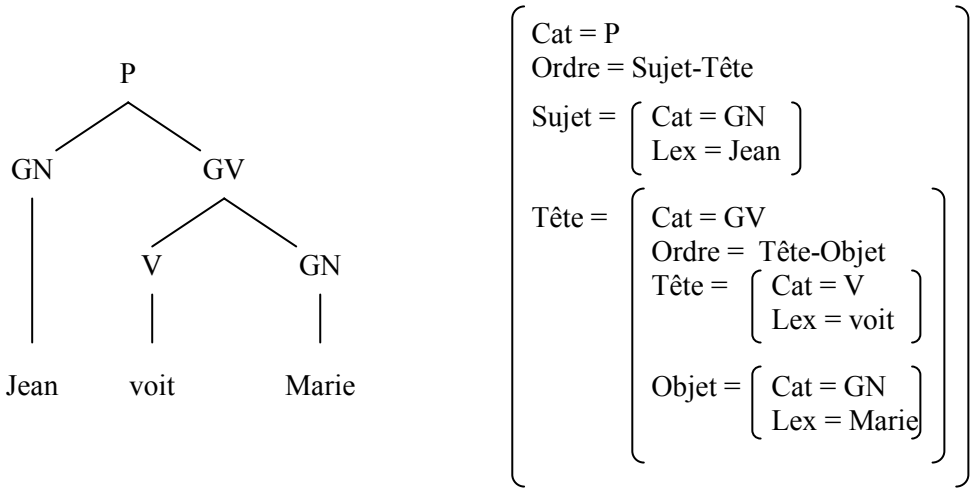


Figure 4-2 Arbre et structure de traits complexe

Chaque grammaire d'unification développe un usage particulier des structures de traits. Le modèle LFG les emploie pour formaliser la notion de fonction grammaticale. Il dispose d'une « structure fonctionnelle » parrallèle à l'arbre de dérivation syntagmatique. Le modèle GPSG utilise les traits pour vérifier des contraintes et exprimer des dépendances entre éléments situés à des niveaux syntagmatiques différents. Les structures de traits complexes remplacent complètement les représentations arborescentes dans le modèle HPSG. Dans le modèle TAG, enfin, elles ne sont pas utilisées en tant que telles, mais la description arborescente dont les noeuds sont occupés par des structures de traits atomiques est une représentation équivalente.

L'ajout des principes de bonne formation linguistique :

[51] Pour une explication plus détaillée, voir [ABE 93] - page 23.

Il s'agit des principes qui par exemple imposent la présence d'un trait, ou concernent la cooccurrence ou la propagation de traits syntaxiques. Ces principes permettent de définir une palette de cas de malformation linguistique.

L'articulation explicite des descriptions syntaxiques, lexicales et sémantiques :

Dans les grammaires d'unification, les entrées lexicales sont enrichies par des propriétés syntaxiques spécifiques codées sous forme de traits. L'intégration du lexique et de la syntaxe peut aller jusqu'à a lexicalisation intégrale de la grammaire comme dans le modèle TAG (LTAG – *Lexicalized TAG*).

Au niveau d'interprétation sémantique, certaines grammaires d'unification reprennent des modèles sémantiques construits indépendamment (GPSG, HPSG), d'autres modèles développent leur propre mode de représentation sémantique.

La compatibilité entre les informations syntaxiques, lexicales et sémantiques est vérifiée naturellement grâce à l'unification des structures de traits correspondantes.

Nous terminons cette section par un tableau comparatif des trois modèles les plus actifs LFG, HPSG et TAG proposé dans Abeillé [ABE 93] sur les critères suivants : la représentation des catégories, les types de traits employés, les principes de combinaison, la représentation des analyses, l'expression des généralisations syntaxiques, les représentations sémantiques et leurs équivalents mathématiques.

	LFG	HPSG	TAG
Catégories	Atomes	Structure de traits typés	Arbres élémentaires avec « tête » lexicale
Traits syntaxiques enchâssés	Structures fonctionnelles	Structure de traits typés	non utilisés
Principes de combinaison	Règles hors contexte (avec annotations fonctionnelles) et principes généraux	Structures de traits (schémas DI) et principes généraux	Adjonction ou substitution et unification
Résultat d'une analyse	Structure en constituants et structure fonctionnelle	Structure de traits typée	Arbre syntagmatique et arbre de dépendance
Représentations sémantiques	Structure argumentale et structure sémantique	Structure de traits typée (sémantique situationnelle)	Arbres sémantiques (synchrones)
Généralisations syntaxiques	Règles lexicales	Règles lexicales	Règles lexicales (familles d'arbres élémentaires)
Équivalent mathématique	Grammaires contextuelles	Grammaires non contraintes	Grammaires « légèrement contextuelles »

Nous avons rappelé ci-dessus les caractéristiques très généraux des formalismes de grammaire, qui décident la modélisation des règles grammaticales d'une langue pour un système d'analyse syntaxique. Dans la section suivante, nous discutons rapidement du développement et de l'évaluation des systèmes d'analyse syntaxique.

4.2.2. Systèmes d'analyse syntaxique et évaluation

Lors du développement d'un système d'analyse syntaxique d'une langue naturelle, deux aspects sont à évaluer constamment : la performance de l'analyseur et la couverture de la grammaire.

4.2.2.1. Systèmes d'analyse syntaxique

Le développement d'un analyseur reste un compromis entre finesse et efficacité : les analyseurs à description syntaxique superficielle peuvent être rapides, ce qui est difficile à obtenir pour les systèmes reposant sur des grammaires à large couverture proposant des analyses détaillées.

Les formalismes de grammaire d'unification donnent lieu à une réelle séparation entre données et programmes, ce qui permet d'utiliser le même analyseur pour des langues ou des grammaires différentes. Cela n'était pas le cas pour les analyseurs basés sur les réseaux de transition ou sur les grammaires chomskyennes. Cette séparation nous permet, durant le travail de thèse, de nous concentrer sur la construction des données, sans avoir besoin de développer des algorithmes d'analyse syntaxique.

La performance des analyseurs syntaxiques peut-être limitée à cause des raisons connues suivantes :

- La couverture lexicale insuffisante
- La couverture syntaxique insuffisante
- La technique d'analyse utilisée
- Le langage de représentation (formalisme) utilisé.

Nous passons maintenant aux méthodes d'évaluation des analyseurs.

4.2.2.2. Méthodes d'évaluation

EAGLES[52] (*Expert Advisory Group on Language Engineering Standards*) distingue trois types d'évaluation des systèmes de TAL suivants (*cf.* Balkan [BAL 94]) :

- Évaluation diagnostique, visant à localiser les déficiences du système ;
- Évaluation progressive pour mesurer l'évolution au fur et à mesure des étapes de développement du système ;
- Évaluation de l'adaptation du système aux spécifications qui ont été établies.

Pour les tests d'un système de TAL, Balkan [BAL 94] aborde également de la distinction de trois types de matériels :

- Corpus de test : c'est-à-dire une collection de textes réels agrandis, sous forme électronique ;
- Ensemble des phrases de test (*cf.* projet TSNLP - *Test Suites for Natural Language Processing*) : Il s'agit d'une collection de données construites (souvent) artificiellement, dont chaque donnée est créée pour vérifier le comportement du système vis à vis un phénomène spécifique. Ces données peuvent être des phrases, des segments de phrase, voire des séquences de phrases ;
- Collections de test : Un ensemble de données d'entrée, associé à un ensemble correspondant de résultats visés.

Les mesures d'évaluation sont habituellement deux mesures statistiques :

[52] Le projet EAGLES, établi en 1993, a pour but à améliorer les méthodes d'évaluation en vue d'élaborer des standards pour les produits de l'ingénierie des langues.

- la précision, c'est-à-dire le rapport du nombre d'analyses proposées correctes sur le nombre d'analyses proposées ;
- le rappel, c'est-à-dire le rapport du nombre d'analyses proposées correctes sur le nombre d'analyses correctes de référence.

Un exemple de la réalisation d'une évaluation des analyseurs syntaxiques est le projet EASY, introduit dans le paragraphe suivant.

4.2.2.3. Exemple d'évaluation : Projet EASY

L'évaluation des analyseurs permet d'une part à ceux qui les utilisent de connaître leurs forces et leurs faiblesses, et d'autre part à ceux qui les développent de disposer d'une référence à laquelle ils peuvent se confronter. Cependant, l'évaluation des analyseurs syntaxiques n'est en rien une tâche facile et ce pour diverses raisons : les sorties des analyseurs varient dans leur forme et leur nature d'un analyseur à l'autre, et même en parvenant à un accord sur ces sorties, des métriques équitables et représentatives ne sont pas actuellement disponibles. Ainsi, a été mis au point un protocole complet d'évaluation des analyseurs syntaxiques, baptisé PEAS, et dont la volonté était de n'écarter la participation d'aucun analyseur (*cf.* Vilnat *et al.* [VIL 04]). Ce protocole PEAS a servi de pré-projet au projet EASY (Évaluation des Analyseurs SYntaxiques) qui rassemble 5 fournisseurs de corpus et 14 analyseurs participants.

Le protocole mis au point dans PEAS puis EASY est articulé autour des modules suivants :

1. La constitution d'un corpus d'un million de mots, composé de textes hétérogènes en genre : des articles de journaux, des extraits de romans, des recueils de questions, des transcriptions de l'oral, des extraits de sites Internet. Et son formatage : normalisation, *tokenization* et découpage en phrases.
2. L'annotation d'un sous-ensemble de 20 000 mots. Elle est effectuée à l'aide d'un éditeur HTML et les résultats sont transcrits dans un format XML.
3. L'analyse par les participants et la transcription des sorties de leur analyseur dans un format XML commun.
4. L'évaluation, prévue actuellement comme un calcul du rappel et de la précision sur les constituants et les relations, est toujours ouverte à de nouvelles propositions.

Nous avons jusqu'ici introduit les bases méthodologiques qui nous permettent d'orienter le plan de travail de la thèse : il s'agit de la construction des données nécessaires pour l'analyse syntaxique et son évaluation.

4.2.3. Plan de la présentation

Nous présentons dans la suite notre travail concernant l'analyse syntaxique du vietnamien. Les problèmes abordés sont les suivants :

- modélisation de la grammaire vietnamienne, dans le but de construire une grammaire à large couverture pour le vietnamien. Pour cela, nous choisissons le formalisme TAG. Nous essayons d'abord de modéliser les groupes nominaux, puis de parcourir les phénomènes syntaxiques spécifiques du vietnamien. Nous parlons également de la construction du lexique syntaxique ;
- construction des ressources linguistiques pour tester et évaluer les systèmes d'analyse syntaxique du vietnamien : un corpus arboré, un ensemble de phrases de test.

Avant de décrire en détail notre travail, nous exposons les notions concernant le formalisme de grammaire et le système d'analyse syntaxique que nous utilisons pour notre projet.

4.3. Formalisme et outils utilisés : LTAG et LLP2

Dans cette section, nous introduisons d'abord les notions de base du formalisme TAG (*Tree Adjoining Grammar*) que nous avons choisi pour modéliser la grammaire vietnamien. Puis nous présentons le système d'analyse syntaxique développé pour les grammaires LTAG (*Lexicalised Tree Adjoining Grammar*) à l'équipe Langue et Dialogue au LORIA.

4.3.1. TAG – formalisme choisi

4.3.1.1. Choix de formalisme

Le formalisme que nous avons choisi dans notre projet pour modéliser la grammaire vietnamienne est LTAG, qui a été bien étudié pour les grammaires anglaise et française (*cf.* [XTA 01], Abeillé [ABE 02]). Ce choix est dû à plusieurs facteurs. Théoriquement, l'interface syntaxe sémantique y est plus simple que dans les grammaires hors-contexte de par le domaine de localité syntaxique étendu proposé par les TAGs, et la complexité au pire pour l'analyse des TAGs reste polynomiale (en $O(n^6)$). D'un point de vue pratique, les outils génériques pour les analyseurs basés sur TAG sont nombreux (par exemple XTAG, Dyalog) et également bien développé au Loria (Crabbé *et al.* [CRA 03]). Un format normalisé pour les ressources syntaxiques est disponible : TAGML (Bonhomme and Lopez, [BON 00b]). De plus, la possibilité de convertir une grammaire du formalisme TAG à un autre formalisme est ouverte (*cf.* Yoshinaga *et al.* [YOS 03]). Ces conditions nous permettent de choisir le formalisme TAG pour construire une grammaire électronique du vietnamien. Ce choix est encore plus motivé par les caractéristiques suivantes de TAG (*cf.* le paragraphe suivant sur les notions de base de TAG) :

- les structures de traits en TAG ne contiennent que des valeurs atomiques,
- la représentation arborescente des structures syntaxiques est plus lisible.

Ces caractéristiques, en simplifiant le travail de modélisation, permettent de focaliser la recherche sur la langue étudiée elle-même plus que sur les spécificités du formalisme employé.

Nous rappelons maintenant les notions de base du formalisme TAG.

4.3.1.2. Notions de base du formalisme TAG

Une grammaire TAG est constituée d'***arbres élémentaires*** (arbres initiaux et arbres auxiliaires), la construction d'arbres d'analyse syntaxique étant réalisée grâce à trois opérations: ***l'adjonction, la substitution et l'unification***.

Les arbres initiaux sont des structures linguistiques minimales non récursives (structures syntagmatiques de phrases simples) :

- tous les nœuds internes sont étiquetés par des non-terminaux[53] ;
- tous les nœuds-feuilles sont étiquetés par des terminaux, ou par des nœuds non-terminaux marqués pour la substitution (*cf.* Figure 4-3).

Les arbres auxiliaires sont des structures récursives représentant des constituants adjoints (des modifieurs) aux structures de base :

- tous les nœuds internes sont étiquetés par des non-terminaux ;

[53] *cf.* GLOSSAIRE : Terminal

- tous les nœuds feuilles sont étiquetés par des terminaux ou par des nœuds non-terminaux marqués pour la substitution, à l'exception du nœud-pied, qui a le même nom que la racine, marqué pour l'adjonction (*cf.* Figure 4-3).

Figure 4-3 L'arbre initial et l'arbre auxiliaire

À chaque nœud peut être associée une structure de traits spécifiant comment les nœuds interagissent entre eux. Les structures de traits ont deux parties (1) l'amont (*top*), qui contient l'information sur le nœud supérieur et (2) l'aval (*bottom*), qui contient l'information sur le nœud inférieur.

Pour la ***substitution***, le nœud-racine d'un arbre élémentaire est combiné avec un nœud-feuille non-terminal marqué pour la substitution (le nœud racine et le nœud de substitution doivent avoir le même nom, *cf.* Figure 4-4).

Pour l'***adjonction***, un arbre auxiliaire est greffé sur un nœud non-terminal n'importe où dans un arbre élémentaire (ce nœud et le nœud racine de l'arbre auxiliaire doivent avoir le même nom, *cf.* Figure 4-5).

Les opérations de substitution et d'adjonction opèrent également sur les structures de traits, en réalisant l'***unification*** (*cf.* Figure 4-4 et Figure 4-5[54]), qui permet de spécifier dynamiquement les contraintes locales, et non statiquement à l'intérieur des arbres.

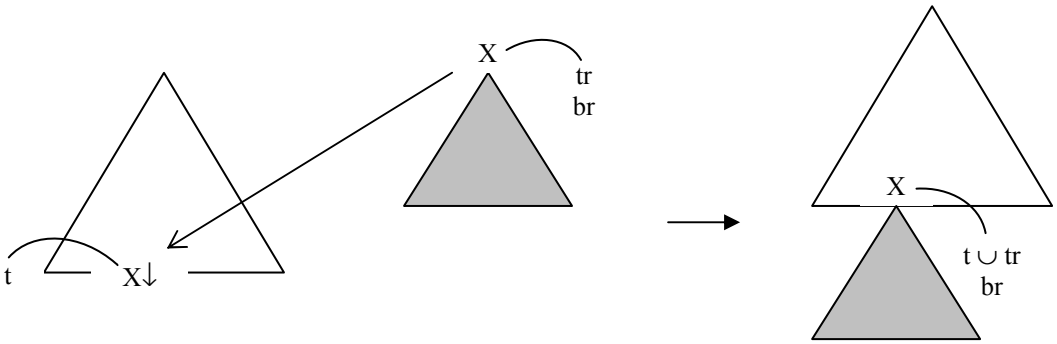


Figure 4-4 La substitution et l'unification des traits

[54] tr, br = les traits amont (t), aval (b) de la racine (r) ; tf, bf = les traits amont (t), aval (b) du nœud pied (f).

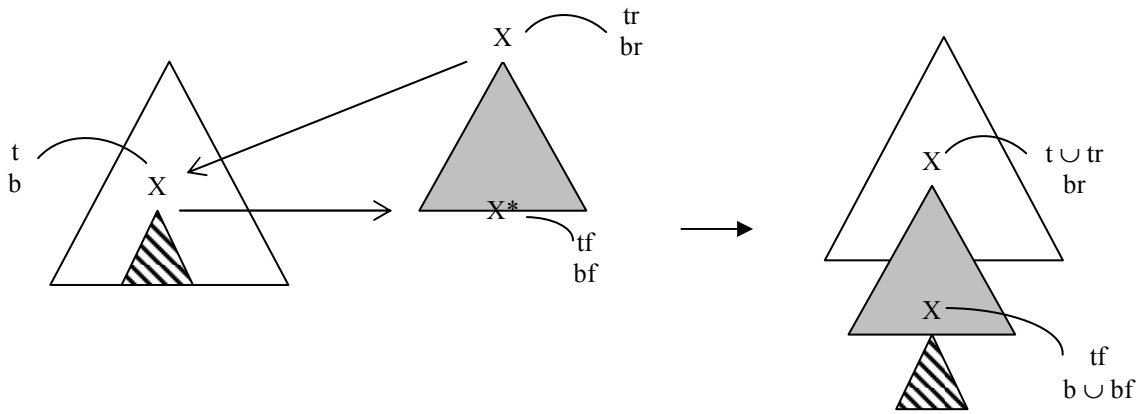


Figure 4-5 L'adjonction et l'unification des traits

La lexicalisation des grammaires permet de mieux guider les analyses par les propriétés syntaxiques spécifiques de chaque mot. Shabes *et al.* [SHA 88] définissent le modèle TAG ***lexicalisé*** (LTAG : *Lexicalized* TAG), dans lequel toute structure élémentaire a au moins une feuille d'ancrage lexical, occupée par un item lexical qui lui sert de tête (notée ◊).

L'analyse syntaxique produit comme résultats un ***arbre dérivé*** et un ***arbre de dérivation***[55].

Considérons l'exemple « *Jean dort* ». Nous ajoutons l'adverbe « *beaucoup* » pour illustrer l'opération d'adjonction.

Exemple simplifié d'une grammaire LTAG :

Arbres élémentaires lexicalisés (Figure 4-6): Deux arbres correspondants aux entrées « *Jean* » et « *dort* » sont des arbres initiaux, l'arbre de l'entrée « *beaucoup* » est auxiliaire.

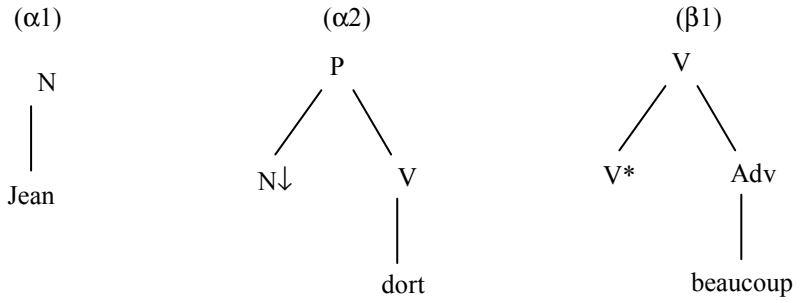


Figure 4-6 Exemples d'arbres élémentaires ([ABE 93])

Résultat d'analyse syntaxique (Figure 4-7): Arbre dérivé et arbre de dérivation correspondant à la phrase « *Jean dort beaucoup* ». Dans l'arbre de dérivation, la branche discontinue représente une opération de substitution, et la branche continue représente une adjonction. À chaque nœud sont associées les informations suivantes : nom de l'arbre participant à l'opération, adresse du nœud où l'opération a eu lieu, items lexicaux de tête de l'arbre.

[55] *cf.* GLOSSAIRE : arbre dérivé, arbre de dérivation

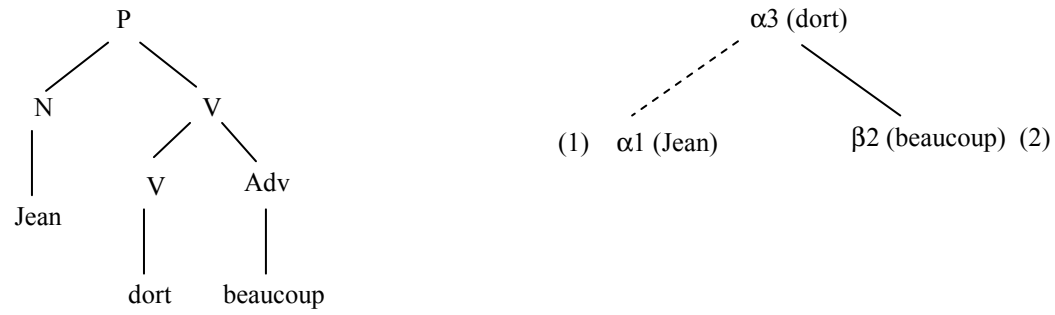


Figure 4-7 Exemples d'arbre dérivé et de dérivation en TAG ([ABE 93])

Les arbres dérivés en TAG correspondent à la notion d'arbre syntaxique dans d'autres grammaires syntagmatiques, alors que les arbres de dérivation sont à la base de l'interprétation sémantique. Dans une TAG, « ils font apparaître explicitement les relations de dépendances entre items lexicaux (têtes des arbres élémentaires), en particulier les relations prédicats/arguments qui peuvent être 'noyées' dans l'architecture syntagmatique de l'arbre dérivé. Ainsi les arguments sont toujours dominés directement par leur prédicat dans l'arbre de dérivation, alors qu'ils peuvent en être infiniment éloignés (en termes d'ordre des mots et de niveau de profondeur) dans l'arbre dérivé. » ([ABE 93]).

La capacité d'exprimer des relations sémantiques en TAG est encore beaucoup plus forte avec le mécanisme des **TAG synchrones** (*cf.* Shieber et Schabes [SHI 90]). Ce mécanisme permet de calculer la structure sémantique de la phrase, qui est un arbre où les nœuds sont les prédicats sémantiques associés aux mots et où les branches codent les phénomènes de dépendance sémantique ou de portée. Il s'agit de mettre en lien deux grammaires d'arbres lexicalisées, une syntaxique et une sémantique, en assurant une synchronisation des dérivations au fur et à mesure de chaque analyse.

La construction d'une grammaire lexicalisée LTAG doit respecter les **principes de bonne formation** des arbres élémentaires suivants (*cf.* Abeillé [ABE 93, 02]) :

- « principe d'**ancrage lexical** : tout arbre élémentaire a au moins une tête lexicale non vide ;
- principe de **cooccurrence prédicat-arguments** : tout prédicat contient dans sa structure élémentaire au moins un nœud pour chacun des arguments qu'il sous-catégorise (sous forme de nœud à substitution ou de nœud pied) ;
- principe d'**ancrage sémantique** : tout arbre syntaxique élémentaire a un correspondant sémantique non vide. Ceci exclut la plupart des éléments fonctionnels (prépositions « vide », complémenteurs, certains pronoms relatifs) en tant qu'entités autonomes de la syntaxe : ces éléments apparaissent comme co-têtes lexicales dans un arbre élémentaire ayant une tête lexicale non vide ;
- principe de **compositionnalité** : un arbre élémentaire correspond à une et une seule unité sémantique. »

Pour finir ce paragraphe, nous introduisons la notions de règle lexicale.

Les **règles lexicales**, utilisées par la plupart des grammaires d'unification, représentent des régularités syntaxiques et sémantiques en mettant en relation des ensembles d'entrées lexicales (généralement les formes verbales). Elles contraignent ainsi la bonne formation du lexique d'une langue : par exemple, si celui-ci contient telle forme verbale active, il devra également contenir une forme verbale passive de même sens, ayant telle construction apparentée (Abeillé [ABE 93, 02]). En TAG, une règle lexicale s'applique à tous les arbres élémentaires dont la description s'unifie avec la description partielle de la partie gauche de la règle.

Du point de vue linguistique, deux types de règles lexicales sont distingués (Abeillé [ABE 02]) : le premier désigne les règles de réalisation des arguments d'un prédicat (phénomènes d'extraction, les réalisations non canoniques ou les variations d'ordre des mots), qui ne changent ni le sens de l'expression ni la sous-catégorisation du prédicat, le second, les règles de redistribution fonctionnelle (phénomènes d'alternance ou de changement de valence).

La section suivante présente la réalisation d'un analyseur LTAG au LORIA.

4.3.2. LTAG à l'équipe Langue et Dialogue

4.3.2.1. Point de vue général

En théorie, les grammaires lexicalisées LTAG sont supposées décrire les propriétés syntaxiques spécifiques de chaque mot. En pratique, on peut optimiser la taille de la grammaire en prenant en considération les possibilités suivantes (*cf.* Crabbé [CRA 03]) :

- Pour les langues flexionnelles, afin d'éviter de multiplier le nombre d'arbres élémentaires en réalisant l'ancrage sur les formes fléchies, on ramène les formes fléchies à leur lemme. En conséquence, on aura une base de données associant à chaque forme fléchie son lemme et une structure de traits morphologique, et une base d'arbres élémentaires dont les ancres sont des lemmes.
- De nombreux lemmes partagent des comportements syntaxiques similaires. On peut donc séparer les lemmes des arbres élémentaires. Les arbres élémentaires ainsi obtenus (appelés schèmes) sont mis en relation avec des lemmes qui peuvent substituer leurs ancres. À chaque association lemme/schème, on ajoute une liste de structures de traits imposées sur les nœuds arguments sélectionnés par le lemme dans le schème.
- Les arbres élémentaires qui ont des relations exprimables par des règles lexicales peuvent être organisés en familles d'arbres (*cf.* XTAG [XTAG 01] et Abeillé [ABE 02]).
- La méta-grammaire permet une description compacte : les schèmes sont factorisés afin de partager au maximum les descriptions partielles en commun (*cf.* Figure 4-8).

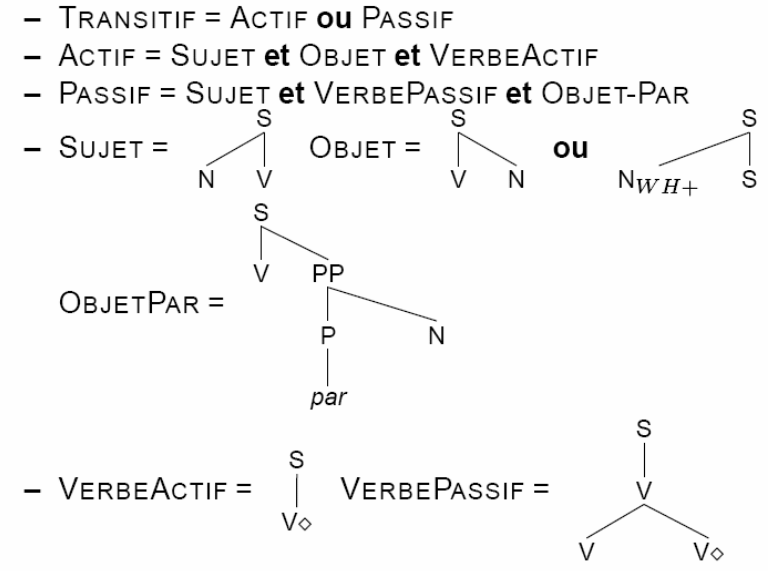


Figure 4-8 Exemple de factorisation de schèmes (*cf.* Crabbé *et al.* [CRA 03, 05])

Le format employé pour le codage des données syntaxiques par l'analyseur LTAG au LORIA est TAGML, format défini en commun avec le projet ATOLL de l'INRIA Rocquencourt et l'équipe TALANA de Paris 7. Nous décrivons maintenant les principes de ce mode de représentation.

4.3.2.2. Format des ressources : TAGML

TAGML (*cf.* Bonhomme et Lopez [BON 00b]) est un formalisme de spécification XML des divers éléments d'une grammaire LTAG.

Un fichier de descriptions d'une grammaire se compose de trois parties. La première partie est une base lexicale, qui comprend les descriptions des unités morphologiques. Par exemple, le code suivant correspond au classificateur des objets « `cái` » en vietnamien :

```
<morph lex="cái">
    <fs>
        <f name="type">
            <sym value="common"/>
        </f>
        <f name="count">
            <sym value="absolute"/>
        </f>
        <f name="unit">
            <sym value="natural"/>
        </f>
        <f name="meaning">
            <sym value="object"/>
        </f>
    </fs>
    <lemmaref name="cái" cat="N"/>
</morph>
```

Dans cet exemple, la structure de traits (**fs**) est constituée de 4 traits (**f**) avec les noms d'attribut et les valeurs balisées par **sym**. Le lemme correspondant au morphème et sa partie du discours sont marqués par le balisage **lemmaref**.

Une deuxième partie spécifie les arbres élémentaires de la grammaire. Par exemple, le code suivant correspond à l'arbre élémentaire décrivant la sous-catégorisation d'un verbe transitif :

```
<tree id="NP0-Vtransitive-NP1">
    <fs>
        <f name="alt"> <sym value="NP0-Vtransitive-NP1"/> </f>
    </fs>
    <node cat="S" type="std">
        <node cat="NP" type="subst" name="np0"/>
        <node cat="VP" type="std" name="vp">
            <node cat="V" type="anchor" name="verb">
                <narg type="top">
                    <fs>
                        <f name="transitivity">  <sym value="transitive"/>   </f>
                    </fs>
                </narg>
            </node>
            <node cat="NP" type="subst" name="np1"/>
        </node>
    </node>
</tree>
```

Le type des nœuds est noté **subst** pour les nœuds de substitution, **anchor** pour les ancres, **adj** pour les adjonctions et **std** pour les autres nœuds.

Et enfin la troisième partie décrit les lexicalisations des arbres. Par exemple, une lexicalisation de l'arbre ci-dessus, avec la lexicalisation de l'ancre (**anchor**) dont le nom (**noderef**) est « *verb* » :

```
<lexicalization>
    <tree><fs>
            <f name="alt">   <sym value="NP0-Vtransitive-NP1"/>   </f>
    </fs></tree>
    <anchor noderef="verb">
        <lemmaref name="ăn" cat="V"/> <!-- manger -->
    </anchor>
</lexicalization>
```

On note deux méthodes de lexicalisation. La première consiste à associer à un nœud précis d'un arbre donné un lemme et sa catégorie syntaxique. La deuxième méthode, propre au TAGML, utilise un *filtrage dynamique* (*cf.* Crabbé *et al.* [CRA 03]), dans laquelle on énonce des règles, décrites par une structure de traits, qui visent à sélectionner pour un lemme donné l'ensemble des arbres qui répondent à certains critères.

Toutes les descriptions dans chaque partie peuvent être organisées en bibliothèques.

4.3.2.3. Analyseur LLP2

L'analyseur LLP2 (*Loria LTAG Parser 2*), un analyseur syntaxique pour les grammaires d'arbres adjoints, est développé et maintenu au sein de l'équipe Langue et Dialogue (LeD) du LORIA. Cependant, un analyseur syntaxique n'est en pratique qu'un élément d'une chaîne plus complexe : outre les indispensables prétraitements nécessaires à l'analyse syntaxique (tels que l'annotation morphologique), il apparaît depuis quelques années qu'une des principales difficultés est la maintenance de la grammaire. Les travaux au sein de LeD portent donc également sur des outils de maintenance et d'exploration de grammaires d'arbres adjoints. LLP2 est constitué de la collection de modules spécialisés suivants :

- ***parser*** : le parseur pour le formalisme LTAG ;
- ***tagviewer*** : le visualiseur des arbres élémentaires ;
- ***graphtag*** : gestionnaire de visualisation des arbres TAG ;
- ***tagml2*** : API gérant l'entrée/sortie sous format TAGML2 des ressources lexicales et syntaxiques ;
- ***segment*** : API gérant l'entrée/sortie sous format XML du prétraitement de textes ;
- ***FeatureStructure*** : API gérant les structures de traits ;
- et d'autres APIs aidant à gérer les différents types d'arbres (arbres élémentaires, arbres dérivés, arbres de dérivation).

L'analyseur avec sa documentation est disponible à http://www.loria.fr/~azim/LLP2/help/fr/.

L'utilisation de cet analyseur nous permet de tester notre modélisation de grammaire vietnamienne, qui est l'objet de la section suivante. Le format des données TAGML2 assure une extraction facile des informations enregistrées dans la grammaire construite.

4.4. Descriptions syntaxiques du vietnamien

Dans le cadre de cette thèse, il nous est impossible de réaliser une modélisation complète de la grammaire vietnamienne. Cependant, nous essayons d'avoir une vue assez détaillée sur les phénomènes syntaxiques du vietnamien, dans un double objectif :

- prévoir une première liste des phénomènes à modéliser et préparer des jeux de phrases de test
- constater la faisabilité d'une représentation de ces phénomènes employant le formalisme TAG.

Nous concrétisons seulement la représentation des groupes nominaux en TAG. Les autres phénomènes font l'objet à la section 4.4.2 d'une description et d'une discussion sur les pistes à suivre pour leur analyse.

Les notations et les conventions utilisées dans cette section sont les suivantes :

- Dans les exemples linguistiques :
  - * : agrammatical
  - ? : douteux
  - (X) : X est optionnel
- Catégories morphosyntaxiques : N – Nom ; V – Verbe ; A – Adjectif ; P – Pronom ; D – Déterminant ; M – Numéral ; Adv – Adverbe ; Prep – Préposition ; Conj - Conjonction ; I - Interjection ; T - Mot modal.
- Groupes syntagmatiques : S – phrase, NP – groupe nominal, VP – groupe verbal, AP – groupe adjectival, PrepP – groupe prépositionnel, PredP – groupe prédicatif

4.4.1. Description en TAG du groupe nominal vietnamien

Dans cette section, nous fondons notre étude sur les descriptions détaillées des groupes nominaux du vietnamien publiées par Nguyễn T. Cẩn [NGU 98b] et le Comité des Sciences Sociales [UYB 83]. Dans le cadre de cette première étude, nous ne nous ne considérons pas le cas des noms propres.

La structure complète d'un groupe nominal en vietnamien est donnée par la séquence :

$C_1\ C_2\ C_3\ N_1\ N_2\ C_4\ C_5$,

où

- **C_1** est un pronom de totalité (par ex. tất cả / *tous*) ;
- **C_2** est un quantificateur (numéral ou déterminant) ;
- **C_3** est le pseudo-article optionnel cái, qui joue le rôle spécial de particule intensive ;
- **N_1** est un nom désignant une unité, c'est-à-dire un classificateur (*cf.* les sections 2.4.1.1 et 3.3.2.1) ou une unité de mesure ;
- **N_2** est un nom désignant le concept dominant ce groupe nominal (N_1 est une unité de mesure ou un classificateur de N_2). N_2 est donc la "tête sémantique" du syntagme ;
- **C_4** est une suite de compléments, chacun de ceux-ci pouvant être un nom, un adjectif, un verbe – ou leur syntagme respectif –, une préposition ou un nombre ;
- **C_5** est un pronom démonstratif.

Voici quelques exemples permettant d'illustrer cette structure :

1) Un groupe nominal présentant la structure complète

- tất cả [C_1] năm [C_2] cái [C_3] quyển [N_1] sách [N_2] cũ [C_4] này [C5]

  =$_{litt.}$ *tout* | *cinq* | <*particule intensif*> | <*classificateur*> | *livre* | *vieux* | *ce*

  = *tous ces cinq vieux livres-là*[56]

2) Un groupe nominal avec un seul élément

- sách [N_2] = *livre* (détermination générique)

3) Un groupe nominal sans N_2

- năm [C_2] quyển [N_1] cũ [C_4] này [C_5]

  =$_{litt.}$ *cinq* | <*classificateur-livre*> | *vieux* | *ce*

  = *ces cinq vieux livres*

Nous analysons maintenant plus en détail les structures possibles du groupe nominal.

4.4.1.1. Structures possibles du groupe nominal

Quand N1 et N2 sont simultanément présents, le noyau syntagmatique nominal peut être N1 ou N2 : il n'existe pas à ce sujet de consensus entre les linguistes. Notre rôle n'est pas de prendre position dans ce débat, c'est pourquoi nous nous contentons de donner dans le Tableau 4-2 ci-dessous l'ensemble des séquences de mots possibles pouvant former un groupe nominal. Le tableau mentionne le combinaisons des constituants C1, C2, N1, N2, C4, C5, sans tenir compte de la particule modale C3 qui peut, dans la plupart des cas, s'ajouter avant les deux éléments N1 et N2.

On distingue deux types de noms d'unité, pouvant se trouver à la position N_1 :

- les unités naturelles, aussi appelées *classificateurs*, spécifiques à certaines langues asiatiques, qui désignent N_2 en tant qu'individu (*un « fruit-orange »*). Ces noms se composent :
  - o des noms utilisés pour désigner l'homme (par ex. người = mot désignant l'homme en général, thằng = mot placé devant certains noms désignant des personnes de rang inférieur ou de conduite indésirable, *etc.*)
  - o des noms utilisés pour les objets et les concepts abstraits qui classifient la nature d'objet. Par exemple, quả est un classificateur pour les fruits, cây pour les arbres, et cam signifie *orange*, alors :
    - một quả cam =$_{litt.}$ *un* <*class-fruit*> *orange* = *une orange*
    - một cây cam =$_{litt.}$ *un* <*class-arbre*> *orange* = *un oranger*
- les unités conventionnelles, communes à toutes les langues, qui considèrent N_2 comme « matériau » (*un kilo d'oranges*), et se composent :
  - o des unités de mesure exactes (par ex. kilogramme, mètre, *etc.*)
  - o des unités de mesure inexactes (par ex. morceau, classe, poignée, *etc.*)

[56] Nous traduisons par "-là" la particule intensive "cái", même si la correspondance n'est pas tout à fait littérale.

	P (tout-C1)	D/M (quantité-C2)	N1 (unité)	N2	N/V/A/M/PrepP (modifieurs-C4)	P (ce-C5)
1	?	-	-	+ (sách [livre] / gạo [riz])	?	?
2	?	?	+ (quyển/ cân/thùng)	+ (sách [livre])	?	?
3	?	?	+ (cân/thùng)	+ (gạo [riz])	?	?
4	?	?	+ (quyển/ cân/thùng)	ε	?	?
5	?	?	-	+ (ý kiến [opinion])	?	?
6	-	? (D)	-	+ (khi [fois])	?	?
7	?	?	?	+ (học sinh [élève])	?	?
8	?	? (grand nombre)	-	+ (quần áo [vêtements])	?	?
9	-	-	-	+ (trước [avant])	-	?
10	+	+ (M)	ε	ε	-	-
11	-	+ (M)	ε	ε	?	?
12	-	+ (M)	ε	+ (bàn [tables])	?	?
13	-	+ (M)	-	+ (tôm [crevettes])	-	-
14	?	?	+ (việc [fait])	-	?	?

+ : présence obligatoire, ? : présence optionnelle, - : présense interdite, ε : élément extrait

Tableau 4-2 Constituants d'un groupe nominal

C_2 est un quantificateur qui peut-être :

- un numéral correspondant à une quantité exacte, par exemple hai (*deux),* ba *(trois)*…
- un numéral correspondant à une quantité approximative comme vài *(quelques*), mươi *(une dizaine)*…
- un déterminant correspondant à une distribution, par exemple mỗi *(chaque)*, mọi *(tout)*…
- un déterminant de nombre singulier ou pluriel, par exemple những *(des/les),* các *(des/les),* một *(un)*.

Ce quantificateur peut être normalement ajouté devant un nom comptable. Cependant, il faut noter qu'il existe des cas irréguliers où un nombre peut être suivi par un nom non dénombrable ou par un mot appartenant à une autre catégorie syntaxique :

- dans les expressions : bảy nổi ba chìm =$_{litt.}$ *sept flotté trois noyé = très tourmentée* (en parlant d'une vie) (Tableau 4-2, ligne 11)

- au restaurant : hai tái =litt. *deux saignant = deux phở*[57] *saignants*, ou
  hai đen nóng =litt. *deux noir chaud = deux cafés noirs chauds* (Tableau 4-2, ligne 11)
- dans les formules de mélange : ba sôi hai lạnh =lit *trois bouillant deux froid = trois mesures d'eau bouillante pour deux mesures d'eau froide* (Tableau 4-2, ligne 11)
- dans les énumérations : chúng tôi cần hai bàn, sáu ghế, một tủ =litt. *nous avoir besoin de deux tables, six chaises, une armoire = nous avons besoin de deux tables, six chaises et une armoire*. (Tableau 4-2, ligne 12)

Nous considérons maintenant la partie droite de la tête sémantique (N2) : C_4 et C_5. C_5 est un pronom démonstratif qui se trouve toujours à la fin du syntagme nominal, son traitement est donc simple. En revanche, C_4 est une suite de modificateurs assez complexe. Nous précisons ci-dessous les modificateurs possibles du nom-noyau du syntagme nominal.

Un modificateur appartenant à C_4 peut être :

- un nom ou un groupe nominal, par exemple
  - một [C_2] cuốn [N_1] sách [N_2] toán [C_4] =litt. *un <class.> livre mathématiques = un livre de mathématiques* ;
- un adjectif ou un groupe adjectival, par exemple :
  - một [C_2] cuốn [N_1] sách [N_2] rất quý [C_4] =litt. *un <class.> livre très précieux* ;
- un verbe ou un groupe verbal, par exemple :
  - cái [C_3-N_1] bàn [N_2] kê trong góc [C_4] =litt. *<class.> table mettre dans coin = la table mise dans le coin*.
  - bàn [N2] học [C4] =litt *table apprendre = table de travail (« étude »)*. Le verbe ici montre l'objectif.
- un pronom, par exemple :
  - đầu [N_2] tôi [C_4] =litt. *tête je = ma tête* ;
- un numéral, par exemple :
  - ngày [N_1] 27 [C_4] =litt. *jour 27 = le 27* (date) ;
- une proposition, par exemple :
  - bức [N_1] thư [N_2] tôi viết [C_4] =litt. *<class.> lettre je écrire = la lettre que j'ai écrite* ;
- un groupe prépositionnel, par exemple :
  - nhà [N_2] của tôi [C_4] =litt. *maison de je = ma maison*. Le groupe prépositionnel se compose normalement d'une préposition suivie par un groupe nominal ou une proposition. Dans plusieurs cas, cette préposition peut-être enlevée et on revient à l'une des formes précédentes.

Ces modificateurs sont normalement ordonnés par longueur croissante, comme nous pouvons constater dans l'exemple à la Figure 4-9. Cet exemple montre la structure arborée d'un groupe nominal dont la traduction littérale est la suivante :

<class.> chat noir de maison ami Nam (que) je venir-de demander hier ce

= Ce chat noir de chez Nam que je viens de demander hier

[57] Soupe vietnamienne.

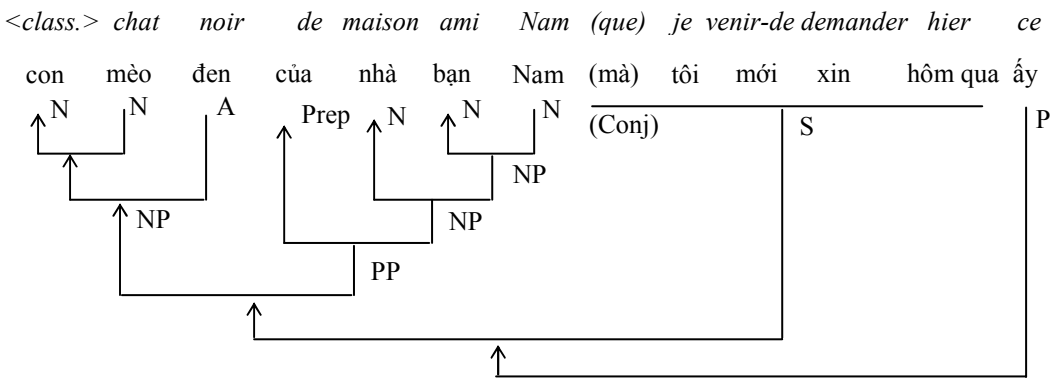


Figure 4-9 Exemple de structure arborée d'un groupe nominal

La présentation théorique exposée ci-dessus nous guide dans la représentation du groupe nominal dans le formalisme TAG, qui est le sujet du paragraphe suivant.

4.4.1.2. Représentation TAG du groupe nominal vietnamien

Nous avons vu dans les paragraphes précédents la structure des groupes nominaux en vietnamien avec une représentation plate. Nous proposons une représentation arborée de cette structure comme dans la Figure 4-10. Pour générer un tel arbre, une grammaire TAG doit contenir les arbres initiaux permettant de produire les différents types de noyaux syntagmatiques nominaux. Tous les autres constituants appartenant aux parties gauche et droite du noyau sont générés grâce aux arbres auxiliaires.

Les arbres initiaux correspondant aux différents types de noyau présentés à la section 4.4.1.1 se trouvent dans la Figure 4-11.

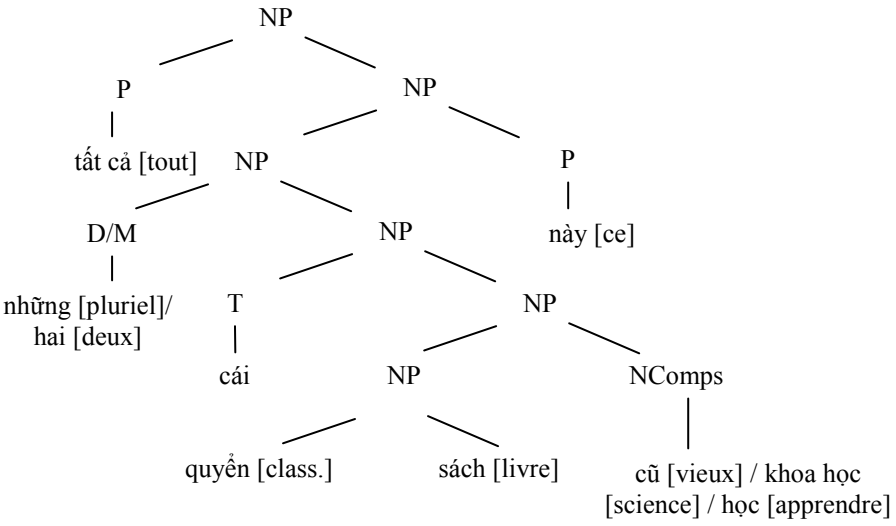


Figure 4-10 Structure arborée général du groupe nominal

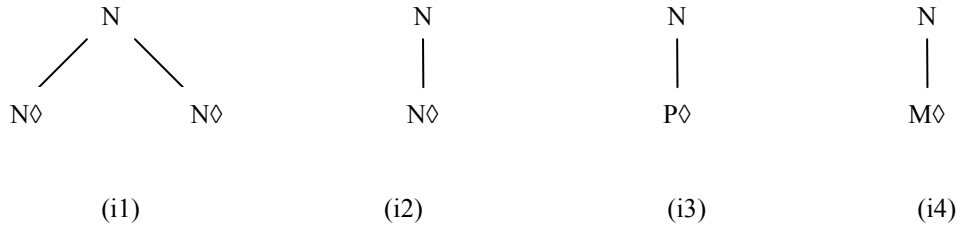


Figure 4-11 Arbres initiaux pour les groupes nominaux

(i1) et (i2) permettent de générer le noyau des groupes nominaux (N_1 ou N_2 ou les deux), dans la plupart des cas

(i3) correspond aux cas où le groupe nominal est représenté par un pronom

(i4) gère le cas où le numéral joue le rôle de tête syntagmatique à cause de l'omission du nom d'unité et de la tête sémantique

La Figure 4-12 montre les arbres auxiliaires permettant de générer les modifieurs d'un groupe nominal ; les opérations présentées sont :

- a1-a7 : adjonction de divers types de compléments en position C4,
  - a1 : N-N (`livre mathématiques` : *livre de mathématiques*)
  - a2 : N-A (`livre vieux` : *vieux livre*)
  - a3 : N-V (`livre apprendre` : *livre d'étude*)
  - a4 : N-M (`chambre trois` : *chambre numéro trois*)
  - a5 : N-P (possessif « elliptique » `chambre je` : *ma chambre*)
  - a6 : N-PrepP (en particulier possessif `chambre de je` : *ma chambre*)
  - a7 : N-mà-S ou N-S (proposition relative)
- a8 : adjonction du pronom démonstratif C5 (`livre ce`: *ce livre*)
- a9 : adjonction d'un déterminant ou adjectif numérique en position C2 (`trois livre` : *trois livres*)
- a10 : adjonction du pronom « *tout* » tất cả (`tout trois livres` : *tous les trois livres*).

Une liste d'attributs est définie pour exprimer les contraintes appliquées pour les opérations de substitution et d'adjonction.

Les attributs pour un groupe nominal :

head = *n1n2* (deux noms du noyau sont présents), *n1* (seul N1 est présent), *n2* (seul N2 est présent), *num* (les noms du noyau sont absents), *pron* (le groupe nominal est un pronom)

det = *generic* (sách), *definite* (quyển sách), *demonstrative* ([quyển] sách này)[58]

number = *plural*, *singular*, *neuter*

count = +, - (contrainte pour l'adjonction du déterminant ou numéral en position C2)

quant = +, - (contrainte pour l'adjonction du quantificateur total)

[58] inspiré de l'attribut de détermination pour le créole martiniquais de Vaillant [VAI 03]

Les attributs pour un nom :

Pour chaque nom, il est important de pouvoir déterminer les traits suivants :

- *classifier* qui enregistre la liste des classificateurs utilisables pour ce nom : classifier = {liste de classificateurs possibles du nom} ;
- *countable*, comme le trait défini dans le lexique morphosyntaxique (*cf.* 3.3.2.1) ;
- *unit* ;
- *meaning* ;

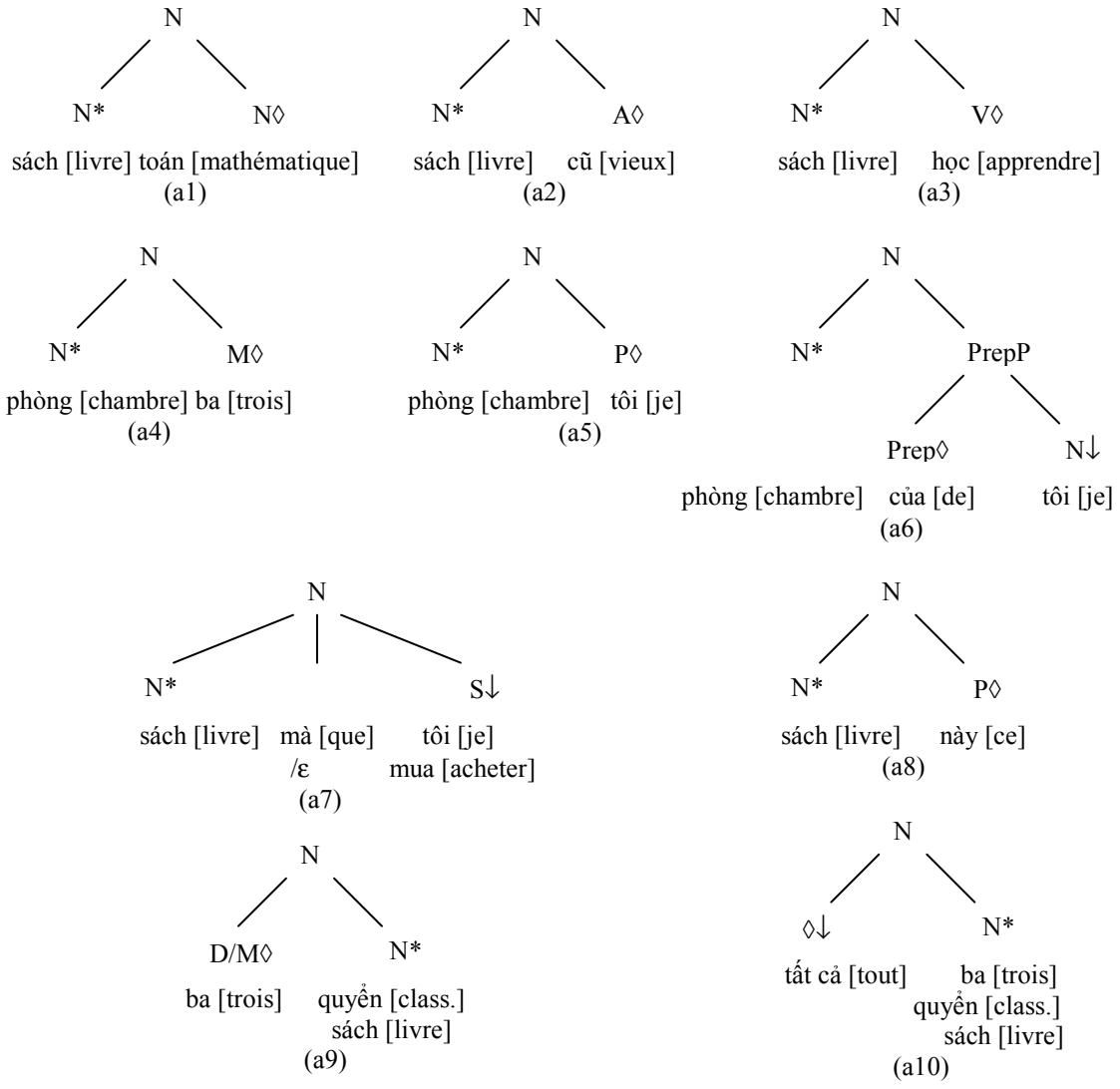


Figure 4-12 Arbres auxiliaires produisant les modifieurs du groupe nominal

Les arbres que nous avons conçus pour l'analyse des groupes nominaux ne sont pas encore exhaustifs, car ils couvrent seulement les règles grammaticales présentées dans la littérature. Nous avons fait le test de la grammaire avec l'analyseur LLP2 sur un jeu d'exemples simples. Le travail en perspective est de développer un jeu de test plus systématique (*cf.* 4.5.2).

Dans la section suivante, nous considérons les autres phénomènes syntaxiques en vue de constater la capacité de représentation de ces phénomènes avec le formalisme TAG.

4.4.2. Parcours des phénomènes syntaxiques à modéliser

Nous parcourons dans cette section les phénomènes syntaxiques du vietnamien en donnant des exemples illustratifs de chacun d'eux : sous-catégorisation des verbes et adjectifs, et composants complémentaires des phrases. Nous nous fondons principalement sur deux sources : un livre de grammaire vietnamienne du Comité de Sciences Sociales du Vietnam [UYB 83] et une référence récente du Département de Linguistique de l'Université Nationale de Hanoi (Nguyễn M. Thuyết et Nguyễn V. Hiệp [NGU 98a]) sur les composants de phrases en vietnamien.

4.4.2.1. Typologie des phrases

La structure nucléaire des phrases simples en vietnamien, analysées en constituants immédiats (CI), peut se composer de :

- un CI simple : Mưa =$_{litt}$ *Pluie = Il pleut.*
- deux CI simples :
  - (Tôi) (đọc sách) = (*Je*) (*lis des livres*)
- deux CI complexes :
  - Nếu ((trời) (mưa)) thì ((tôi) (đọc sách))
    S'((*il*) (*pleut*)), *alors* ((*je*) (*lis des livres*)).

Des phrases plus complexes peuvent ensuite être construite par coordination et juxtaposition de phrases complètes, mécanismes que nous n'abordons pas ici. De plus, nous nous limitons également aux phrases de deux CI simples au plus.

Le découpage en deux CI d'une phrase correspond à la frontière devant le prédicat de cette phrase. On peut distinguer les phrases en fonction du type de prédicat comme suit :

a) Le prédicat est un verbe ou un adjectif[59], lié directement au sujet
   - Exemples :
     - Tôi **đọc** sách =$_{litt}$ *Je* ***lire*** *livre = Je lis des livres (un livre)*
     - Chiếc đồng hồ ấy **đẹp** = <*class.*> *montre ce* ***jolie*** *= Cette montre est jolie.*
   - Négation : **không** devant le prédicat, par exemple :
     - Tôi **không đọc** sách = *Je* ***ne lis pas*** *de livre.*
     - Chiếc đồng hồ ấy **không đẹp** = *Cette* ***montre n'est pas jolie****.*

b) Le prédicat est un groupe nominal, un groupe verbal ou une proposition, lié au sujet grâce au copule là (*être*)[60]
   - Exemples :
     - Tôi là **sinh viên** =$_{litt}$ *Je être* ***étudiant*** *= Je suis étudiant.*
     - Học cũng là **làm việc** =$_{litt}$ *Apprendre aussi être travailler = Apprendre est également travailler.*

---

[59] Certains linguistes considèrent que le vietnamien n'a pas de verbe ni d'adjectif (*cf.* Cao X. Hạo [CAO 00]). Cette classification, d'après ces linguistes, est purement une reprise forcée des langues indo-européennes. Ils nomment « prédicat » tous les mots que nous avons appelé jusqu'ici verbe et adjectif. Pour nous, ce n'est pas une question vitale en TAL – cela n'est qu'une question de nomination, car la distinction entre les classes « verbe » et « adjectif » considérées est claire et réalisée proprement dans le dictionnaire du vietnamien Hoàng Phê [HOA 02].

[60] *Cf.* section 4.4.2.4 pour une explication plus détaillée sur le prédicat de ce type de phrase.

- Mình nói dối là **mình dại** =litt. Je mentir être **je stupide** = Je serais stupide si je mentais.

o Négation : không phải devant là. Par exemple : Tôi **không phải là sinh viên** = *Je **ne suis pas étudiant**.*

c) Le prédicat est une proposition, un groupe nominal (souvent nombre + nom), un groupe prépositionnel ou une locution, lié directement au sujet

o Exemples :

- Nó **tên là Nam** =litt *Il **nom être Nam** = Il s'appelle Nam*.
- Tôi **30 tuổi** =litt. *Je **30 ans** = J'ai 30 ans*.
- Sách này **của tôi** =litt *Ce livre **de moi** = Ce livre est à moi*.
- Nó **đầu bò đầu bướu** lắm = litt *Il **tête-bœuf-tête-bossue** très = Il est têtu comme une mule*.

o Négation : **không phải** devant le prédicat. Par exemple : Nó **không phải tên là Nam** = *Il ne s'appelle pas Nam*.

On peut également avoir des prédicats composés à partir des trois types de prédicats précédents :

- (aa) Chị nhìn anh, cười =litt *Elle regarder il, sourire = Elle le regarde en souriant*.
- (bb) Nó là lớp trưởng, cũng là sinh viên giỏi nhất =litt *Il être classe chef, aussi être étudiant bien premier = Il est délégué de classe et aussi le meilleur étudiant*.
- (cc) Cô bé tên là Oanh, 18 tuổi =litt *Fille petit nom être Oanh, 18 an = La jeune fille s'appelle Oanh, elle a 18 ans.*
- (ab/ba) : Nó thông minh, là sinh viên Tin học =litt *Il intelligent, être étudiant informatique = Il est intelligent et est étudiant en informatique.*
- (ac/ca) Anh cười, khuôn mặt rạng rỡ =litt *Il sourire, visage rayonnant = Il sourit, le visage rayonnant.*
- (bc/cb) Sách này của tôi, là cuốn cũ nhất =litt *Livre ce de moi, être <class.> vieux premier = Ce livre est le plus vieux de mes livres.*

En cas de prédicat composé, si un prédicat est un verbe avec modificateur, on peut le déplacer avant le sujet. À partir de l'exemple (aa), on peut ainsi avoir :

- * Cười, chị nhìn anh =litt *Souire, elle regarder il.*
- Cười hiền, chị nhìn anh = litt *Sourire gentil, elle regarder il = Elle le regarde en souriant gentiment.)*

Outre les composants noyaux de la phrase comme le prédicat, son sujet et ses compléments obligatoires, il existe en vietnamien quatre types de constituants de subordination dépendant du noyau de la phrase :

- Le thème (topique) de la phrase, qui se trouve toujours avant son noyau ;
- Le constituant de modalité, qui se trouve toujours à la fin de la phrase ;
- Les adverbes de phrases, qui ont deux positions possibles : devant le noyau de la phrase ou entre le sujet et le prédicat ;
- Les locutions adverbiales, qui ont trois positions possibles : soit devant le noyau de la phrase, soit après le noyau de la phrase, soit entre le sujet et le prédicat de la phrase.

Pour une modélisation de la structure d'une phrase en TAG, se pose la question suivante : la représentation arborescente du cadre de sous-catégorisation du prédicat noyau de la phrase doit-elle être plate (par exemple, formule S = NP1 V NP2, au cas où V est un verbe transitif) ou à deux niveaux (par exemple, formule S = NP VP). Pour l'anglais, le projet XTAG [XTA 01] choisit la représentation à deux niveaux. Pour le français, Abeillé [ABE 02] a montré que la représentation plate est la plus appropriée. Pour le vietnamien, nous sommes favorable à une représentation à deux niveaux. Notre choix vient du fait qu'il existe deux types d'adverbes modifiant un prédicat noyau : les pré-modifieurs et les post-modifieurs. Les pré-modifieurs se trouvent juste avant le prédicat noyau, alors que les post-modifieurs se trouvent toujours derrière le noyau de la phrase, après les compléments d'objets, ce qui empêche qu'ils soient insérés par adjonction sur le verbe. La représentation du noyau de la phrase à deux niveaux NP PredP est ainsi la seule qui permette d'adjoindre les adverbes à leur bonne position en assurant la cohérence entre le traitement des pré- et post-modifieurs (*cf.* Figure 4-13). Notons que PredP peut être aussi bien un groupe verbal qu'un groupe adjectival ou un groupe nominal ou même un groupe prépositionnel (*cf.* les types de prédicat a, b et c).

Figure 4-13 Exemples d'adjonction des adverbes de temps et d'aspect au groupe prédicatif

Nous présentons dans les sections suivantes chaque type de prédicat : prédicat verbal (4.4.2.2), prédicat adjectival (4.4.2.3), et d'autres prédicats de type c) (4.4.2.4) ; puis les constituants subordonnés (4.4.2.5).

4.4.2.2. Phrases avec prédicat verbal

Cette section introduit la sous-catégorisation de différents verbes, à l'exception des verbes copules, jouant le rôle de prédicat de type (a) mentionné dans la section précédente.

Comme nous en avons déjà vu au Chapitre 2, le vietnamien ne comporte pas de variation morphologique. La caractérisation fonctionnelle des verbes, qui est *a priori* universelle, sert de base pour décrire les structures syntaxiques des phrases. Dans la plupart des cas, l'ordre des composants des phrases est Sujet – Verbe – Objet, comme nous le constatons ci-après.

Ordre « standard » des arguments

Nous pouvons schématiser l'ordre « standard » des arguments des verbes comme suit :

1) N_1 V, où N_1 est le sujet du verbe V (cadre intransitif)
   - Exemple : `Tôi | ngủ` $=_{litt}$ *Je dormir = Je dors*.
2) V N_2, où N_2 est le complément d'objet du verbe impersonnel V (verbe d'existence)
   - Exemple : `Có | một quyển sách trên bàn`
     $=_{litt}$ *Avoir un <class.> livre sur table = Il y a un livre sur la table*.
3) N_1 V PrepP, où N_1 est le sujet du verbe locatif V, et PrepP est un groupe nominal ou prépositionnel désignant une localisation.
   - Exemple : `Nhà tôi | ở | đây`
     $=_{litt}$ *Maison je* | *se trouver* | *ici = Ma maison est ici*.

4) N_1 V N_2, où N_1 est le sujet, et N_2 est le complément d'objet du verbe V (cadre transitif).
    - Exemple : Tôi | đọc | sách =$_{litt}$ *Je* | *lire* | *livre = Je lis des livres*.

5) N_1 V N_2 [Prep] N_3 ou N_1 V [Prep] N_3 N_2, où N_1 est le sujet, N_2 est l'objet, et N_3 est la destination ou l'expédition du verbe V (cadre ditransitif), la préposition devant N_3 est souvent facultative.
    - Exemples :
        - Tôi | trả | sách | [cho] thư viện =$_{litt}$ *Je* | *rendre* | *livre* | *[à] bibliothèque = Je rends des livres à la bibliothèque*.
          Equivalent à « Tôi | trả | [cho] thư viện | sách ».
        - Tôi | mượn | sách | [của] thư viện =$_{litt}$ *Je* | *emprunter* | *livre* | *[de] bibliothèque = J'emprunte des livres à la bibliothèque*.
          Equivalent à « Tôi | mượn | [của] thư viện | sách ».

6) N_1 V N_2 N_3, où N_1 est le sujet, N_2 est le complément d'objet du verbe de comparaison V, et N_3 est le « résultat » de la comparaison.
    - Exemple : Tôi | kém | nó | 3 tuổi =$_{litt}$ *Je* | *moins* | *il* | *3 an = J'ai 3 ans de moins que lui*.

7) N_1 V N_2 PredP, où N_1 est le sujet, N_2 est le complément d'objet du verbe impératif ou causatif V, et PredP est le complément prédicatif (phrase avec sujet vide).
    - Exemple : Mẹ | bảo | tôi | đi chợ =$_{litt}$ *Maman* | *demander* | *je* | *aller marché = Maman me demande d'aller faire les courses*.

8) N_1 V N2 PrepP ou N_1 V PrepP N_2, où N_1 est le sujet, N_2 est le complément d'objet du verbe ditransitif V, PrepP est le syntagme prépositionnel.
    - Exemple : Tôi buộc trâu vào cột =$_{litt}$ *Je lier buffle à poteau = J'attache le buffle au poteau*.

9) N_1 V VP, où N_1 est le sujet, et VP est le complément verbal (phrase avec le sujet vide) du verbe V (verbe modal, verbe d'impression).
    - Exemples :
        - Tôi | định | mua một quyển sách mới =$_{litt}$ *Je* | *<avoir l'intention de>* | *acheter un <class.> livre nouveau = J'ai l'intention d'acheter un nouveau livre*.
        - Tôi | thích | mua một quyển sách mới =$_{litt}$ *Je* | *aimer* | *acheter un <class.> livre nouveau = J'aimerais acheter un nouveau livre*.

10) N_1 V [C] S, où N_1 est le sujet du verbe V (qui peut être soit un verbe d'impression, soit un verbe d'énoncé, *etc.*), S est le complément phrastique, et la conjonction C est souvent facultatif.
    - Exemple : Tôi | biết | [rằng/là] trời mưa =$_{litt}$ *Je* | *savoir* | *[que] ciel pleuvoir = Je sais qu'il pleut*.

11) N_1 V N_2 (VP), où N_1 est le sujet, N_2 est la direction du verbe de mouvement avec direction V, VP est le complément verbal (phrase avec le sujet vide).
    - Exemple : Tôi | vào | hiệu sách | (mua một quyển sách mới) =$_{litt}$ *Je* | *entrer* | *magasin livre (acheter un <class.> livre nouveau) = J'entre à la librairie (acheter un nouveau livre)*.

12) N_1 V (VP), où N_1 est le sujet du verbe de mouvement sans direction V, VP est le complément verbal (phrase avec le sujet vide) dont le verbe est un verbe de mouvement avec direction.

- Exemples :
  - Tôi | chạy (về nhà) =$_{litt}$ *Je* | *courir* (*rentrer maison) = Je rentre en courant à la maison.*
  - Tôi | chạy (vào hiệu sách mua một quyển sách mới) =$_{litt}$ *Je courir (entrer magasin livre acheter un <class.> livre nouveau) = J'entre à la librairie en courant pour acheter un nouveau livre*.

13) S_1 V N_2, où S_1 est un sujet phrastique, et N_2 est le complément d'objet du verbe V (verbe de transformation, verbe d'une action agissant sur un patient).

- Exemple : Phụ nữ lái tắc-xi | đã trở thành | chuyện bình thường =$_{litt}$ *Femme conduire taxi* | *déjà devenir* | *phénomène normal = Il n'est plus exceptionnel de voir une femme conduire un taxi.*

14) VP_1 V N_2, où VP_1 est un sujet verbal (phrase avec sujet vide), et N_2 est le complément d'objet du verbe V (verbe de transformation, verbe d'existence, verbe de passivité).

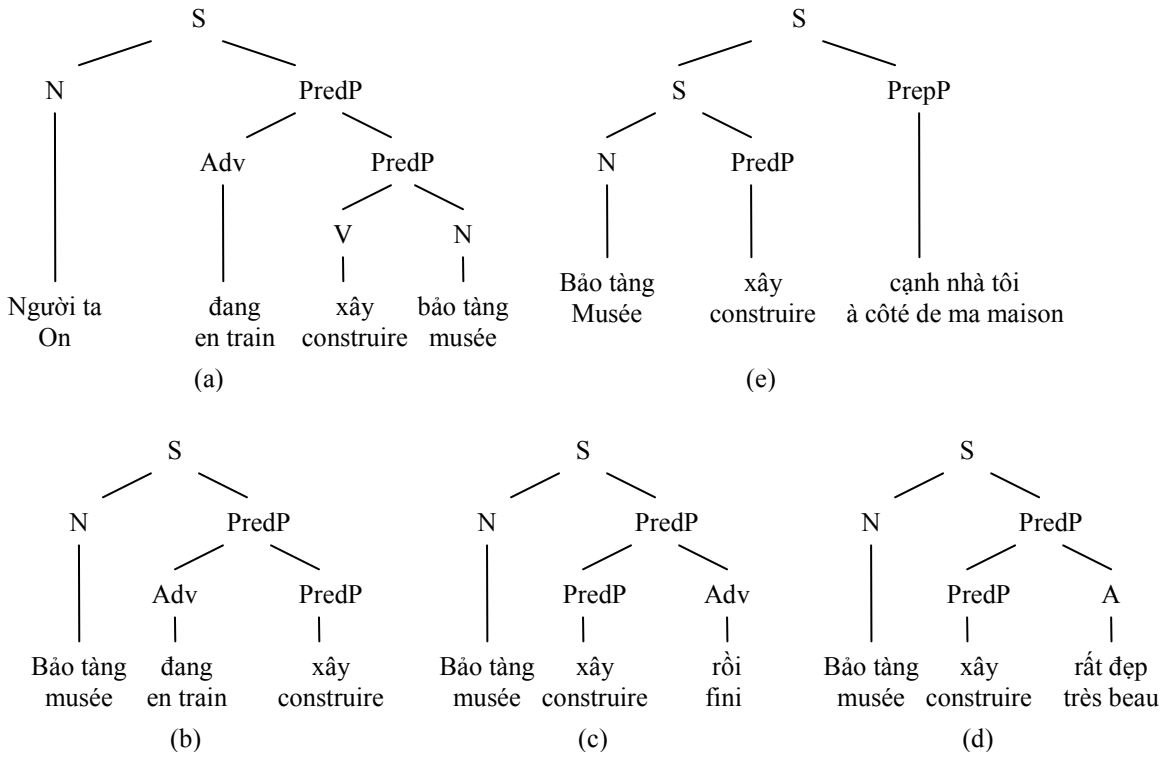
- Exemple : Lái tắc-xi | đã trở thành | một nghề phổ biến =$_{litt}$ *Conduire taxi* | *déjà devenir* | *un métier répandu =Chauffeur de taxi est devenu un métier répandu.*

Phrases sans sujet agentif

Outre les structures suivant l'ordre Sujet-Verbe-Objet comme ci-dessus, on rencontre assez fréquemment en vietnamien un type de phrases dont le sujet agentif est absent, et le sujet grammatical (se trouvant au début de la phrase) est le complément d'objet du prédicat verbal (qui sous-catégorise, bien entendu, au moins un complément d'objet). Dans ces cas, le locuteur n'aborde pas l'agent de l'action, mais insiste sur les informations concernant le temps, la manière ou l'aspect de l'action. Pour cette raison, le prédicat verbal de ces phrases doit être obligatoirement modifié par un complément circonstanciel. Cela est illustré dans l'exemple suivant.

Exemple :

- Sous-catégorisation N_1 V N_2 : Người ta đang xây bảo tàng =$_{litt}$ *On <adverbe de temps : présent> construire musée = On est en train de construire un musée*. (*cf.* Figure 4-14 (a))
- Phrases du type N_2 V :
  - Bảo tàng đang xây =$_{litt.}$ *Musée <adverbe de temps : présent> construire = Le musée est en train d'être construit*. (*cf.* Figure 4-14 (b))
  - Bảo tàng xây rồi =$_{litt.}$ *Musée construire <adverbe d'aspect : parfait> = Le musée a été construit*. (*cf.* Figure 4-14 (c))
  - Bảo tàng xây rất đẹp =$_{litt}$ *Musée construire très beau = Le musée est construit de très belle manière*. (*cf.* Figure 4-14 (d))
  - Bảo tàng xây cạnh trường tôi =$_{litt}$ *Musée construire à_côté_de école je = Le musée est construit à côté de mon école*. (*cf.* Figure 4-14 (e))



(*Note : Nous ne détaillons pas les niveaux intermédiaires qui ne sont pas sujet de l'illustration*)

Figure 4-14 Exemples de phrases dont le sujet grammatical est l'objet logique du verbe noyau

Un autre type de phrases sans sujet est celui des phrases ayant la structure PrepP PredP, où PrepP est un groupe prépositionnel désignant une localisation, et PredP est un groupe verbal correspondant à une phrase sans sujet. Par exemple, `Trên bàn đặt cuốn sách` =$_{litt}$ *Sur table poser <class.> livre = Un livre est posé sur la table*.

Sujet inversé

Plusieurs types de verbe permettent le déplacement du sujet derrière le groupe verbal prédicatif. Dans ces phrases, le constituant se trouvant au début de la phrase est souvent un groupe prépositionnel, qui joue le rôle de complément circonstanciel, ou un groupe nominal qui joue le rôle de sujet thématique. Dans ce dernier cas, la relation entre le sujet du verbe et le sujet thématique est souvent la possession du premier par le second. Nous l'illustrons par les exemples suivants :

- `Nhà cháy` =$_{litt}$ *Maison brûler = La maison brûle*.
  - `(Trong làng)`[61] `cháy nhà` =$_{litt}$ *Dans village brûler maison = Dans le village, une maison brûle*.
  - `(Tôi) cháy nhà` =$_{litt}$ *Je brûler maison = Ma maison brûle*[62].
- `Cơn sốt lại lên` =$_{litt}$ *Fièvre de_nouveau monte = La fièvre remonte.*
  - `Tôi lên cơn sốt` =$_{litt}$ *Je monter fièvre = Ma fièvre monte*.

Nous introduisons maintenant le phénomène de passivation.

---

[61] Ce qui est entre parenthèses peut être éliminé.

[62] Le verbe `cháy` vietnamien n'est pas ergatif comme le français « brûler », la phrase exemple ne peut donc pas être traduite comme « je brûle la maison ».

Passivation

Etant donné une forme active 4) N_1 V N_2, on peut obtenir une ou plusieurs des formes passives suivantes :

- N_2 « Verbe de passivité : bị/được » N_1 V
  - Exemple :
    - Actif : Tôi chọn cuốn sách này =$_{litt}$ *Je choisir <class.> livre ce = Je choisis ce livre.*
    - Passif : Cuốn sách này **được** tôi chọn =$_{litt}$ *<class.> livre ce **<verbe de passivité>** je choisir = Ce livre a été acheté par moi*.
- N_2 « Verbe de passivité : bị/được » V « par : bởi » N_1
  - Exemple :
    - Actif : Một nhà sư Ấn Độ xây chùa này =$_{litt}$ *Un bonze Inde contruire pagode ce = Un bonze indien a construit cette pagode*
    - Passif : Chùa này **được** xây **bởi** một nhà sư Ấn Độ =$_{litt}$ *Pagode cette **<verbe de passivité>** construire **par** un bonze Inde = Cette pagode a été construite par un bonze indien.*
- N_2 « Verbe de passivité : bị/được » V
  - Exemple :
    - Actif : Thầy giáo khen nó =$_{litt}$ *Maître féliciter il = Le maître l'a félicité*.
    - Passif : Nó **được** khen =$_{litt}$ Il ***<verbe de passivité>** féliciter = Il est félicité.*

Les formes actives 5), 7) et 8) ont également des transformations passives que nous ne détaillons pas ici.

Certaines classes de verbes ayant un cadre transitif (par exemple des verbes de transformation, de modalité, *etc.*) n'admettent pas de transformation passive.

Bilan

L'étude des structures syntaxiques principales présentées ci-dessus montre qu'il n'existe pas d'obstacle particulier à leur représentation dans le formalisme TAG. Pour la modélisation en TAG des phrases ayant un prédicat verbal, il serait nécessaire de définir les traits sémantiques contraignant la sous-catégorisation d'un verbe, ainsi que les traits sémantiques des arguments pour pouvoir repérer les cas de déplacement du sujet ou d'autres arguments du verbe. Les rôles thématiques devraient également être annotés afin d'éclaircir ces structures.

Nous passons maintenant aux phrases construites autour d'un prédicat adjectival.

4.4.2.3. Phrases avec prédicat adjectival

A la différence du français, le vietnamien peut recevoir un adjectif comme prédicat sans avoir besoin du verbe copule « être ». Dans ce cas, il est nécessaire que l'adjectif soit accompagné d'un adverbe, ou que le sujet nominal soit bien déterminé. Comparons par exemple :

- Chiếc đồng hồ **đẹp** = *<class.> montre **jolie*** peut être reconnu seulement comme un groupe nominal.
- Chiếc đồng hồ **rất đẹp** = *<class.> montre **très jolie** = La montre est **très jolie*** peut être reconnu soit comme un groupe nominal, soit comme une phrase.

- Chiếc đồng hồ ấy **đẹp** = <*class.*> *montre ce* ***jolie*** = *Cette montre est* ***jolie*** peut être reconnu seulement comme une phrase (car l'adjectif se trouve après le pronom démonstratif ấy).

Le sujet des phrases attributives peut être phrastique ou prédicatif. Par exemple,

- Nói chuyện với anh ta **rất chán** =$_{litt}$ *Parler avec il* ***très ennuyeux*** = *C'est très ennuyeux de parler avec lui*.
- Anh nói thế **không đúng** =$_{litt}$ *Vous dire cela* ***pas vrai*** = *Ce que vous dites n'est pas vrai*.

Les phrases attributives avec sujet nominal peuvent recevoir un adverbe de phrase qui est un verbe. Par exemple :

- Chiếc đồng hồ rất đẹp = *La montre est très jolie* => Chiếc đồng hồ **trông** rất đẹp =$_{litt}$ *La montre regarder très jolie = La montre est très jolie à regarder*[63]. Cette phrase est équivalente à : **Trông** chiếc đồng hồ rất đẹp =$_{litt}$ ***Regarder*** *la montre très jolie*.
- Chuối này không ngon =$_{litt}$ *Banane ce pas délicieux = Cette banane n'est pas bonne*. => Chuối này **ăn** không ngon =$_{litt}$ *Banane ce* ***manger*** *pas délicieux = Cette banane n'est pas bonne à manger*[63]. Cette phrase est équivalente à : **Ăn** chuối này không ngon.

Les adjectifs quantitatifs (comme par exemple « cao = haut ») sous-catégorisent un groupe nominal de mesure. Par exemple :

- Nhà này **cao 5m** =$_{litt}$ *Maison ce haut 5m = Cette maison fait 5m de haut*.

Dans ce cas, l'ajout des adverbes de degré doit être interdit :

- Nhà này **rất cao** =$_{litt}$ *Maison ce très haut = Cette maison est très haute*.
- ***Nhà này **rất cao 5m** =$_{litt}$ *Maison ce très haut 5m*

Comme les prédicats verbaux, les prédicats adjectivaux peuvent être modifié par les adverbes de temps et d'aspect (par exemple, « đã = déjà », « rồi = aspect parfait (action finie) »). Par ailleurs, un prédicat adjectival peut être modifié par un adverbe qui est un verbe de mouvement avec direction ; dans ce cas, le prédicat obtenu désigne un processus. Par exemple :

- Oanh rất **đẹp** =$_{litt}$ *Oanh très beau = Oanh est très belle*. => Oanh **đẹp lên** nhiều =$_{litt}$ *Oanh beau s'élever beaucoup = Oanh a beaucoup embelli*.
- Ông ấy rất **già** =$_{litt}$ *Homme ce très vieux = Cet homme est très vieux*. => Ông ấy **già đi** rất nhanh =$_{litt}$ *Homme ce vieux aller très vite = Cet homme vieillit très vite*.

Ces caractéristiques font partie des raisons pour lesquelles certains linguistes ont avancé que le vietnamien n'a pas de catégories « verbe » et « adjectif » (*cf.* note 59, page 120). Afin de rester ouvert vis-à-vis de cette possibilité théorique, nous choisissons, quoiqu'ait retenu la distinction verbe/adjectif dans les descriptions morphosyntaxiques, d'adopter une dénomination commune « PredP » pour les groupes prédicatifs, qu'ils soient verbaux ou adjectivaux. Cela est en outre cohérent avec le choix de représentation en deux niveaux de la structure de phrase, présenté à la section 4.4.2.1.

Le dernier type de phrase nous restant à étudier est celui des phrases copulatives (à la forme affirmative ou négative) du vietnamien.

[63] Cette formulation est redondante en français, mais elle est naturelle en vietnamien.

4.4.2.4. Phrases utilisant un copule à la forme affirmative ou négative

Le copule « là » (*être*) est considéré soit comme un verbe spécial selon le Comité de Sciences Sociales [UYB 83], soit comme un mot outil selon certains autres auteurs (*cf.* Nguyễn M. Thuyết [NGU 98a]). Dans le premier cas, le prédicat des phrases copulatives doit être le verbe copule « là », dans le deuxième cas, le prédicat est la partie suivant le copule (*cf.* le type de phrase (b) à la section 4.4.2.1).

La structure d'une phrase copulaire peut être :

- N_1 « là » N_2
- S/PredP/PrepP « là » N_1 ou N_1 « là » S/PredP/PrepP, où S et PredP sont les phrases complètes ou avec le sujet vide, et PrepP est le groupe prépositionnel désignant une localisation.

Le groupe prédicatif noyau de ces phrases peut être modifié par des adverbes de temps et d'aspect comme les verbes « normaux » : les pré-modifieurs se trouvent devant le copule « là », les post-modifieurs se trouvent après le noyau de la phrase. La forme négative, à la différence des verbes, est réalisée par le groupe copulatif « không phải » (*non-être*) au lieu de « không » (*non*). En revanche, ce groupe copulatif s'ajoute également devant la copule « là ». Nous avons donc adopté la solution consistant à considérer la copule « là » comme un verbe particulier, et ce mot appartient donc au groupe prédicatif noyau.

Le copule « không phải » sert également à construire le négatif des phrases du type (c) introduit à la section 4.4.2.1, et dont les structures générales sont les suivantes :

- N_1 N_2, où N_2 est un groupe nominal quantifié (souvent le temps).
  - Tôi **không phải** 30 tuổi =$_{litt.}$ *Je **non-être** 30 ans = Je n'ai pas 30 ans*.
- N_1 PrepP
  - Sách này **không phải** của tôi =$_{litt}$ *Ce livre **non-être** de moi = Ce livre n'est pas à moi*.
- N_1 S, où le sujet de S est généralement à la possession de N_1. Exemples :
  - Nó **không phải** tên là Nam =$_{litt}$ *Il **non-être** nom être Nam = Il ne s'appelle pas Nam*.

Pour les constructions de phrases de type (c), sous leur forme affirmative ou négative, la partie droite de la phrase (en enlevant N_1), constitue comme pour les autres types de phrases un groupe prédicatif pouvant être précisé par des pré- et post-modifieurs.

Nous avons jusqu'ici considéré tous les types de phrases déclaratives simples en vietnamien. L'ordre des constituants noyaux étant comme nous l'avons constaté assez figé, la modélisation de cet aspect de la grammaire ne devrait pas soulever de difficulté particulière. En revanche, il convient de sélectionner et définir les traits sémantiques lexicaux les mieux à même de contraindre la sélection des cadres de sous-catégorisation. Cette problématique est *a priori* indépendante du formalisme syntaxique employé, et devrait faire l'objet d'une étude rigoureuse dans un autre projet à plus long terme.

Ayant réalisé un tour d'horizon rapide de l'ensemble des phénomènes syntaxiques à l'œuvre dans les phrases simples du vietnamien, nous nous proposons d'aborder dans la sous-section suivante la description des constituants subordonnés de la structure nucléaire de la phrase.

4.4.2.5. Constituants subordonnés

Thème et thématisation

En vietnamien, comme en chinois (Chen *et al.* [CHE 03]), un thème peut être un argument du prédicat noyau ou non. Par exemple :

- **Áo này** (thì[64]) tôi không có tiền =$_{litt}$ ***Chemise ce*** *(alors) je non avoir argent = Je n'ai pas d'argent pour cette chemise*.
- **Sinh viên** họ rất năng động =$_{litt}$ *Etudiant ils très dynamique = Les étudiants sont très dynamiques*.
- **Thuốc** (thì) ông ấy không hút =$_{litt}$ ***Cigarette*** *(alors) homme ce non fumer = Il ne fume pas de cigarettes*.

Dans l'idéal, une formalisation grammaticale du vietnamien devrait permettre de distinguer les thèmes en fonction du rôle qu'ils remplissent vis-à-vis du prédicat (complément circonstanciel dans le premier exemple, sujet dans le second, objet dans le troisième).

Au contraire du constituant thématique, qui se trouve toujours devant la structure nucléaire de la phrase, le constituant de modalité se trouve toujours à la fin de la phrase.

Modalité (langue parlée)

Le constituant de modalité est réalisé soit par une particule modale simple, soit par un syntagme modal toujours introduit par les particules de modalité thì ou là. Ce constituant est donc relativement facile à reconnaître.

- Particules modales :
  - Anh cho em đi theo với! =$_{litt}$ *Vous accorder je aller suivre <mot modal> = Laissez moi vous suivre, s'il vous plaît*.
  - Tôi rất tin tưởng anh mà =$_{litt}$ *Je très faire confiance vous <mot modal> = Mais je vous fais très confiance*.
  - Anh về à! =$_{litt}$ *Vous rentrer <mot modal> = Vous rentrez déjà !* (ou *Vous rentrez ?*, suivant l'intonation du locuteur).
- Syntagmes modaux :
  - Trời mưa thì[65] phải =$_{litt}$ *Ciel pleuvoir alors <mot modal> = Il pleut, à ce qu'il paraît*.
  - Nó nặng 20 cân là cùng =$_{litt}$ *Il peser 20 kg être extrême = Il pèse 20kg au maximum*.
  - Nó lười thì có =$_{litt}$ *Il paresseux alors oui = Dîtes plutôt qu'il est paresseux/C'est plutôt lui qui est paresseux (selon l'intonation du locuteur)*.

Adverbe de phrase

Les adverbes de phrases se composent des mots, expressions ou idiomes qui complémentent toute la phrase et qui peuvent avoir deux positions : soit devant la structure nucléaire de phrase, soit entre le sujet et le groupe prédicatif de la phrase. Par exemple :

---

[64] Dans ce contexte, « thì » est un particule modal, qui ne se traduit pas. Il peut être absent.

[65] En vietnamien, « thì = alors, là = être » sont des marqueurs de modalité.

- Hình như anh ấy đã về = *Il semble qu'il est rentré.* / Anh ấy hình như đã về = *Il semble être rentré.*
- Thỉnh thoảng tôi về thăm anh ấy = *De temps en temps je lui rends visite.* / Tôi thỉnh thoảng về thăm anh ấy = *Je lui rends visite de temps en temps.*

Locution adverbiale

Les locutions adverbiales peuvent occuper couramment trois positions différentes : devant la structure nucléaire, après cette structure, ou entre le sujet et le groupe prédicatif.

Ces locations sont généralement soit un groupe nominal de temps, soit un groupe prépositionnel désignant un lieu, un objectif, une cause, *etc.*

- Trong vườn, một lũ trẻ chơi đùa vui vẻ. = *Dans le jardin, un groupe d'enfants s'amusent joyeusement.* = Một lũ trẻ trong vườn chơi đùa vui vẻ. = Một lũ trẻ chơi đùa vui vẻ trong vườn.
- Khi ấy, trời rất lạnh. = *En ce moment là, il fait très froid.* = Trời khi ấy rất lạnh. = Trời rất lạnh khi ấy.
- Đối với tôi, cuộc sống thật là đẹp. = *Pour moi, la vie est vraiment belle.* = Cuộc sống, đối với tôi, thật là đẹp. = Cuộc sống thật là đẹp đối với tôi.
- Vì trời mưa, tôi không đi chơi. = *Puisqu'il pleut, je ne sors pas.* = Tôi, vì trời mưa, không đi chơi. = Tôi không đi chơi, vì trời mưa.
- Bất thình lình, trời đổ mưa. = *Soudain, la pluie tombe.* = Trời bất thình lình đổ mưa. = Trời đổ mưa bất thình lình.

Tous ces quatre constituants sont reconnaissables grâce aux mots marqueurs ou au lexique.

4.4.3. Bilan

Dans cette section sur la description syntaxique du vietnamien, nous avons étudié la modélisation en TAG du groupe nominal en vietnamien. Nous avons modélisé les différentes compositions d'un syntagme nominal du vietnamien et fait les tests avec l'analyseur LLP2. Pour faire fonctionner cet analyseur sur les exemples du vietnamien sans le modifier, il est nécessaire de coller les syllabes des mots composés par des « _ », afin que l'analyseur les reconnaisse comme des mots. Ce problème peut être résolu en intégrant l'outil de segmentation en unités lexicales pour le vietnamien à l'analyseur LLP2. Ce travail a été réalisé dans le cadre du stage de LE Hong Phuong ([LEH 05]), qui a également réalisé une mise à jour de la gestion des structures de traits du système LLP2, suivant la représentation normalisée ISO/DIS 24610-1 proposée par l'ISO/TC 37/SC 4 (*cf.* document [ISO 06]).

Nous avons ensuite étudié les principales structures de phrase du vietnamien, en nous fondant sur les descriptions présentées dans la littérature. Nous avons présenté un aperçu assez large, quoique inégalement approfondi, des phénomènes grammaticaux du vietnamien :

- La structure nucléaire de phrase : <Sujet> <Prédicat> <Compléments obligatoires du prédicat>, où le prédicat peut être verbal, adjectival, nominal ou phrastique. Le sujet grammatical peut être l'objet logique du prédicat noyau. La possibilité d'absence d'un ou plusieurs arguments et l'ordre dans lequel ils apparaissent sont décidés par le prédicat. Nous suggérons, au vu notamment des mécanismes d'introduction d'adverbes dans les phrases, de représenter la structure nucléaire des phrases par une hiérarchie à deux niveaux : Sujet PredP, où PredP regroupe le prédicat et les arguments à sa droite. Cette structure nous permet en outre d'unifier sous le nom PredP la représentation des prédicats verbaux et adjectivaux, qui peuvent jouer en vietnamien des rôles très similaires.

- Les constituants subordonnés de la structure nucléaire de phrase sont classés en quatre types, en fonction des positions auxquelles ils peuvent s'adjoindre à la structure nucléaire, et de leur catégorie syntaxique.

Précisons enfin quelques caractéristiques générales de la grammaire vietnamienne :

- Les propositions relatives s'insèrent de manière tout à fait naturelle (éventuellement récursive) dans une structure initiale, sans modification de cette dernière. Leur modélisation en TAG ne soulève donc aucune difficulté.
- Un composant de phrase (que ce soit le sujet logique de phrase ou un autre argument du prédicat) peut être totalement absent s'il est mentionné précédemment ou peut être déduit du contexte. Si en français, l'extraction d'un composant est faite en le remplaçant par un pronom, en vietnamien elle est faite tout simplement en supprimant ce composant. Ce phénomène permet également de rendre compte des propositions relatives, qui ne comprennent pas en vietnamien de pronom relatif.
- Les questions en vietnamien sont formées en remplaçant l'élément d'information inconnu par un pronom interrogatif, sans aucune modification à la structure de la phrase. L'analyse des structures interrogatives ne suppose donc pas de travail particulier par rapport à celle des phrases « ordinaires ».

Suite à notre travail décrit ci-dessus, les structures verbales de base ont été implémentées dans le système LLP2 par LE Hong Phuong pendant son stage (*cf.* [LEH 05, 06]). Les données que nous avons créées sont accessibles sur le site Internet de l'équipe Langue et Dialogue au Loria.

Notre objectif à moyen terme est non seulement d'approfondir l'étude de la modélisation de la grammaire, mais aussi de mener à bien la construction des ressources linguistiques pour l'analyse syntaxique du vietnamien, dont le lexique syntaxique, les corpus arborés, *etc.* Nous présentons à la section suivante, en conclusion de ce chapitre, un aperçu du travail réalisé à cette fin.

4.5. Bilan et perspectives

Dans ce chapitre, nous avons étudié la modélisation de la grammaire vietnamienne par le formalisme TAG. Une première proposition de modélisation du groupe nominal a été définie, nous permettant de montrer l'adaptation du formalisme adopté aux mécanismes syntaxiques à l'œuvre en vietnamien. Nous avons également recensés l'ensemble des phénomènes syntaxiques majeurs du vietnamien devant être pris en compte par une telle grammaire, en accompagnant ceux-ci d'exemples illustratifs, ce qui doit permettre de guider et d'évaluer le progrès des développements futurs. C'est un travail aussi difficile que passionnant, car le problème d'analyse les constituants de phrase en vietnamien n'a pas encore fait l'objet d'un consensus au sein de la communauté linguistique au Vietnam.

Le travail que nous avons engagé pour l'analyse syntaxique du vietnamien n'est encore qu'un début, et il reste un long chemin à parcourir avant d'achever la construction des ressources et d'outils pour l'annotation syntaxique du vietnamien. Nous tentons, dans ce qui suit, de concrétiser les pistes à suivre pour construire les ressources syntaxiques qui n'existent pas encore pour le vietnamien : un lexique syntaxique, une grammaire et une base de phrases de test, et un corpus arboré du vietnamien. La construction de ces ressources étant un travaux très coûteux à tous points de vue, nous mettons donc l'accent sur la définition des schémas de codage de ces ressources afin qu'elles soient extensibles, réutilisables et faciles à transporter d'un système à l'autre. Ce « plan de travail » constitue une base de réflexion utile pour l'élaboration de projets d'ingénierie des langues de grande ampleur au Vietnam.

4.5.1. Construction du lexique syntaxique

Toute analyse syntaxique a besoin d'un lexique fournissant des informations sur l'usage grammatical de chaque unité lexicale la langue. D'un formalisme syntaxique à l'autre, ces informations peuvent être plus ou moins riches, représentées de manières différentes. Mais une information donnée doit pouvoir être convertie aisément d'une application à l'autre. Bien que notre tentative actuelle soit de modéliser la grammaire vietnamienne avec le formalisme TAG, nous visons la construction d'un lexique syntaxique opérationnel pour le vietnamien qui ne dépende pas d'un formalisme de grammaire donné. Nous sommes convaincue que cela est possible grâce à une représentation normalisée qui facilite la conversion d'une même ressource pour l'utilisation dans différentes applications. Les efforts pour élaborer une telle norme sont menés par de nombreuses activités dans la communauté du TAL, et en particulier, entrepris par le projet LMF de l'ISO/TC 37/SC 4 que nous avons eu l'occasion de présenter au Chapitre 3 (*cf.* section 3.3.4.3).

4.5.1.1. Modèle de représentation du lexique syntaxique

Le méta-modèle LMF pour la représentation d'un lexique opérationnel se compose d'un modèle noyau (*cf.* Figure 4-15) partagé entre les lexiques qui s'organisent selon les sens de chaque unité lexicale, et des extensions lexicales (par exemple morphologie, syntaxe, ou sémantique).

La Figure 4-16 montre les extensions lexicales pour la syntaxe. Les descriptions syntaxiques sont attachées à chaque sens de l'entrée lexicale. Le composant « constructionSet » représente l'ensemble de réalisations syntaxiques (composant « syntacticConstruction ») correspondant à la formule logique (composant « semanticFormula ») exprimant le sens du mot. Le composant « syntacticPosition » représente la position d'un constituant sous-catégorisé d'un prédicat : il donne les informations concernant ce constituant, y compris le lien avec l'argument sémantique. Le modèle sémantique est spécifié dans la Figure 4-17. Pour une discussion détaillée sur l'interaction de ces composants, on peut se référencer au document [ISO 05b].

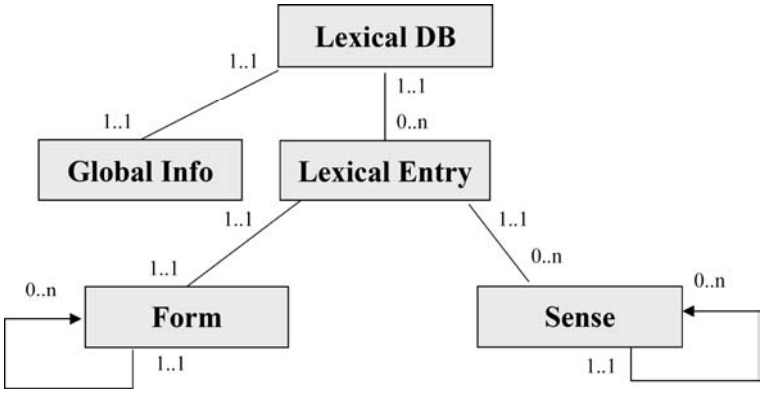


Figure 4-15 LMF – modèle noyau

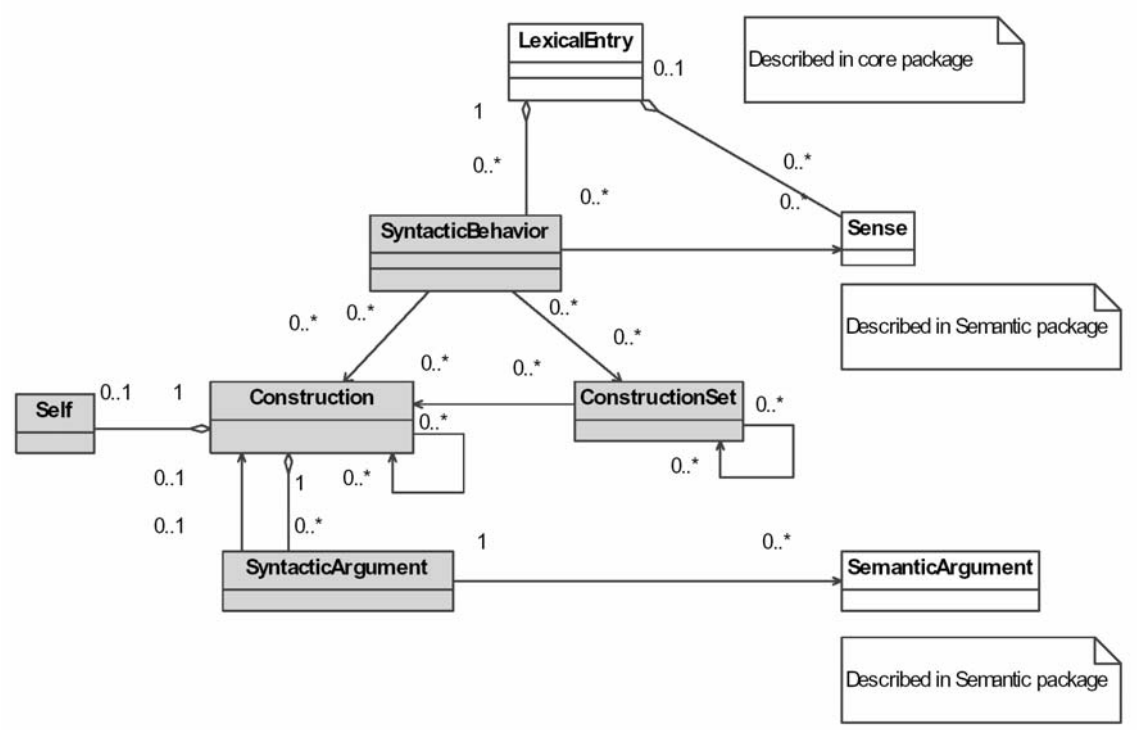


Figure 4-16 LMF – Extensions lexicales pour la syntaxe [ISO 05b]

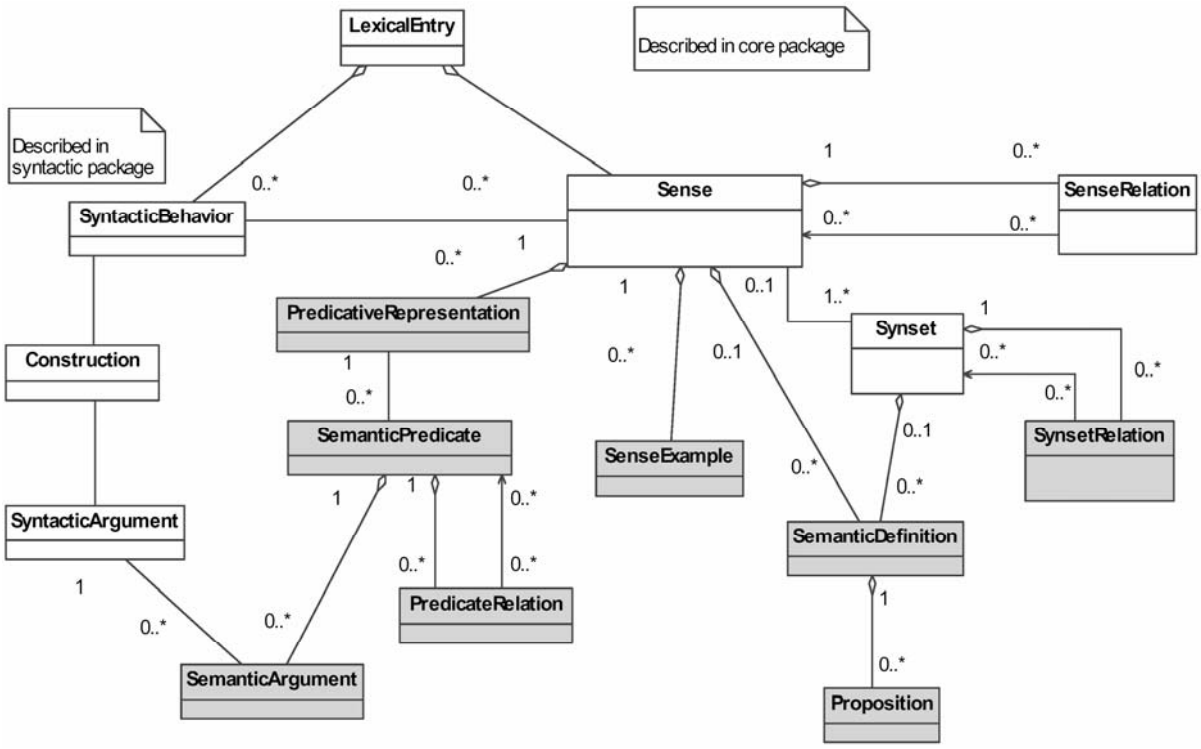


Figure 4-17 LMF – Extensions lexicales pour la sémantique [ISO 05b]

La Figure 4-18 exemplifie l'instanciation XML de la description syntaxique de l'entrée « regretter$_1$ » du DEC (*Dictionnaire Explicatif et Combinatoire*, *cf.* Figure 1-14).

REGRETTER (verbe)
1.
FP : LA personne X ~ SON action Y(X)
TR : X = I = N ; Y = II = N, de V-inf
EX : C'est une décision qu'il va regretter cruellement. Il ne regrette pas d'avoir investi 4 000 F.

```
<lexicalEntry>
    <lemma>regretter</lemma> <grammaticalCategory>verb</grammaticalCategory>
    <sense>
        <semanticFormula>
            <propositionalForm>LA personne X ~ SON action Y(X)</propositionalForm>
             <semanticArgument> <role>X</role> <properties>person</properties> </semanticArgument>
             <semanticArgument> <role>Y</role> <genitive>X</genitive> <properties>action</properties> </semanticArgument>
        </semanticFormula>
        <constructionSet>
            <syntacticConstruction referenceConstruction="X = I = N ; Y = II = N" exampleConstruction="C'est une décision qu'il va regretter cruellement " >
                <syntacticPosition>
                    <syntacticCategory>nominalPhrase</syntacticCategory>
                    <syntacticFunction>subject</syntacticFunction>
                    <semanticArgument>X</semanticArgument>
                </syntacticPosition>
                <syntacticPosition>
                    <alt>
                        <syntacticCategory>nominalPhrase</syntacticCategory>
                        <brack>
                            <syntacticCategory>nonFiniteClause</syntacticCategory>
                            <syntacticIntroducer>de</syntacticIntroducer>
                        </brack>
                    </alt>
                    <syntacticFunction>directObject</syntacticFunction>
                    <semanticArgument>Y</semanticArgument>
                </syntacticPosition>
            </syntacticConstruction>
        </constructionSet>
    </sense>
```

Figure 4-18 LMF : composant syntaxique – Exemple de l'instanciation XML [SAL 05]

Nous n'en sommes pas encore au point de pouvoir proposer un codage concret pour le lexique syntaxique du vietnamien : cela exige une définition de l'ensemble des catégories de données utilisées ainsi que des extensions lexicales à prendre en compte, et enfin l'instanciation XML du modèle ainsi défini.

Takenobu *et al.* [TAK 06] présentent un projet de construction d'un lexique multilingue de base pour les langues asiatiques. Le papier expose des problèmes en appliquant le modèle MILE (*cf.* Projets multilingues de la section 1.1.1), un des modèles à visées pré-normatives mis en considération pour la définition du LMF, pour les langues asiatiques telles que le chinois, le coréen, le japonais et le thai. Nous pouvons constater que les spécificités de ces langues sont également celles que nous rencontrons pour le vietnamien, que nous avons présentées au Chapitre 2 et analysées aux Chapitres 3 et 4. Nous espérons que la définition d'un modèle conforme au LMF pour les langues asiatiques avec les relations étendues et des catégories de données communes spécifiques pour les langues asiatiques serait possible grâce à ce projet.

4.5.1.2. Un lexique syntaxique pour le vietnamien

Au Chapitre 3, nous avons construit un lexique pour l'annotation morphosyntaxique. Ce lexique contient des descriptions lexicales de chaque mot, qui indique la cooccurrence des mots appartenant aux différentes catégories grammaticales.

Ces descripteurs sont naturellement réutilisables pour le lexique syntaxique. Certaines classes de mots, notamment les classes des noms et des verbes, ainsi que des adjectifs, ont besoin d'informations plus détaillées. En effet, à chaque verbe ou adjectif (prédicat) devraient être associées les informations sur sa structure argumentale, ses compléments circonstanciels potentiels, *etc*. Pour un nom commun, on a besoin de connaissances sur ses classificateurs possibles. Nous considérons par exemple les verbes, dont les informations sont les plus riches.

Les verbes sont caractérisés, dans le lexique existant, par deux attributs : la gradabilité (*gradability*) et le sens. Le premier attribut reflète la capacité d'un verbe qui peut être modifié par un adverbe de degré (comme « très »). La classification selon l'attribut *sens* a en réalité pour but de regrouper des verbes qui partagent les mêmes structures argumentales (*cf.* 3.3.2.4). Pour le lexique syntaxique, il nous faut donc détailler la description de ces structures, en partant du lexique existant et en s'appuyant sur le modèle de représentation du lexique introduit dans la section 4.5.1.2.

Au niveau de gestion du nouveau lexique, nous suivons le même principe de partage des ressources et d'amélioration coopérative que nous avons adopté pour le lexique morphosyntaxique ou le corpus annoté (*cf.* 3.5.1).

Nous abordons maintenant la construction de la grammaire à large couverture et des jeux de phrases de test qui peuvent servir ultérieurement à l'évaluation des grammaires.

4.5.2. Construction de la grammaire et des jeux de phrases de test

Le processus de construction d'une grammaire par parcours méthodique des phénomènes syntaxiques du vietnamien peut être accompagnée d'une construction de jeux de phrases de test illustrant ces phénomènes, qui servent à l'évaluation des analyseurs syntaxiques (*cf.* 4.2.2.2). En particulier, les exemples que nous avons analysés comme illustrations des phénomènes syntaxiques principaux du vietnamien font partie des exemples que nous pouvons utiliser pour établir l'ensemble de phrases de test pour l'analyse syntaxique du vietnamien. Notons que les phrases agrammaticales appartiennent également à l'ensemble de phrases de test, car elles permettent de contrôler la surgénération d'une grammaire.

Si la grammaire que nous tentons de construire est fondée sur le formalisme TAG, les jeux des phrases de test, tout comme le lexique syntaxique, doivent en revanche être élaborés de manière à être réutilisables, indépendamment du formalisme retenu.

La réutilisabilité et la spécificité d'un ensemble de phrases de test font partie de l'objectif du projet TSNLP de la Communauté européenne (*cf.* Lehmann *et al.* [LEH 96]). Ce projet a proposé une méthodologie pour la construction systématique et progressive de phrases-test, ainsi qu'un schéma d'annotations détaillé et neutre tant vis-à-vis des théories linguistiques spécifiques et que des types d'applications particuliers.

Une base de phrases de test selon la méthodologie du TSNLP peut être établie en se fondant sur les éléments suivants :

- Une liste canonique des catégories et des fonctions utilisées pour la représentation de constituants et de dépendances, idéalement indépendante des particularités présentées par telle ou telle théorie linguistique.
- Une liste de phénomènes linguistiques dont une grammaire doit être à même de rendre compte. Par exemple pour les trois langues française, anglaise et allemande sont sélectionnés ces phénomènes : complémentation, accord (non applicable au vietnamien), modification, diathèse, mode, temps, aspect, type de phrase (interrogative, exclamative), ordre des mots, coordination, négation, phénomènes extra-grammaticaux (ponctuation, abréviations, *etc.*).
- Ces phénomènes sont classés selon leur domaine, c'est-à-dire le cadre syntaxique dans lequel ils se produisent (phrase, proposition, syntagme nominal, *etc.*)
- Les paramètres d'un phénomène caractérisent ses propriétés. La dérivation des phrases de test agrammaticales est effectuée par génération de variantes : des variations sont appliquées de façon systématique sur les exemples par l'application d'une ou plusieurs opérations de test linguistique suivantes : remplacement, ajout, suppression, permutation.

L'élaboration des phrases de test elles-mêmes doit être sous contrôle : le vocabulaire doit être le moins ambigu possible, chaque phrase doit être le plus simple possible, par exemple les modifieurs, s'il y en a, servent uniquement à l'illustration des phénomènes testés.

Pour clore ce chapitre, nous étudions enfin la problématique de construction d'un corpus arboré du vietnamien.

4.5.3. Construction du corpus arboré

Par rapport aux corpus annotés morpho-syntaxiquement, qui associent à chaque mot des informations morphologiques et sa catégorie syntaxique, les corpus annotés syntaxiquement fournissent des informations supplémentaires : découpage des constituants (proposition, syntagme, *etc.*), fonction grammaticale des mots ou des constituants, dépendance entre mots ou constituants. Ils peuvent avoir plusieurs applications directes en TAL :

- évaluation des étiqueteurs et des analyseurs syntaxiques ;
- entraînement des étiqueteurs ou des analyseurs probabilistes ;
- recensement des constructions négligées dans la littérature linguistique ;
- enrichissement des dictionnaires (extraction de collocations, cadres de complémentation) ;
- *etc.*

Ainsi, pour chaque langue, la disponibilité d'un corpus arboré de référence de grande taille est précieuse pour toutes les applications en TAL qui font usage d'informations syntaxiques (extraction d'information, résumé de texte, alignement de texte multilingue, *etc.*). L'annotation syntaxique est aussi une préparation à l'annotation sémantique.

Avant d'étudier le codage normalisé des corpus arborés, nous présentons dans la section suivante la structure (les informations annotées) des corpus arborés.

4.5.3.1. Structure des corpus arborés

L'ouvrage édité par Anne Abeillé ([ABE 03a]) présente de nombreux projets de construction des corpus arborés (*Treebank*) d'une assez grande variété de langues. En général, les corpus arborés sont construits en plusieurs étapes :

- Annotation morphosyntaxique
- Analyse syntaxique partielle
- Analyse syntaxique profonde

Les corpus suivant le modèle du *Penn Treebank* (introduit à la section 1.1.4) représentent les informations syntaxiques en réalisant un parenthésage hiérarchique des constituants de la phrase, et en associant les rôles sémantiques de la structure prédicat-argument à ces constituants (*cf.* Figure 4-19).

```
((S      (NP-SBJ-1 Jones)
         (VP followed)
         (NP him)
         (PP-DIR       into
                (NP the front room))
         ,
         (S-ADV (NP-SBJ *-1)
                (VP    closing
                       (NP the door)
                       (PP    behind
                              (NP him)))))
         .))
```

Figure 4-19 Exemple d'annotation syntaxique dans le corpus Penn Treebank

D'autres corpus arborés, dont le modèle *Prague Dependency Treebank* est représentatif, reposent uniquement sur la grammaire de dépendance (dépendance et rôle sémantique des mots). Les Figure 4-20 et Figure 4-21 montrent des exemples de l'annotation de dépendances.

```
subj(intend,Paul,_)
xcomp(intend,leave,to)
subj(leave,Paul)
dobj(leave,IBM,_)
```

Figure 4-20 Exemple d'annotation de dépendances ([CAR 03])

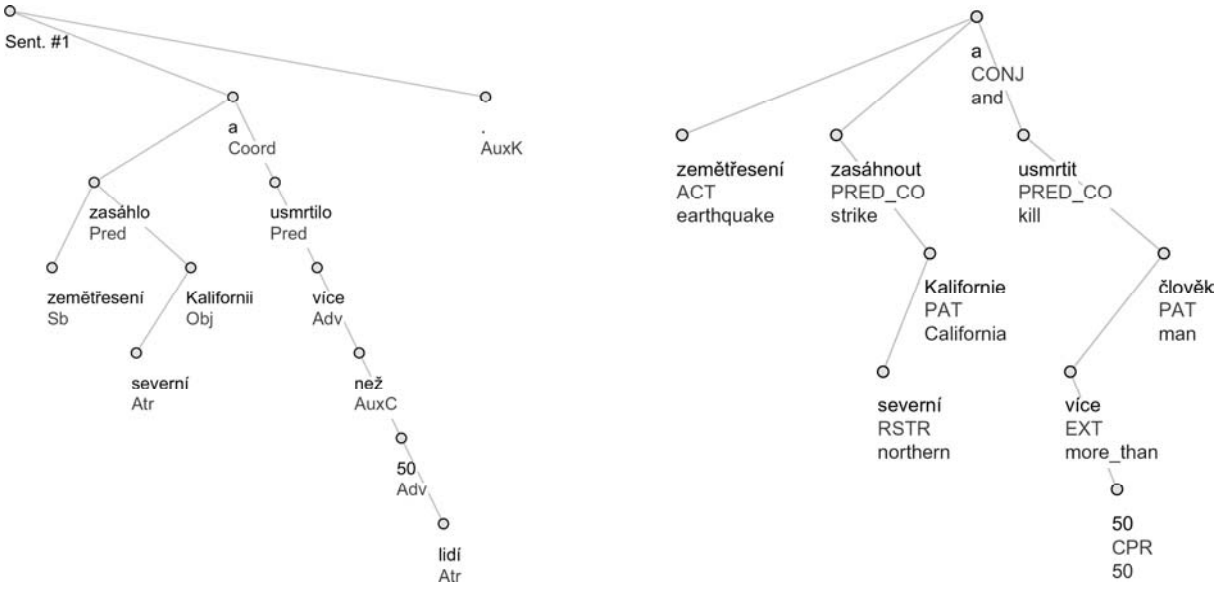


Figure 4-21 Exemple de l'annotation de dépendances du tchèque [CME 04]

Le projet NEGRA/TIGER (*cf.* Brants *et al.* [BRA 03]) adopte un modèle hybride, qui combine l'annotation de constituants et de dépendances pour l'allemand. Ce modèle autorise des relations de dépendances croisées. Ce choix permet notamment de simplifier la représentation des constituants discontinus de l'allemand, langue pour laquelle l'ordre de mot est très flexible. Ce projet a également donné lieu à la définition du format TIGER-XML pour le codage des formes et fonctions syntaxiques, qui est un des formats reconnus au niveau international.

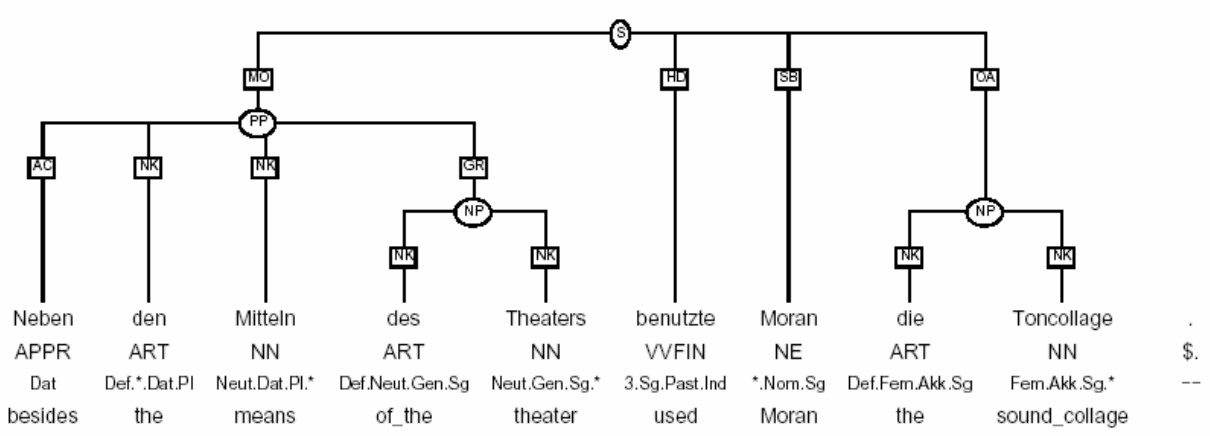


Figure 4-22 Exemple de l'annotation du corpus NEGRA/TIGER

Un autre corpus richement annoté est le *Sinica Treebank* pour le chinois (*cf.* [CHE 03]). Il contient l'annotation des constituants de phrase ainsi que, comme le *Prague Dependency Treebank*, des rôles sémantiques de chaque mot. On peut citer également le corpus arboré du polonais, qui vise à devenir une base de test pour une grammaire HPSG de cette langue (*cf.* Marciniak *et al.* [MAR 03]).

En résumé, un corpus arboré richement annoté contiendra des informations suivantes :

- Catégorie syntaxique/rôle grammatical des mots/constituants
- Parenthésage (ou dépendance) des constituants
- Dépendance des rôles grammaticaux des mots/constituants.

Le paragraphe suivant présente une proposition de codage normalisé des corpus arborés, dans le but de maximiser l'échange, l'évaluation et la réutilisabilité de ces ressources.

4.5.3.2. Codage normalisé du corpus arboré

Ide et Romary ([IDE 03]) proposent un modèle abstrait pour différents types d'annotation (morphosyntaxique, syntaxique, co-référence, *etc.*) qui peut être instancié de plusieurs façons selon l'approche et le but de l'annotateur. Ce modèle, ainsi que plusieurs instanciations, ont été implémentées en employant des schémas XML et RDF (*Resource Definition Framework*), et incorporés dans XCES[66] (*cf.* Ide *et al.* [IDE 00b]), qui est en relation étroite avec le travail sur la définition du cadre d'annotation linguistique mené par le comité ISO/TC 37/SC 4 (*cf.* 1.2.2).

Suivant un principe similaire aux travaux menés par le SC 4, l'objectif cette recherche est de définir un méta-modèle et une spécification des catégories de données communes pour chaque type d'annotation. Le modèle concret sera obtenu par une instanciation du méta-modèle et une interprétation des catégories de données.

Le modèle sous-jacent pour l'annotation syntaxique spécifie des relations constitué/constituant entre des composants grammaticaux ou syntaxiques. Ces relations peuvent être soit modélisées par une structure arborescente, soit données explicitement (*cf.* les exemples de la section 4.5.3.1).

L'instanciation XML du méta-modèle proposé utilise les balises suivantes pour représenter les annotations syntaxiques utilisant des arbres (*cf.* Ide et Romary [IDE 03] pour les détails concernant les attributs de ces balises) :

- `<struct>` représente un nœud dans l'arbre ;
- `<feat>` inclut l'information attachée à un nœud ;
- `<alt>` permet de représenter les alternatives d'annotation en cas nécessaire ;
- `<rel>` est utilisé pour identifier un élément relié non adjacent ;
- `<seg>` référence aux données auxquelles est associée l'annotation, car il est recommandé d'utiliser une annotation externe (*stand-off*, *cf.* Bonhomme [BON 00a])

La hiérarchie de l'élément `<struct>` correspond à la structure de constituants de la phrase annotée. Ainsi, on peut dans un sens utiliser la grammaire sous-jacente de l'annotation syntaxique pour vérifier la grammaticalité de la phrase, et dans l'autre sens détecter les nouvelles structures non enregistrées dans la grammaire en utilisant des outils pour la génération automatique de la DTD.

Le modèle abstrait impose une distinction claire entre les informations implicite et explicite (par exemple des relations fonctionnelles déduites des relations structurelles des constituants), entre les relations syntagmatiques et fonctionnelles. Cela permet une comparaison plus aisée des schémas d'annotation.

La Figure 4-23 (*cf.* Ide et Romary [03]) présente le codage XML de l'annotation exemplifiée à la Figure 4-19. Les têtes (*head*) des relations, marquées en gras, sont implicites dans l'annotation originale. L'élément `<feat>`, associé au sujet implicite ([Jones]) de la phrase subordonnée, marqué en gras, est présent pour refléter le contenu de l'annotation originale.

Une annotation de dépendance telle que celle présentée à la Figure 4-20 est codée par une hiérarchie plate, comme le montre la Figure 4-24 (*cf.* Ide et Romary [03]).

Ce cadre XCES et ses supports XML et RDF doivent permettre à l'annotateur de se concentrer sur la spécification du schéma d'annotation syntaxique (c'est-à-dire des étiquettes morphosyntaxique, des types de constituant syntaxique, et des structures selon une théorie/modèle linguistique). L'application du schéma de codage XCES nous aide à obtenir des ressources annotées cohérentes, d'accès et d'utilisation faciles.

---

[66] Instanciation XML du Corpus Encoding Standard, faisant partie des recommandations développées par le groupe EAGLES

```
<struct id="s0">
 <feat type="Cat">S</feat>
 <struct id="s1">
     <rel type="SBJ" head="s2"/>
       <feat type="Cat">NP</feat>
     <seg target="xptr(substring(/p/s[1]/text(),1,5))"/>      <!-- Jones -->
 </struct>
 <struct id="s2">
     <feat type="Cat">VP</feat>
     <seg target="xptr(substring(/p/s[1]/text(),7,8))"/>      <!-- followed -->
    </struct>
    <struct id="s3">
     <feat type="Cat">NP</feat>
     <seg target="xptr(substring(/p/s[1]/text(),16,3))"/>      <!-- him -->
    </struct>
    <struct id="s4">
     <rel type="DIR" head="s2"/>
     <feat type="Cat">PP</feat>
     <seg target="xptr(substring(/p/s[1]/text(),20,4))"/>      <!-- into -->
     <struct id="s5">
            <feat type="Cat">NP</feat>
            <seg target="xptr(substring(/p/s[1]/text(),25,14))"/> <!--the room -->
     </struct>
    </struct>
 <struct id="s6">
     <rel type="ADV" head="s2"/>
     <feat type="Cat">S</feat>
     <struct id="s7" ref="s1">
         <rel type="SBJ" head="s8"/>                          <!-- [Jones] -->
         <feat type="Cat">NP</feat>
     </struct>
     <struct id="s8">
         <feat type="Cat">VP</feat>
         <seg target="xptr(substring(/p/s[1]/text(),41,7))"/>  <!-- closing -->
          <struct id="s9">
         <feat type="Cat">NP</feat>
         <seg target="xptr(substring(/p/s[1]/text(),49,8))"/>  <!-- the door -->
             <struct id="s10">
             <rel type="DIR" head="s8"/>
             <feat type="Cat">PP</feat>
             <seg target="xptr(substring(/p/s[1]/text(),57,6))"/> <!-- behind -->
                 <struct id="s11">
                 <feat type="Cat">NP</feat>
                 <seg target="xptr(substring(/p/s[1]/text(),64,3))"/><!-- him -->
             </struct>
         </struct>
     </struct>
    </struct>
</struct>
```

Figure 4-23 Codage XML abstrait pour l'exemple Penn TreeBank [IDE 03]

```
<struct>
  <rel type="subj"  head="mySentence.xml#w2" dependent="mySentence.xml#w1"/>
  <rel type="xcomp" head="mySentence.xml#w2" dependent="mySentence.xml#w4"
                introducer="mySentence.xml#w3"/>
  <rel type="subj"  head="mySentence.xml#w4" dependent="mySentence.xml#w1"/>
  <rel type="dobj"  head="mySentence.xml#w4" dependent="mySentence.xml#w5"/>
</struct>
```

Figure 4-24 Codage XML abstrait pour l'exemple de dépendances [IDE 03][67]

[67] Ce codage suppose que la phrase en question appartient à un document séparé `mySentence.xml` sous forme: `<s1><w1>Paul</w1><w2>intends</w2><w3>to</w3><w4>leave</w4><w5>IBM</w5></s1>`.

S'appuyant sur le principe du modèle de représentation de l'annotation syntaxique proposé ci-dessus, la définition du codage d'un corpus arboré du vietnamien est en cours dans le cadre de notre nouveau projet national de recherche en TAL.

4.5.3.3. Annotation syntaxique des textes vietnamiens

Nous souhaitons construire un corpus arboré richement annoté pour le vietnamien. Un tel corpus doit contenir, comme nous l'avons précisé plus haut, les informations suivantes :

- partie du discours pour chaque mot, accompagnée par une structure de traits contenant les informations venant du lexique syntaxique ;
- structures des constituants ;
- rôles thématiques reflétant les relations de dépendance entre les mots/constituants.

Le processus d'annotation habituel est :

- étiquetage de catégories syntaxiques ;
- analyse syntaxique par un analyseur syntaxique « profond » (*deep parser*) comme par exemple l'analyseur LTAG, ou par un analyseur partiel (*shallow parser*) ;
- révision manuelle.

Suivant le même principe que nous avons adopté pour la constitution d'autres ressources linguistiques de taille importante, nous devons assurer la disponibilité des ressources construites sur la Toile, en créant et maintenant un mécanisme pour le partage et la construction coopérative de ces ressources (*cf*. 3.5.1).

Pour tout cela, il faut étudier et développer les outils suivants :

- analyseur partiel pour le vietnamien ;
- documentation et guide d'annotation syntaxique ;
- outils d'accès et d'édition des annotations syntaxiques en ligne ;
- formulaire de contribution de nouvelles annotations en ligne.

Une licence pour assurer une bonne distribution de ces ressources est également nécessaire.

En conclusion, afin de construire des outils et ressources linguistiques pour l'analyse syntaxique du vietnamien, nous avons présenté les premiers efforts vers la construction d'une grammaire à large couverture avec le formalisme TAG : il s'agit de modéliser les groupes nominaux en vietnamien, et d'étudier la structure noyau des phrases pour pouvoir identifier les spécificités du vietnamien par rapport aux langues indo-européennes comme le français. Cela nous permet de prévoir la possibilité de modéliser la grammaire vietnamienne avec le formalisme TAG. Nous avons ensuite présenté les pistes à suivre pour construire les ressources linguistiques importantes pour l'analyse syntaxique en TAL comme le lexique syntaxique, la base de phrases de test, le corpus annoté syntaxiquement, en étudiant l'état de l'art sur la construction de ces ressources. Cette construction ne peut pas être, bien entendu, réalisée dans le cadre de notre thèse, mais le plan de travail proposé pourra trouver son cadre de réalisation dans les prochains projets de recherche en TAL au Vietnam.

Dans le chapitre suivant, nous présentons les travaux sur l'alignement multilingue, le sujet original de notre projet de thèse.

Chapitre 5

Traitement de corpus multilingues français - vietnamiens

Nous présentons dans ce chapitre les problèmes concernant l'alignement de textes multilingues. Nous nous intéressons en particulier à deux aspects de cette problématique : l'alignement au niveau des phrases et celui au niveau des mots (unités lexicales), pour lesquels nous avons développé deux outils spécialisés. Pour l'alignement au niveau des phrases, nous disposons d'un outil fondant son analyse sur la structure hiérarchique des documents, qui s'est montré d'une grande efficacité pour le couple de langues français-anglais dans le cadre de la campagne d'évaluation ARCADE I. Notre première tâche est donc d'évaluer l'adaptation de cet outil aux textes français-vietnamiens. Nous développons ensuite un outil d'alignement au niveau des unités lexicales. Dans le temps limité de la thèse, nous ne réalisons qu'une rapide évaluation de l'application de cet outil à chaque couple de langues d'un texte multilingue français - vietnamien - anglais, dont chaque texte est soumis à un pré-traitement lexical, afin de montrer la perspective de la technique utilisée. Nous présentons également l'évaluation de notre outil sur des corpus en dix langues différentes dans le cadre du projet ARCADE II.

- *Introduction : Alignement multilingue*
- *Méthodologie d'alignement*
- *Construction de corpus multilingues et codage de données*
- *Alignement structurel*
- *Alignement lexical*
- *Combinaison des approches structurelle et lexicale*
- *Participation à la campagne ARCADE II*
- *Bilan et perspectives*

5.1. Introduction

L'alignement de corpus multilingues, ou textes parallèles, consiste à apparier automatiquement des unités de différents niveaux dans deux textes qui sont la traduction l'une de l'autre: paragraphes, phrases, mots et expressions, *etc*.

Un texte multilingue aligné constitue une source d'information utilisable pour un vaste ensemble d'applications : traduction, recherches en terminologie et lexicographie multilingue, recherche d'information multilingue, désambiguïsation du sens des mots dans des textes, enseignement des langues, recherches linguistiques comparatives, étude de la traduction, *etc*. (Véronis [VER 00b]).

Les techniques d'alignement peuvent intervenir lors de la création même de textes parallèles, et peuvent ainsi fournir un support à la création et à la maintenance de documents multilingues (Isabelle et *al*. [ISA 93]).

Pour les langues occidentales, les systèmes automatiques d'alignement au niveau des phrases atteignent aujourd'hui des résultats plus que satisfaisants. Ils utilisent, pour beaucoup d'entre eux, uniquement des informations d'ordre statistique concernant les corpus comparés, ce qui leur procure une quasi-indépendance vis-à-vis des couples de langues étudiés. Les résultats qu'ils procurent sont tout à fait pertinent puisqu'en fonction des types de textes concernés et leurs qualités, le taux d'erreur atteint rarement les 8% et se situe habituellement en-dessous de 5%. Il est envisagé d'effectuer un alignement plus fin, c'est à dire au niveau des syntagmes et des mots. C'est une inflexion qu'a intégré le projet Arcade I (*cf.5.2.2*) de l'Agence Universitaire de la Francophonie, avec le lancement d'une campagne d'évaluation des systèmes d'alignement au niveau des mots à partir de l'année 1998. Ce nouveau type d'alignement demande cependant un usage plus important de la composante linguistique et nécessite une prise en compte des spécificités, notamment morphosyntaxique, de chacune des langues concernées.

L'équipe Langue et Dialogue du LORIA a participé au projet ARCADE I avec un système d'alignement au niveau des phrases qui se fond sure la structure logique hiérarchique des documents (*cf.* 1.2.1). Nous avons amélioré ce système (Nguyen [NGU 99]) afin d'introduire une plus grande flexibilité en lui permettant de détecter les cas où le codage structurel du corpus dans les différentes langues considérées n'est pas homogène, et d'adapter son fonctionnement en conséquence. Nous avons également travaillé à l'extension du domaine d'applicabilité du système, en introduisant le support de l'UTF-8, d'une part, et surtout en définissant des métriques statistiques ne s'appuyant que sur les observations réalisées sur le corpus, et donc totalement indépendantes des langues traitées.

Notre premier objectif est d'évaluer l'application de ce système sur le couple de langues français-vietnamien et, par intérêt comparatif, anglais-vietnamien. Le deuxième objectif est d'implémenter et évaluer un système d'alignement lexical, qui fait l'appel à l'information de lemmatisation des textes français ou anglais, et d'étiquetage morphosyntaxique des textes vietnamiens. Nous décrivons et évaluons également un système original combinant ces deux alignements structurel et lexical, qui peut permettre d'atteindre des résultats de qualité supérieure dans le cas où les textes présentent des différences importantes (traductions parcellaires, inexactes, *etc.*).

Dans ce chapitre, nous présentons dans un premier temps la méthodologie de l'alignement multilingue (techniques d'alignement, mesures d'évaluation). Dans un second temps, nous présentons le travail que nous avons accompli au cours de cette thèse, consistant à :

- collecter et normaliser le codage d'un corpus de bitextes français – vietnamien et anglais – vietnamien
- développer un système d'alignement multilingue basé sur la structure hiérarchique des documents et sur les informations lexicales (textes lemmatisés).

5.2. Méthodologie d'alignement

5.2.1. Méthodes d'alignement

L'état de l'art de l'alignement multilingue est abondamment décrit dans l'ouvrage *Parallel Text Processing* édité par Jean Véronis [VER 00]. Nous présentons dans cette section les traits les plus importants des méthodologies rencontrées dans la littérature.

5.2.1.1. Alignement de phrases

Deux familles d'approches différentes peuvent être distinguées, dans la lignée de deux études initiales qui, malgré leurs différences, reposent sur un certain nombre d'hypothèses simplificatrices communes.

Kay et Röscheisen [KAY 93] font l'hypothèse que, pour que des phrases soient en correspondance de traduction, il faut que les mots qui les composent soient également en correspondance. Cette hypothèse ne fait appel qu'à une information interne, c'est-à-dire que toute l'information nécessaire (et en particulier les correspondances lexicales) est dérivée des textes à aligner eux-mêmes. Les auteurs utilisent le fait qu'un tel alignement des mots, même très grossier et très imparfait, peut conduire à un alignement satisfaisant au niveau des phrases. Le point de départ de l'algorithme est un ensemble initial de phrases raisonnablement candidates à l'alignement : la première phrase et la dernière ont de bonnes chances de se correspondre dans chaque texte, et les phrases intermédiaires sont certainement en correspondance dans un couloir diagonal relativement étroit. L'algorithme compare ensuite la distribution des mots de cet ensemble de phrases dans chacun des textes et fait l'hypothèse que si ces distributions sont similaires au-delà d'un certain seuil pour un couple de mot donné, ces mots ont de bonnes chances d'être en relation de traduction. Les mots en question fournissent alors un ensemble de points d'ancrage qui permette de réduire le couloir diagonal des alignements de phrases candidats. La procédure est itérée jusqu'à convergence vers une solution minimale.

Gale et Church [GAL 91, 93] proposent une méthode qui n'utilise également qu'une information interne, mais ne fait aucune hypothèse directe sur le contenu lexical des phrases. Les auteurs partent de la constatation que la longueur des phrases dans le texte source et de leurs traductions dans le texte cible sont fortement corrélées. De plus, il semble exister un rapport assez constant entre les longueurs de phrases d'une langue à l'autre en termes de nombre de caractères (ainsi, il est connu que les textes français sont plus longs que leurs équivalents anglais : ce rapport est de l'ordre de 1,1 et varie peu selon le genre des textes). Cette observation permet de construire un modèle probabiliste et une mesure de dissimilarité entre phrases des deux textes à aligner, qui prennent en compte la proportion des types d'alignements attendus $m : n$ (m phrases dans le texte source correspondent aux n phrases dans le texte cible). Pour de raisons de calculabilité, Gale et Church sont amenés à faire des hypothèses simplificatrices, et en particulier à réduire le cas ($m : n$) à m, $n \leq 2$ (*cf.* Tableau 5-1). L'alignement optimal peut alors être calculé de façon efficace par un algorithme classique de programmation dynamique. Brown et *al.* [BRO 91] utilisent également la même idée de corrélation entre les longueurs de phrases, mais ils formulent le problème à l'aide de modèles de Markov cachés.

5.2.1.2. Alignement de mots et expressions

L'alignement ou l'extraction de lexiques consiste théoriquement en deux phases :

- détecter les mots et les expressions dans le texte source et le texte cible,
- mettre ces mots en correspondance.

Nombre de phrases du texte source	Nombre de phrases du texte cible	Type de traduction
1	1	Substitution
2	1	Compression
1	2	Extension
2	2	Mélange
1	0	Destruction
0	1	Insertion
> 1	0	Large destruction
0	> 1	Large insertion

Tableau 5-1 Différents types de traduction

Plusieurs méthodes statistiques ont été proposées pour choisir des expressions complexes d'une langue. Pourtant les méthodes purement statistiques ne peuvent pas facilement découvrir des opérations linguistiques réalisées sur des expressions « semi-figées » qui sont très fréquentes. En conséquence, certaines approches linguistiques ont été proposées seules ou en combinaison avec des méthodes statistiques. Ces méthodes se basent normalement sur des expressions régulières et des grammaires locales.

5.2.1.3. Alignement de clauses et de structures de phrase

L'alignement des textes à un niveau supérieur aux mots ou expressions et inférieur à la phrase, comme par exemple des clauses ou des fragments d'arbres syntaxiques, pourrait être très utile pour les applications comme la traduction fondée sur l'exemple, l'étude comparative des langues *etc*. Mais l'alignement à ce niveau soulève de grandes difficultés, car pour cela il faut d'abord détecter les frontières des clauses ou les structures syntaxiques des textes, ce qui est une tâche très complexe. Un second problème, encore plus délicat, naît de la grande différence de structure syntaxique pouvant exister entre deux langues.

Plusieurs références sur ce problème sont listées dans Véronis [VER 00b].

5.2.2. Évaluation - Projets ARCADE I & II

L'action ARCADE I (1996-1999) financée par l'AUPELF-UREF (maintenant AUF) visait deux objectifs principaux (*cf*. Véronis et Langlais [VER 00a]) :

- produire un grand corpus standardisé de textes multilingues alignés ;
- réaliser une avancée méthodologique sur les techniques et algorithmes d'alignement.

Elle a montré que la qualité de l'alignement de phrases était fortement dépendante du degré de parallélisme structurel des documents concernés. Sur des textes traduits avec un soin extrême de parallélisme, la performance des meilleurs systèmes atteint environ 98% d'alignements corrects. Par contre, face aux cas de non-parallélisme du à des causes diverses : omissions du traducteur, différences de version, traductions abrégées, des glossaires techniques en ordre différent dans les différentes langues, *etc*., tous les systèmes présentés dans l'ARCADE I ont montré une dégradation rapide et très importante.

La campagne ARCADE II a pris la suite d'ARCADE I en octobre 2002. Dans ce nouveau projet, deux tâches d'alignement sont évaluées : alignement des phrases et alignement des entités nommées.

Pour la première tâche, les évaluations de l'alignement du corpus JOC ((Journal officiel de la Communauté Européenne, *cf*. 1.1.5.1) de l'action d'Arcade I seront reproduites, afin d'identifier les évolutions réalisées depuis 1998, et de fournir une base de comparaison aux systèmes participants. En plus de l'anglais, trois autres langues (allemand, espagnol, italien) sont intégrées, le français demeurant la langue pivot. Une deuxième évaluation porte ensuite sur l'alignement d'un corpus extrait des archives du mensuel *Le monde diplomatique*, où le français est toujours la langue pivot, et 6 autres langues sont prises en compte : arabe, chinois, grec, japonais, persan et russe.

Concernant la deuxième tâche, l'évaluation porte sur un corpus bilingue français-arabe, dont les entitées nommées dans la partie en français sont précisément annotées. Les systèmes participants ont pour but d'identifier les entitées nommées correspondant dans la partie en arabe, qui, pour sa part, n'est pas annotée.

Le format d'annotation des corpus est défini en prenant en compte les différentes normes et recommandations applicables : TEI (*Text Encoding Initiative*), CESAlign (*Corpus Encoding Standard*) et le standard TMX (*Translation Memory Interchange*).

Les métriques d'évaluation : Pour évaluer un alignement de phrases A par rapport à un alignement de référence A_{ref}, on utilise les mesures de précision et rappel :

$$\text{Rappel} = \frac{\left|A \cap A_{ref}\right|}{\left|A_{ref}\right|} \qquad \text{Précision} = \frac{\left|A \cap A_{ref}\right|}{\left|A\right|}$$

Le cardinal d'un alignement est calculé comme étant sa surface dans l'espace à deux dimensions formé par le produit cartésien des deux textes, l'unité de longueur étant calculée en nombre de phrases, de mots ou de caractères. Il a été montré dans le cadre du projet ARCADE I que ces mesures sont fortement corrélées. En pratique, une évaluation reposant sur le nombre de caractères est donc préférable, car elle ne dépend pas de la segmentation du texte. Pour ARCADE II, la mesure utilisée prend en compte le nombre de caractères hors espaces, ce qui se justifie par le fait que certaines des langues considérées n'utilisent pas (ou très peu) d'espaces.

ARCADE favorise l'utilisation de la *F*-mesure, qui combine le rappel et la précision dans une seule mesure :

$$\text{F} = 2\frac{\text{rappel} \times \text{précision}}{\text{rappel} + \text{précision}}$$

Pour l'évaluation de la détection des zones non parallèles (omissions, ajouts, interversions), les mesures de rappel et de précision ne concerneront qu'un côté du bi-texte, et donc une seule dimension. Pour le repérage de traduction (ainsi que le repérage des cognats), précision et rappel sont calculées au niveau de chaque appariement :

$$\text{Rappel} = \frac{\text{Mots corrects}}{\text{Mots de l'appariement de référence}}$$

$$\text{Précision} = \frac{\text{Mots corrects}}{\text{Mots proposés par le système}}$$

Un appariement vide est considéré comme un appariement avec un mot spécial (null). Des moyennes différentes en fonction des classes d'unités (partie du discours) peuvent être calculées pour l'ensemble des appariements.

5.2.3. Plan de la présentation

Dans le cadre de cette thèse, nous avons été amenée d'une part à nous intéresser aux problèmes d'alignement au niveau des phrases, puis des mots, d'autre part à proposer une méthode permettant de combiner ces deux approches complémentaires. La présentation de nos travaux est donc réalisée en trois parties :

- Nous traitons dans un premier temps la problématique d'alignement au niveau de phrases (section 5.4). Nous disposons d'un outil fondant son analyse sur la structure hiérarchique des documents, qui s'est montré d'une grande efficacité pour le couple de langues français-anglais dans le cadre de la campagne d'évaluation ARCADE I. Notre tâche est donc d'évaluer l'adaptation de cet outil aux textes français-vietnamiens, et au passage, anglais-vietnamiens, en comparant les résultats obtenus sur ces textes avec ceux sur les mêmes textes français-anglais. Afin d'améliorer l'adaptation de l'algorithme au cas de l'alignement au vietnamien, nous sommes amenée à introduire une nouvelle méthode de mesure de probabilité d'association entre phrases indépendante des langues considérées.
- Dans un deuxième temps, nous abordons la question de l'alignement au niveau des mots (section 5.5). Nous développons à partir d'une méthode classique un outil d'alignement au niveau des unités lexicales. Un alignement à ce niveau peut être utilisé pour améliorer l'alignement phrastique, au cas de rupture fréquente ou de codage grossier du corpus parallèle à aligner. Dans le temps limité de la thèse, nous ne faisons qu'une petite évaluation de l'application de cet outil à chaque coupe de langues d'un texte multilingue français - vietnamien - anglais, dont chaque texte est passé à un pré-traitement lexical, afin de montrer la perspective de la technique utilisée.
- Enfin, nous présentons à la section 5.6 une expérience de mise en place d'une boucle de rétroaction par laquelle le résultat de l'alignement lexical (au niveau des mots) permet de renforcer l'alignement structurel (au niveau de segments de texte), et réciproquement. Nous évaluons l'apport de cette méthode par rapport aux algorithmes originaux présentés auparavant.

Nous terminons ce chapitre sur une présentation de la campagne d'évaluation ARCADE II, et des résultats obtenus par notre soumission. Avant de présenter ces divers points, nous discutons brièvement à la section suivante de la construction et du codage des corpus multilingues alignés.

5.3. Construction de corpus multilingues et codage de données

5.3.1. Construction de corpus multilingues

La construction de corpus parallèles est relativement facile pour les langues occidentales beaucoup pratiquées, comme l'anglais et le français : une assez longue « histoire » de numérisation de documents textuels et de constitution de corpus monolingues pour les langues considérées rend souvent possible l'exploitation de données existantes – même si subsistent dans ces conditions les difficultés liées à l'existence de traduction inexactes ou incomplètes. Pour les langues moins pratiquées, ou pratiquées dans les pays moins développés technologiquement, la constitution de corpus parallèles peut demander beaucoup plus d'efforts (*cf*. Singh *et al.* [SIN 00]) : la quantité de textes parallèles (numérisés ou non) peut être beaucoup moins importante, les documents sont moins couramment numérisés, et enfin, dans le cas des langues n'employant pas l'alphabet latin, les codages des textes dans les différentes langues considérées peuvent s'avérer incompatibles.

Notre but est de collecter les textes existant dans le couple de langues français-vietnamien et aussi anglais-vietnamien pour construire un corpus de référence pour l'alignement multilingue. Si le problème de codage incompatible de caractères vietnamiens peut toujours être réglé par de petits outils de conversion *ad hoc*, nous avons plus de difficulté pour trouver les textes bilingues sous forme électronique[68]. Nous exploitons, dans un premier temps, les sources suivantes pour le couple de langues français-vietnamien :

- Internet : des nouvelles françaises traduites en vietnamien, et inversement ;
- Maison du Droit Vietnamo-Française à Hanoï : des textes de droit vietnamien-français;
- ADETEF-Vietnam (Association pour le Développement des Échanges en Technologies Économiques et Financières) : des textes d'économie français-vietnamien et vietnamien-français.

Depuis quelques années, les textes multilingues contenant le vietnamien sont de plus en plus disponibles sur l'Internet. Nous en reparlons en conclusion de ce chapitre, dans la section 5.8.

5.3.2. Codage des corpus multilingues et alignés

Chaque texte de notre corpus multilingue est codé en XML selon le schéma recommandé par la TEI (*cf*. 1.2.1.3). Les informations sur le texte sont donc enregistrées dans la partie TEI-header, tandis que la structure du document est codée grâce à trois niveaux de balisage : les divisions correspondant aux chapitres, sections *etc*. qui peuvent être récursives, les paragraphes correspondant aux paragraphes physiques du texte, et les segments correspondant aux phrases ou aux fragments de texte similaires à une phrase (*cf*. l'exemple de la Figure 5-1).

Le codage des alignements fait usage des éléments <linkGrp> (groupe d'éléments de lien <link>) pour représenter l'ensemble des alignements (*cf*. Romary et Bonhomme [ROM 00a]).

[68] Sous forme papier, il existe en revanche naturellement beaucoup de traductions d'ouvrages connus au niveau internationnal.

```
<p id="d1p7" TEIform="p">
  <seg id="n16" part="N" TEIform="seg">Les grandes personnes m'ont conseillé de laisser de côté les dessins de serpents boas ouverts ou fermés, et de m'intéresser plutôt à la géographie, à l'histoire, au calcul et à la grammaire.</seg>
  <seg id="n17" part="N" TEIform="seg">C'est ainsi que j'ai abandonné, à l'age de six ans, une magnifique carrière de peintre.</seg>
  <seg id="n18" part="N" TEIform="seg">J'avais été découragé par l'insuccès de mon dessin numéro 1 et de mon dessin numéro 2.</seg>
  <seg id="n19" part="N" TEIform="seg">Les grandes personnes ne comprennent jamais rien toutes seules, et c'est fatigant, pour les enfants, de toujours et toujours leur donner des explications.</seg>
 </p>
```

Figure 5-1 Exemple de codage d'une version de notre corpus suivant les recommandations TEI

Nous produisons également les résultats d'alignement sous le format utilisé dans le cadre du projet ARCADE II, dont voici un exemple :

```
<ALIGN n="1.1" id="JOC081D1A1">
  <LANG lang="fr">
    <S>
      Objet: Garanties nucléaires soviétiques à prévoir dans les accords énergétiques
      <BR />
    </S>
  </LANG>
  <LANG lang="en">
    <S>
      Subject: Nuclear guarantees to be provided by the Soviet Union in respect of energy agreements
      <BR />
    </S>
  </LANG>
</ALIGN>
```

Figure 5-2 Exemple de codage d'alignement multilingue selon le format défini pour ARCADE II

Dans les sections suivantes, nous exposons les différentes méthodes d'alignement implémentées et leur évaluation.

5.4. Alignement structurel

Dans cette section, nous rappelons dans un premier temps la méthode mise en œuvre pour l'alignement au niveau de phrases au sein du LORIA (*cf.* Nguyen T. M. Huyen [NGU 99]) : modèle statistique de Gale et Church [GAL 93] appliqué à des documents structurés. Dans un deuxième temps, nous présentons l'évaluation des résultats sur le corpus français-vietnamien.

5.4.1. Méthode mise en œuvre

5.4.1.1. Algorithme de Gale et Church pour l'alignement des phrases

Gale et Church proposent d'utiliser l'algorithme de DTW (*Dynamic Time Warping*) pour l'alignement des phrases, en se fondant sur deux hypothèses simplificatrices :

- Les longueurs en caractères des phrases dans les deux langues source et cible sont fortement corrélées ;
- Seulement quatre types d'alignements sont pris en compte (*cf.* Tableau 5-2).

Catégorie	Probabilité P(*match*)
1-1 (substitution)	0,89
1-0 ou 0-1 (suppression ou insertion)	0,0099
2-1 ou 1-2 (compression ou extension)	0,0089
2-2 (mélange)	0,011

Tableau 5-2 Probabilités des types d'alignement

La mesure de distance est estimée par la formule suivante :

$$-\log P(match \mid \delta)$$

où :

$\delta = (l_2 - l_1 c) / \sqrt{l_1 s^2}$, ayant une distribution normale,

l_1 et l_2 sont les longueurs respectives des phrases considérées dans les langue L_1 et L_2,

c est le nombre moyen de caractères de la langue L_2 par caractère de la langue L_1,

et s^2 est la variance de cette distribution.

La moyenne c et la variance s^2 sont calculées empiriquement à partir des données disponibles. La valeur de c pour le couple anglais-allemand est de 1,1, pour le couple anglais-français est de 1,06. Les valeurs de s^2 pour ces deux couples sont respectivement 7,3 et 5,6. L'expérience montre qu'un choix de paramètres communs pour ces deux couples : c = 1 et s^2 = 6,8, n'influe que très marginalement sur le résultat d'alignement.

$P(match \mid \delta)$ est calculé par l'intermédiaire de $P(match)$ et $P(\delta \mid match)$, qui est, à son tour, estimée par la formule $2(1 - \int_0^{\delta} Norm(t)dt)$ (Norm représentant la distribution normale).

La fonction de distance *d* est définie sur 4 arguments : x_1, x_2, y_1, y_2 :

$d(x_1, y_1; 0, 0)$ est le coût de substitution de x_1 par x_2 ;

$d(x_1, 0; 0, 0)$ est le coût de suppression de x_1 ;

$d(0, y_1; 0, 0)$ est le coût d'insertion de y_1 ;

$d(x_1, y_1; x_2, 0)$ est le coût de compression de x_1 et x_2, appariée à y_1 ;

$d(x_1, y_1; 0, y_2)$ est le coût d'extension de x_1 en y_1 et y_2 ;

$d(x_1, y_1; x_2, y_2)$ est le coût de mélange de x_1 et x_2, apparié à y_1 et y_2.

L'algorithme de DTW vise à minimiser la distance cumulée des alignements (*match*) effectués :

$$D(0,0) = 0$$

$$D(i,j) = \min \begin{cases} D(i, j-1) + d(0, t_j; 0, 0) \\ D(i-1, j) + d(s_i, 0; 0, 0) \\ D(i-1, j-1) + d(s_i, t_j; 0, 0) \\ D(i-1, j-2) + d(s_i, t_j; 0, t_{j-1}) \\ D(i-2, j-1) + d(s_i, t_j; s_{i-1}, 0) \\ D(i-2, j-2) + d(s_i, t_j; s_{i-1}, t_{j-1}) \end{cases}$$

5.4.1.2. Alignement hiérarchique

La méthode utilisée peut être appliquée sur les textes dont la structure logique de document est explicitement codée.

Soient S et T les textes source et cible à aligner, avec

$$S = [s_1, s_2, ..., s_n], \qquad T = [t_1, t_2, ..., t_m]$$

où s_i, t_j ($i=1,...,n$; $j=1,...,m$) sont des fragments (division, paragraphe, phrase, *etc*.) de chaque texte, contenant potentiellement eux-mêmes d'autres fragments de niveau hiérarchique inférieur.

Un alignement Align(S, T) peut être décrit comme une suite de paires (σ_j, τ_j), signifiant que σ_j dans le texte S est aligné à τ_j dans le texte T :

$$Align(S,T) = [(\sigma_1, \tau_1), ..., (\sigma_r, \tau_r)]$$

où :

$$\bigcup_{j=1}^{r} \sigma_j = S, \quad \bigcup_{j=1}^{r} \tau_j = T$$

σ_j et τ_j peuvent être nuls, auquel cas on a affaire à un alignement correspondant à une traduction du type 0-n (insertion), ou m-0 (suppression).

La procédure d'alignement est réalisée de façon récursive. Au départ, on aligne les racines des documents par un appariement 1-1. Après *i* étapes d'alignement, on obtient les alignements *i*+1 par le rafinement des alignements obtenus à l'étape *i*. Le processus s'arrête quand on obtient des alignements au niveau de phrases.

De plus, certaines heuristiques ont été introduites afin de pouvoir prendre en compte les incohérences de représentation de la structure du document entre les textes source et cible. En effet, l'algorithme décrit ci-dessus ne fonctionnera pas de manière satisfaisante si par exemple l'un des documents à aligner indique sections et sous-sections par des tags `<div>` imbriqués, tandis que l'autre passe directement du niveau des sections à celui des paragraphes. Dans ce cas, on se trouvera à devoir aligner des sous-sections avec des paragraphes, ce qui n'a aucun sens et aboutira sans doute à des résultats catastrophiques. Une autre difficulté possible est le cas où les fils d'un nœud ne sont pas tous du même type (par exemple un mélange de `<div>` et de `<par>`).

Deux mécanismes permettent de contourner ces deux obstacles :

- le premier cas est détecté en vérifiant à chaque étape de l'itération que le rapport *r* entre nombres d'éléments à aligner est « raisonnable ». *r* doit être au maximum égal au triple du rapport moyen constaté sur l'ensemble du texte à aligner pour le niveau hiérarchique considéré, et au minimum égal au tiers de cette valeur. Sinon, on descend d'un niveau hiérarchique du côté où le nombre d'éléments est trop faible, ignorant de ce fait un niveau structurel.
- Le second cas se détecte très simplement, par simple observation des balises. Pour le résoudre, nous introduisons dans ce cas des nœuds artificiels rassemblant les éléments de bas niveau isolés, puis nous appliquons le test précédent pour assurer la pertinence de ce choix.

L'algorithme d'alignement de chaque étape est celui de Gale et Church présenté précédemment, avec une nouvelle fonction de coût que nous décrivons maintenant.

5.4.1.3. Fonction de coût développée

L'algorithme original proposé par Gale et Church [GAL 93] calcule le coût d'une mise en correspondance de phrases de longueur l_1 en langue L_1 et l_2 en langue L_2 à partir de la probabilité $p(l_1, l_2)$: probabilité qu'une phrase de longueur l_1 (en nombre de caractères) en langue L_1 soit la traduction d'une phrase de longueur l_2 en langue L_2. Afin de calculer cette probabilité, Gale et Church font usage de constantes acquises sur un corpus d'entraînement : le rapport d'échelle entre les longueurs de phrases de même sens dans les deux langues, ainsi que l'écart type de cette valeur.

Ne disposant pas de corpus de référence aligné au vietnamien de taille suffisamment conséquente pour évaluer ces constantes, nous avons été amenée à proposer un autre mode de calcul de la probabilité $p(l_1, l_2)$, qui permet un fonctionnement de l'algorithme indépendant des langues employées. Les valeurs statistiques employées sont cette fois acquise sur le corpus à aligner lui-même ; l'hypothèse de travail retenue est que les deux langues étudiées formuleront typiquement le même nombre de phrases pour exprimer un même enchaînement d'idées, même si des déviations par rapport à cette règle sont ponctuellement observables.

À partir des données textuelles disponibles, on calcule :

$\overline{l_1}$: longueur moyenne des phrases dans le corpus en langue L_1,

σ_{l_1} : écart type des longueurs des phrases dans le corpus en langue L_1.

De même, on définit $\overline{l_2}$ et σ_{l_2} pour le texte cible.

Ces valeurs permettent, lors de la comparaison des longueurs l_1 et l_2 pour le calcul de $p(l_1, l_2)$, d'effectuer une translation et une mise à l'échelle de ces valeurs (centrage et réduction) grâce auxquelles elles peuvent par la suite être étudiées comme des variables aléatoires répondant à la loi normale. On définit l_{1n} et l_{2n} (« n » pour « normal ») par :

$$l_{1n} = \frac{l_1 - \overline{l_1}}{\sigma_{l_1}} \qquad l_{2n} = \frac{l_2 - \overline{l_2}}{\sigma_{l_2}}$$

Le calcul de l'aire située "sous" la loi de distribution statistique normale entre les valeurs l_{1n} et l_{2n} donne alors la probabilité *a priori* de pouvoir trouver une valeur de l_1 plus proche de l_2 (et réciproquement) après centrage et réduction, autrement dit, une "meilleure" traduction du point des vue des longueurs de phrases[69]. Cette aire est définie formellement par :

$$\frac{1}{\sqrt{2\pi}} \int_{l_{1n}}^{l_{2n}} e^{\frac{-t^2}{2}} dt$$

On peut l'exprimer informellement comme "la probabilité de trouver mieux que l_1 pour traduire l_2" (ou réciproquement). C'est cette valeur que nous utilisons pour évaluer le coût $c(l_1, l_2)$ d'une transition alignant des phrases de longueur l_1 et l_2.

Les résultats que cette nouvelle fonction de coût permet d'atteindre sont, sur un bitexte français-anglais de référence, tout à fait équivalents à ceux de la mesure originale de Gale et Church. Comme nous l'avons déjà précisé, ce n'est pas la recherche d'un accroissement de performance qui nous a poussée à la développer, mais la recherche d'une méthode indépendante des langues étudiées. Le maintien d'une qualité identique est donc un résultat satisfaisant.

5.4.2. Évaluation du résultat

Dans cette section, nous évaluons d'abord le résultat de l'application de notre système à un sous-corpus parallèle anglais-français utilisé pour la compagne d'évaluation ARCADE I. Ensuite, nous réalisons une évaluation comparative sur les résultats d'alignement d'un texte trilingue pour 3 couples de langues (anglais – en, français – fr, vietnamien – vn).

5.4.2.1. Corpus de référence

Le tableau Tableau 5-3 montre la taille de notre corpus d'évaluation. Il s'agit du roman *Le petit prince* dans trois langues : français, anglais et vietnamien, et du corpus JOC, qui nous a été fourni pour le test blanc de la campagne ARCADE II, en français et en anglais.

	Corpus JOC		**Le petit prince**		
	fr	en	fr	en	vn
Nombre d'unités lexicales	277 616	238 189	18 286	20 881	18 500
Nombre de phrases	9025	9035	1 674	1 660	1 663

Tableau 5-3 Dimensions du corpus de référence

Les alignements de référence sont, d'une part, fournis par la campagne ARCADE II (corpus JOC), et d'autre part, obtenus par l'alignement automatique avec révision manuelle grâce à notre outil de concordance multilingue (*Le petit prince*).

[69] Rigoureusement, il serait nécessaire de calculer l'aire de la zone située de part et d'autre de l_{2n}, à une distance inférieure à $|l_{1n} - l_{2n}|$, mais cela présente l'inconvénient de rendre la mesure asymétrique – le résultat n'est pas le même si l'on considère que la cible est l_1 ou l_2 –, et ne change pas les ordonnancements qualitatifs impliqués par la valeur calculée sur les paires de longueurs possibles.

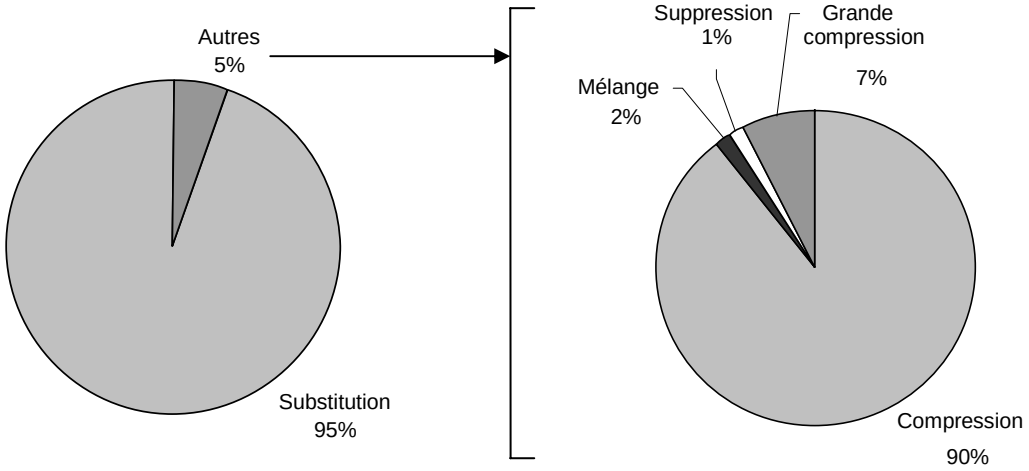


Figure 5-3 Proportion des types d'alignement du corpus JOC fr – en

La Figure 5-3 présente la proportion des types d'alignement du corpus JOC. Environ 95% des alignements sont des substitutions (alignements 1-1) ; des 5% qui restent, la plupart sont des alignements 2-1 ou 1-2 (compression ou extension). Un dixième des alignements non 1-1 concerne les « grande compression », c'est-à-dire les alignements m : 1 ou 1 : m, où $m \geq 3$. Nous avons remarqué que la valeur au pire de m est 5, et que la plupart des cas d'alignements sont du type m : 1 (m phrases françaises alignées à une phrase anglaise). Cela vient en particulier du fait que ces « phrases » anglaises se composent d'une séquence de « phrases » courtes séparées par des « ; ». Le même phénomène est observé dans le texte *Le Petit Prince* en anglais, ce qui explique en partie la proportion des alignements 1-1 est plus élevée pour le couple de langues fr-vn (*cf.* Figure 5-4, Figure 5-5, Figure 5-6). Cela illustre l'impact possible sur l'alignement des choix de segmentation : peut-être serait-il préférable, bien qu'il ne s'agisse pas « officiellement » d'une ponctuation forte, de considérer en anglais le point virgule comme une limite de phrase.

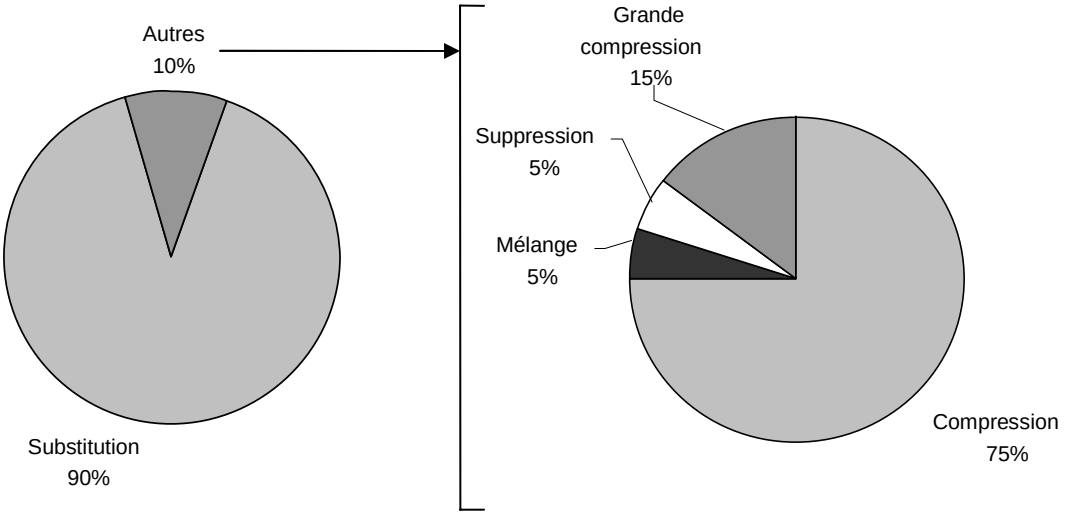


Figure 5-4 Proportion des types d'alignement du texte *Le Petit Prince* français - anglais

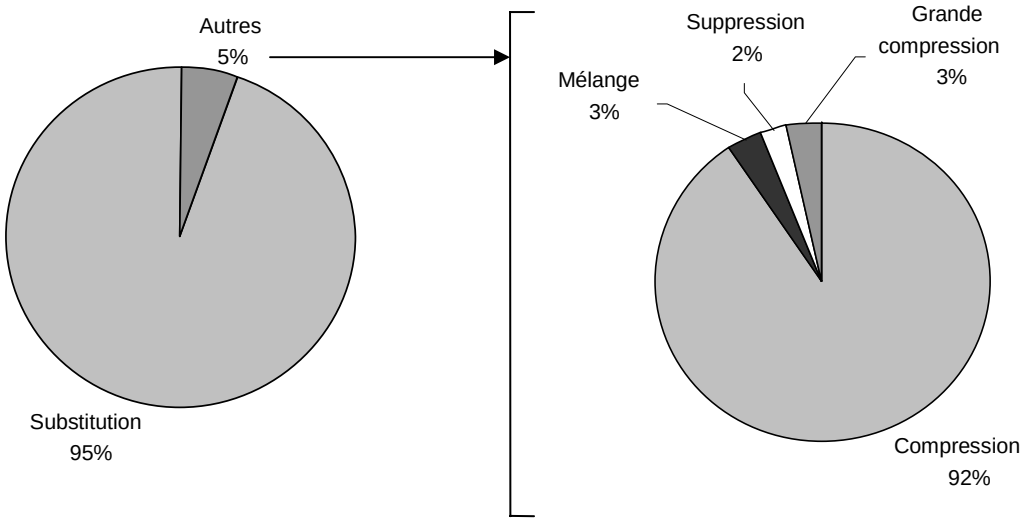


Figure 5-5 Proportion des types d'alignement du texte *Le Petit Prince* français – vietnamien

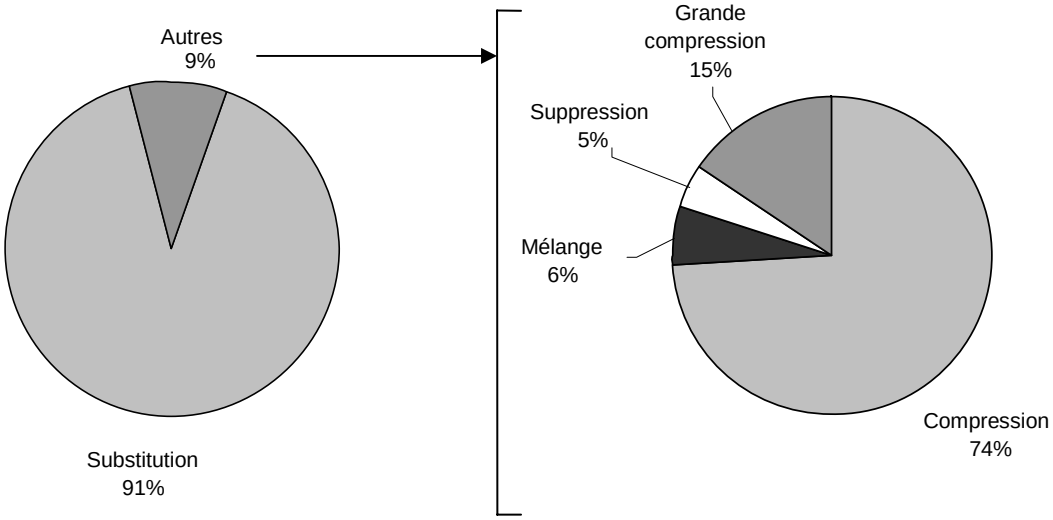


Figure 5-6 Proportion des types d'alignement du texte *Le Petit Prince* anglais – vietnamien

Une autre source de complexité de l'alignement dans le corpus *Le Petit Prince*, pour les couples fr-en et en-vn, est que la représentation des dialogues dans la version anglaise est différente par rapport au français et au vietnamien, ce qui implique une segmentation non cohérente entre ces trois langues. Nous avons également noté que toutes les suppressions recensées apparaissent dans les versions anglaises et vietnamiennes – ce qui est logique puisqu'elles correspondent majoritairement à des défauts de traduction.

Nous passons maintenant aux résultats d'alignement de ces textes, obtenus par notre système d'alignement structurel.

5.4.2.2. Résultat

Pour l'évaluation du résultat d'alignement, nous faisons appel aux taux de précision et de rappel, ainsi qu'à la *F*-mesure, présentés à la section 5.2.2. Les mesures de précision et de rappel sont calculées avec les longueurs définies en nombre de caractères.

Le Tableau 5-4 présente les valeurs de ces trois indices pour l'alignement structurel appliqué sur le corpus JOC dans les deux langues française et anglaise, ainsi que le corpus *Le Petit Prince* dans les trois langues fr, en et vn.

	Corpus JOC	Le Petit Prince		
	fr – en	fr – en	fr – vn	en – vn
Précision	99,09%	96,02%	90,46%	81,42%
Rappel	97,52%	90,96%	87,73%	76,21%
F-mesure	98,30%	93,42%	89,08%	78,73%

Tableau 5-4 Évaluation du résultat de l'alignement structurel

Comme nous pouvons le constater, l'alignement fr-en est en général d'assez bonne qualité, en particulier pour le texte institutionnel JOC. Pour le texte trilingue *Le Petit Prince*, le résultat se dégrade progressivement en passant du couple fr–en au couple fr–vn, puis en–vn. Il peut paraître surprenant au premier abord que l'alignement fr-vn soit de qualité si moyenne alors que, d'après la Figure 5-5, il s'agit du bitexte pour lequel la plus grande partie des alignements sont de « simples » substitutions (1 : 1). L'explication de ce phénomène est donné par la Figure 5-7 (page suivante), qui présente la répartition des rapports de longueur entre phrases alignées pour les trois couples de langue fr-en, fr-vn et en-vn. Le couple fr-en est celui pour lequel les rapports sont les plus cohérents, montrant un « pic » marqué vers la valeur 0,95 ; ensuite vient le couple fr-vn, pour lequel la variance de ce rapport est plus importante, et enfin en-vn, pour lequel les valeurs sont très étalées. Cette caractéristique est naturellement un obstacle au fonctionnement de notre méthode, qui se fonde sur l'hypothèse simplificatrice que ce rapport est sensiblement constant. Le Tableau 5-5 donne un aperçu synthétique de cette différence en présentant les moyennes et écarts types des rapports observés.

On constate à l'étude des textes parallélisés qu'il arrive que certaines phrases soient « résumées » dans la traduction vietnamienne par rapport à leur contenu original, ce qui explique la plus grande dispersion des rapports de longueur entre phrases alignées dans les couples fr-vn et en-vn. L'alignement en-vn est naturellement le plus « irrégulier » (et aussi celui pour lequel les résultats sont le moins bon), puisque il y a entre eux la distance de deux traductions, le texte original étant en français.

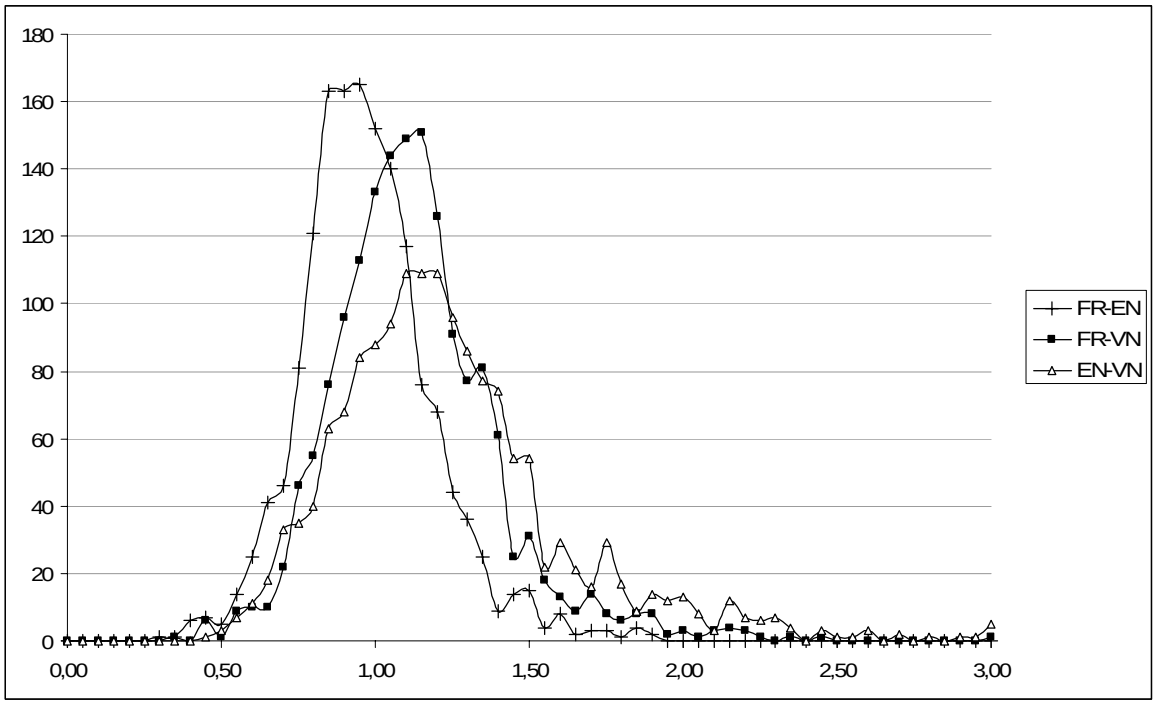


Figure 5-7 Densités de répartition des rapports entre longueurs de phrases alignées dans *Le Petit Prince*

	fr-en	fr-vn	en-vn
Moyenne	0,95	1,10	1,22
Ecart type	0,22	0,27	0,39

Tableau 5-5 Moyenne et écart type des rapports entre longueurs de phrases alignées dans *Le Petit Prince*

5.5. Alignement lexical

Dans cette section, nous présentons notre expérience sur l'alignement des mots en utilisant la méthode DK-vec. Cette méthode ne permet pas un alignement exhaustif de tous les mots mais seulement de ceux comptant plusieurs occurrences. Notre but est de repérer les bons candidats de traduction mutuelle, et à partir de ces points d'ancrage, d'améliorer l'alignement au niveau des divisions, des paragraphes et des phrases. Ces points d'ancrage peuvent être également les points de départ pour un alignement lexical détaillé. Nous faisons une évaluation des résultats obtenus en appliquant la méthode sur les textes multilingues français-anglais, français-vietnamien et anglais-vietnamien, dont chaque texte monolingue a reçu un prétraitement lexical.

5.5.1. Méthode mise en œuvre

Pour l'alignement des mots dans un corpus bilingue, nous empruntons la méthode d'alignement basée sur l'algorithme DK-vec (*cf*. Fung [FUN 97]), qui constitue un bon point de départ pour aligner, au niveau des mots, des paires de textes de deux langues distantes (Choueka et *al.* [CHO 00]). Il s'agit d'un algorithme statistique simple, qui tente de mettre en correspondance les mots ayant une distribution similaire dans les deux textes préalablement traités. Le prétraitement de chaque texte consiste à les segmenter en mots, et à étiqueter ceux-ci avec des informations lexicales comme le lemme et la catégorie grammaticale. Par exemple, pour les langues flexionnelles et/ou agglutinantes, les mots peuvent être lemmatisés avant d'être alignés. Pour les langues isolantes comme le vietnamien, un étiquetage morpho-syntaxique est nécessaire.

Nous présentons d'abord l'algorithme DK-vec, puis le système que nous avons mis en œuvre.

5.5.1.1. Algorithme DK-vec

L'algorithme DK-vec fonctionne comme suit :

A chaque mot *w* d'un texte est associé un vecteur représentant les distances relatives entre ses occurrences : $D^w = <d_1^w, ..., d_n^w>$, où *n* est le nombre d'occurrences du mot *w* dans le texte, d_i^w est la distance (en nombre de mots) entre la *i*ème occurrence et la *i*-1ème occurrence de *w* dans le texte.

Ces vecteurs de distance permettent d'évaluer la similarité de distribution d'une paire de mots quelconques dans les deux textes à aligner. L'algorithme DK-vec suppose que les textes considérés ont des tailles similaires en nombre de mots (unités lexicales). Cependant, ce nombre peut être variable en fonction des langues. Donc, avant d'appliquer l'algorithme, on doit disposer d'un coefficient de proportion de langues LPC (*Language Proportion Coefficient*). Ce coefficient peut être calculé grâce aux statistiques des longueurs de textes d'un corpus multilingue de grande taille (déjà segmenté)[70]. Les vecteurs de distance sont normalisés selon la valeur du LPC.

Pour déterminer la relation de traduction entre un mot du texte source et un mot du texte cible, on utilise l'algorithme de programmation dynamique pour calculer la similarité entre deux vecteurs de distance de ces deux mots. Pour éviter une explosion de la complexité des calculs, inévitable en prenant en compte toutes les paires de mots, on peut effectuer un prétraitement pour éliminer les paires de mots dont les vecteurs de distance sont clairement différents. Les critères de sélection d'une paire de mots - *s* dans le texte source et *t* dans le texte cible sont les suivants :

[70] Dans notre application, le LPC est calculé simplement par le rapport de tailles des textes étudiés.

- Les fréquences d'occurrences de *s* et de *t* sont supérieures à 2 (pour des fréquences inférieures ou égales à 2, le résultat n'est pas fiable), et le rapport entre ces deux fréquences ne doit pas dépasser un seuil donné, habituellement 2 (Choueka et *al.* [CHO 00]) – c'est cette valeur que nous avons retenue.
- L'indice de dissimilarité entre leurs deux vecteurs de distribution, mesuré par la formule suivante, ne dépasse pas le seuil de 200 (Fung [FUN 97], Choueka et *al*. [CHO 00]):

$$\varepsilon(s,t) = \sqrt{(m_{Ds} - m_{Dt})^2 + (\sigma_{Ds} - \sigma_{Dt})^2}$$

où *Ds* et *Dt* désignent les vecteurs de distribution respectifs de *s* et *t*, et *m* et σ désignent respectivement la moyenne et l'écart type.

La sélection des paires de mots « intéressantes » ayant été réalisée avec les critères ci-dessus, on peut appliquer l'algorithme DTW pour identifier les paires de vecteurs de distance ayant la similarité la plus grande, en utilisant la fonction de coût d'appariement *C* calculée comme suit (Choueka et *al.* [CHO 00]) :

$$C(0,0) = 0, \quad d_0^s = 0, \quad d_0^t = 0$$

$$C(i,j) = \mid d_i^s - d_j^t \mid + \min \begin{cases} (i)\ C(i-1, j-1) \\ (ii)\ C(i-1, j) \\ (iii)\ C(i, j-1) \end{cases} \qquad (i + j > 0)$$

Les paires de mots correspondant aux vecteurs ayant le coût d'appariement le plus faible sont des candidats à l'alignement.

5.5.1.2. Système mis en œuvre

Voici les caractéristiques de notre système d'alignement :

- Entrées : deux textes ayant subi un prétraitement lexical, codés au format XML. Les informations de lemme et catégorie de chaque unité lexicale dans un texte se trouvent dans les attributs des balises <w> (*cf*. 5.3.2)
- Sortie : un fichier de liens entre des unités lexicales, également codés au format XML (*cf*. 5.3.2).

L'algorithme implémenté est identique à celui présenté, suivi d'une étape de filtrage consistant à ne retenir pour chaque mot étudié que la paire correspondant à l'appariement ayant obtenu le coût minimal. Les paires ayant un coût supérieur à un seuil fixé à 100 sont automatiquement écartées. Les paires retenues sont ensuite classées par ordre de coût croissant.

Les textes français et anglais sont étiquetés et lemmatisés avec le logiciel d'étiquetage TreeTagger librement distribué sur Internet[71] pour les langues français, anglais, allemand, italien, et espagnol. Pour l'instant nous nous en tenons au résultat brut de l'étiqueteur et ignorons les erreurs potentielles de l'étiquetage automatique.

Les textes vietnamiens considérés sont étiquetés et vérifiés manuellement dans le cadre de notre projet de constitution d'un corpus étiqueté de référence (*cf*. 3.4.2).

La section suivante présente le résultat obtenu du système sur les langues française, anglaise et vietnamienne.

[71] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html

5.5.2. Évaluation du résultat

Dans le temps imparti à cette recherche, nous n'avons pu évaluer notre système que sur un seul corpus *Le Petit Prince*, dont les caractéristiques sont présentés au paragraphe 5.4.2.1. L'évaluation du résultat d'alignement des trois couples de langues (français-anglais, français-vietnamien et anglais-vietnamien) nous permet d'avoir un premier aperçu de la perspective de la technique utilisée.

Les trois figures ci-dessous présentent, d'une part, la proportion des mots pertinemment alignés (en ordonnée) parmi les *n* (en abscisse) premières paires de la liste proposée par le système (courbe pleine), d'autre part la proportion de mots pertinemment alignés parmi les 20 paires autour de la position *n* dans la liste (en pointillés).

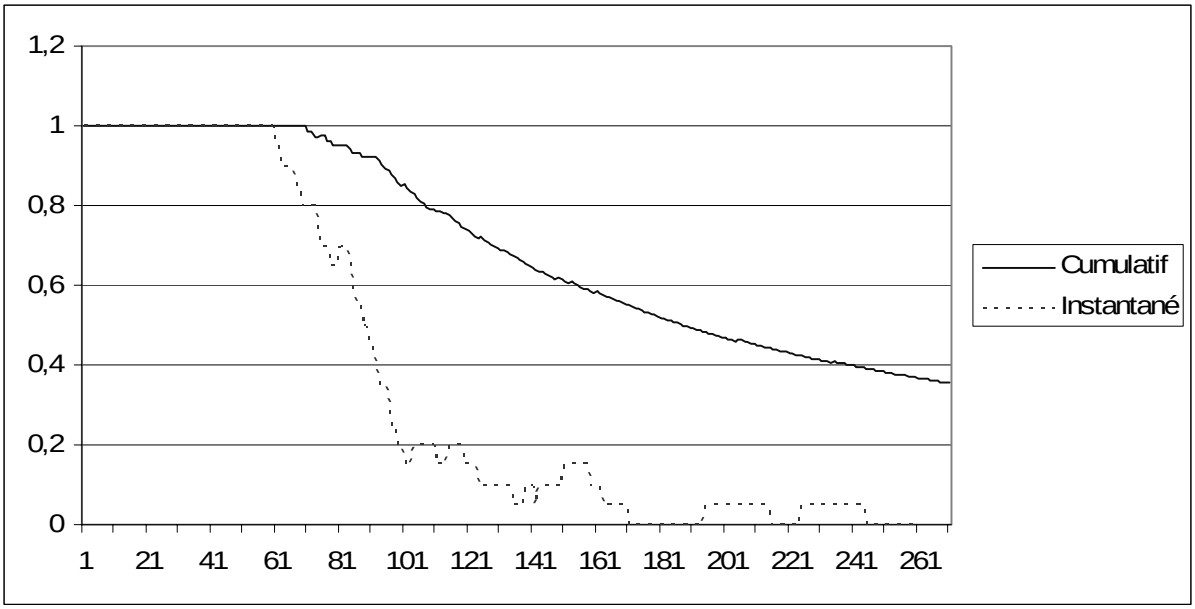


Figure 5-8 Qualité de l'alignement lexical fr–en

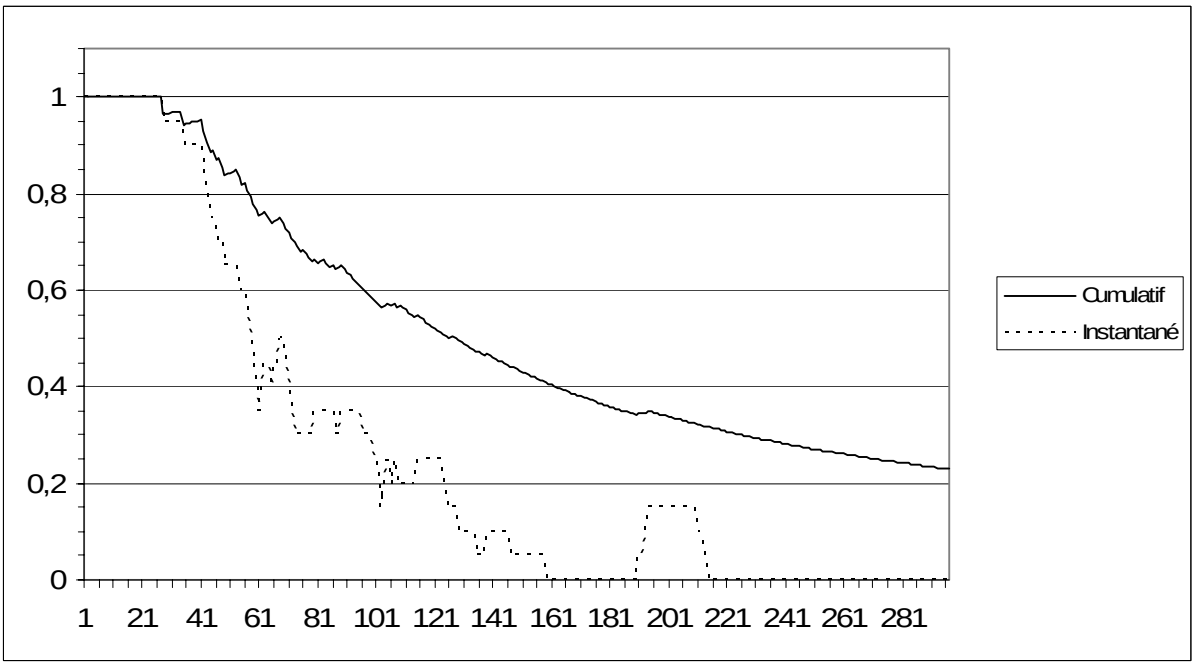


Figure 5-9 Qualité de l'alignement lexical fr-vn

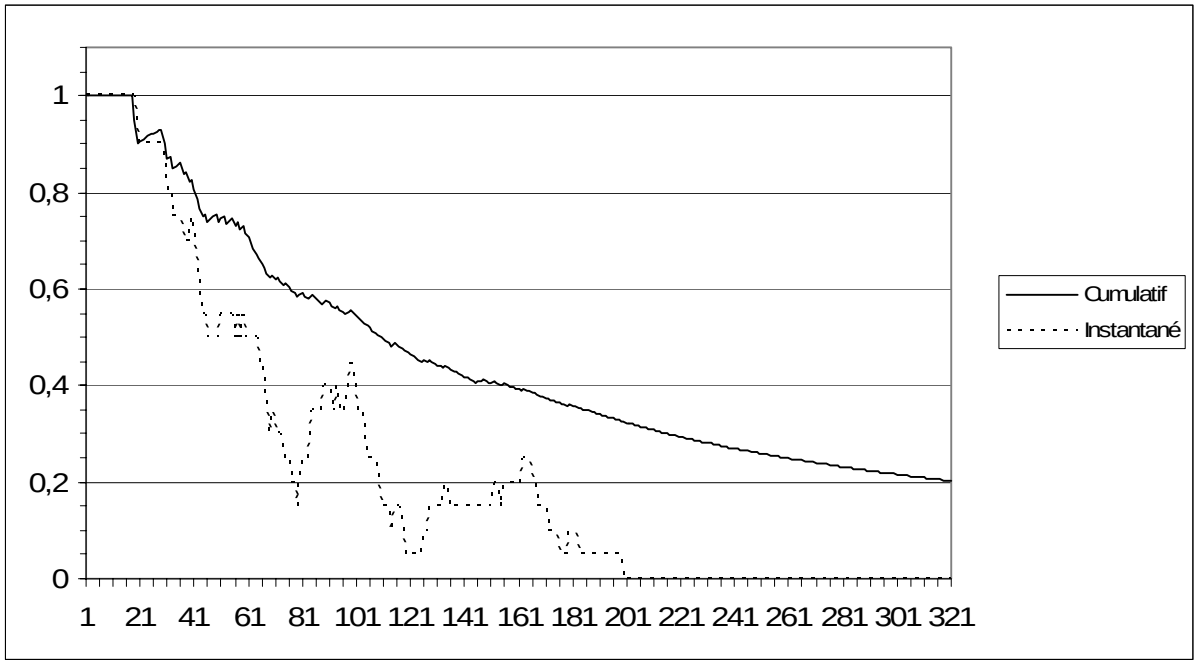


Figure 5-10 Qualité de l'alignement lexical en-vn

La première constatation remarquable est que la fonction de coût employée permet en effet d'ordonner les paires de mots candidates à l'alignement de manière pertinente : les trois graphiques présentent en effet tous une première phase constituée de paires ayant un faible coût d'alignement et toutes, ou presque, « bonnes », suivie d'un chute brutale de la précision pour aboutir à une seconde phase où tous les appariements, ou presque, sont erronés. La longueur de la première phase « fiable », donnant la quantité d'appariements réellement exploitables, ainsi que la brutalité de la transition, sont les principaux signes de la qualité de l'alignement.

La qualité de l'alignement lexical reflète assez directement les résultats obtenus pour l'alignement structurel : en tête, la paire français-anglais, suivi de français-vietnamien et enfin anglais-vietnamien. Outre le fait que les difficultés s'opposant à l'alignement structurel pèsent également ici (en « bruitant » les vecteurs de distribution des mots), les résultats médiocres de l'appariement avec le vietnamien peuvent s'expliquer en particulier par la grande différence pouvant exister en vietnamien entre catégorie morphosyntaxique et rôle des mots : lorsque le français comprend deux mots distincts pour un verbe et un nom (« développer » - « développement »), le vietnamien construit une dérivation du verbe (« développer » - « fait de développer ») ; le phénomène réciproque est également possible (« transformer industrie » pour « industrialiser »). Dans ces cas, les occurrences des mots, tels que leur lemme et leur catégorie permettent de les identifier, sont naturellement très différentes en français/anglais et en vietnamien.

La meilleure solution consisterait naturellement à amender la segmentation du corpus afin de considérer les formes dérivées comme des entités lexicales à part entière. Il est également possible de permettre l'alignement de groupes de mots, détectés par exemple grâce à une méthode fondée sur la recherche de collocations.

5.6. Combinaison des approches structurelle et lexicale

Notre principale contribution personnelle au chapitre de l'alignement multilingue, outre la définition d'une fonction de coût originale pour l'alignement structurel, indépendante des langues considérées, est une proposition de combinaison des deux méthodes classiques présentées aux sections précédentes, d'alignement structurel et lexical.

Le principe de l'évolution que nous introduisons est d'établir une boucle de rétroaction entre ces deux méthodes : l'alignement structurel permet de définir une distorsion temporelle des positions d'occurrences de mots, et ainsi d'augmenter la précision de l'alignement de ces occurrences. Réciproquement, l'alignement lexical permet de définir, sinon des points d'ancrages pour l'alignement (solution qui nous semblait hasardeuse en présence d'erreurs toujours possibles dans l'alignement lexical), du moins des "attirances" entre segments de texte, ou "présomptions d'alignement" permettant de guider le processus de mise en correspondance des segments, en modifiant le calcul de la fonction de coût d'alignement.

Nous précisons dans les deux sous-sections suivantes les mécanismes employés pour prendre en compte au cours de l'alignement lexical les résultats d'un alignement structurel, puis réciproquement, avant de détailler à la section 5.6.3 la structure du programme développé.

5.6.1. Utilisation des résultats d'un alignement structurel pour enrichir l'alignement lexical

Lors de l'alignement lexical, il est fait usage des « coordonnées » des occurrences des mots, comptées en nombre de positions depuis le début du texte à aligner. La technique présentée ici consiste à modifier ces coordonnées afin de prendre en compte l'information apportée par l'alignement structurel, avant d'appliquer l'algorithme d'alignement lexical sans modification majeure.

Nous considérons l'alignement de segments de texte correspondant à un palier textuel quelconque (chapitre, paragraphe, phrase, *etc.*), que nous notons $S_{s1\ldots m}$ pour le texte source et $S_{t1\ldots n}$ pour le texte cible. L'alignement est représenté sous forme de paires mettant en correspondance un ensemble de segments source avec un ensemble de segments cible. Par exemple, ($\{S_{s4}\}$, $\{S_{t5}\}$) est un alignement simple « un vers un », ($\{S_{s10}, S_{s11}\}$, $\{S_{t13}\}$) un alignement « deux vers un », et ($\emptyset$, $\{S_{t8}\}$) représente une insertion. On note $P_{1\ldots p}$ l'ensemble des paires permettant de décrire l'alignement de tous les segments des textes source et cible, la numérotation des P_i suivant l'ordre d'apparition des segments dans les textes.

La coordonnée transformée d'une occurrence de mot est une valeur réelle calculée en deux temps :

- sa partie entière E est le numéro de la paire d'alignement contenant le segment de texte où l'occurrence apparaît
- sa partie fractionnelle reflète la position relative de l'occurrence dans l'ensemble des segments rassemblés par la paire d'alignement P_E.

La Figure 5-11 illustre à l'aide d'un exemple concret le résultat de la « distorsion temporelle » ainsi appliquée aux positions d'occurrences. Son effet est d'une part de rapprocher numériquement les coordonnées d'occurrences de mots apparaissant dans des segments alignés tout en tenant compte des phénomènes d'insertion, délétion, *etc.*, d'autre part de permettre une représentation « continue » des positions d'occurrences intégrant de manière transparente un facteur de proportionnalité entre longueurs d'énoncés dans les deux langues considérées.

Longueur des segments du texte source, en nombre de mots :

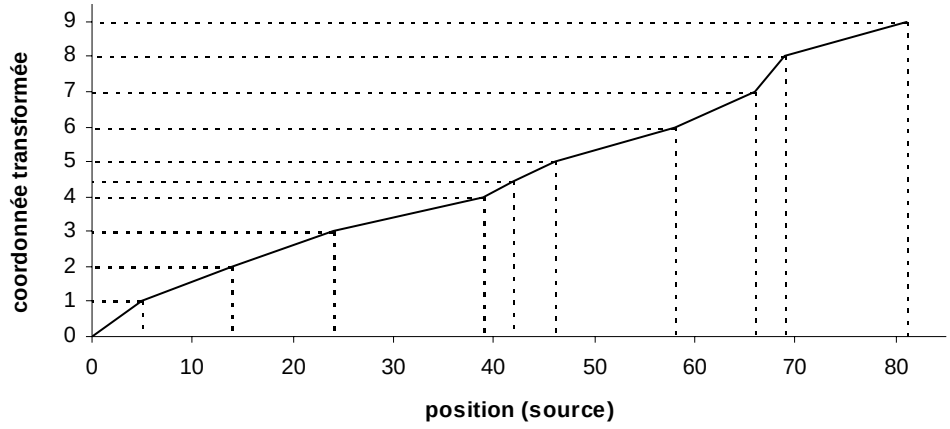
S_{s1}	S_{s2}	S_{s3}	S_{s4}	S_{s5}	S_{s6}	S_{s7}	S_{s8}	S_{s9}	S_{s10}
5	9	10	15	3	4	12	8	3	12

Longueur des segments du texte cible, en nombre de mots :

S_{t1}	S_{t2}	S_{t3}	S_{t4}	S_{t5}	S_{t6}	S_{t7}	S_{t8}	S_{t9}
4	11	13	8	8	12	7	5	15

Alignement structurel :

N° alignement	1	2	3	4	5	6	7	8	9
N° segment source	1	2	3	4	5, 6	7	8	9	10
N° segment cible	1	2	3	4	5	6	7	Ø	8, 9



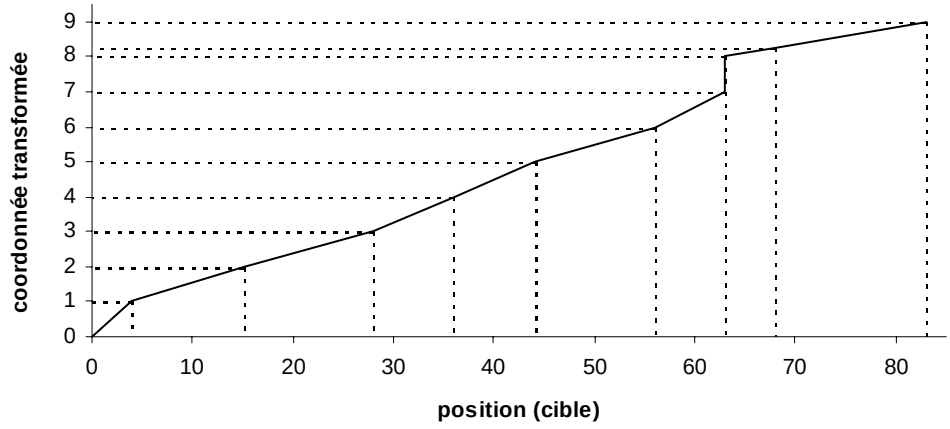


Figure 5-11 Exemple de résultat de transformation des coordonnées de positions d'occurrences de mots

(L'exemple concerne un alignement factice purement illustratif. Les deux courbes donnent l'aspect de la fonction de conversion pour les mots des textes source et cible, respectivement ; les lignes pointillées indiquent les limites des segments de texte alignés.)

5.6.2. Utilisation des résultats d'un alignement lexical pour enrichir l'alignement structurel

La méthode employée pour tenir compte, au cours d'un alignement de segments de textes, de la connaissance d'un certain nombre d'alignements entre occurrences de mots consiste à moduler la fonction de coût employée par l'algorithme DTW par un facteur dépendant de la proportion des occurrences de mots des segments de phrases considérés qui sont alignées.

Pour un palier textuel quelconque (chapitre, paragraphe, phrase, *etc.*), nous notons une fois encore $S_{s1\ldots m}$ les segments du texte source et $S_{t1\ldots n}$ ceux du texte cible. On définit pour toute paire (S_{si}, S_{tj}) une mesure d' « attirance lexicale » *a* par :

$$a(S_{si}, S_{tj}) = \frac{2\ nbOccurrencesAlignées(S_{si}, S_{tj})}{nbOccurrences(S_{si}) + nbOccurrences(S_{tj})}$$

où $nbOccurrencesAlignées(S_{si}, S_{tj})$ désigne le nombre d'occurrences de mots des phrases S_{si} et S_{tj} alignées par la donnée d'alignement lexical connue. L'alignement des segments de texte est ensuite réalisé par l'algorithme classique, la fonction de coût *c* étant remplacée par une variante *c'* qui, contrairement à *c*, ne dépend plus uniquement des longueurs des segments mais également de leur contenu :

$$c'(S_{si}, S_{tj}) = \frac{a_{\max}\ c(l_{si}, l_{tj})}{1 + K\ a(S_{si}, S_{tj})}$$

où $a_{\max}$ désigne la valeur maximale constatée de la mesure *a* sur toutes les paires de segments du bitexte à aligner, et K est une constante permettant de régler l'importance donnée à la prise en compte du critère d'alignement d'occurrences de mots au cours de l'alignement de segments de texte. Typiquement, la valeur K=3 (le coût brut est au maximum divisé par 4) permet d'obtenir des résultats optimaux sur l'ensemble des tests que nous avons réalisés. Pour calculer un coût d'alignement entre ensembles de segments, la valeur de *a* est calculée sur les segments résultant de la fusion des segments des ensembles comparés.

5.6.3. Mise en œuvre de la boucle de rétroaction entre alignements structurel et lexical

Si nous nommons « profondeur » d'un palier de segmentation textuelle son degré d'imbrication dans la structure du document (typiquement, les segments de profondeur 1 sont les chapitres, de profondeur 2, les paragraphes, et de profondeur 3, les phrases), le principe de l'algorithme mis au point consiste à considérer qu'un alignement lexical réalisé en prenant en compte un alignement structurel de profondeur *n* fournira des informations permettant de réaliser avec succès un alignement structurel de profondeur *n+1*. Partant d'un alignement lexical réalisé sans pré-alignement structurel, on peut ainsi de manière récursive réaliser un alignement structurel à la granularité maximale proposée par la représentation des documents considérés.

L'algorithme développé est donc le suivant :

```
AL = alignement lexical (bitexte brut)
Pour profondeur de 1 à max_profondeur répéter
   répéter 2 fois
        AS = alignement structurel (bitexte, profondeur, AL)
        AL = alignement lexical (bitexte, AS)
Résultat : AL, AS
```

Pour chaque palier textuel, la séquence alignement lexical / alignement structurel est répétée 2 fois, afin de permettre une stabilisation des résultats.

5.6.4. Évaluation du résultat

Afin d'évaluer l'apport de la méthode proposée par rapport à l'alignement structurel classique par la méthode de Church et Gale, nous avons mis au point une procédure de test consistant à dégrader de manière automatique le corpus JOC anglais en supprimant aléatoirement un pourcentage variable de ses phrases, puis à observer l'évolution de la f-mesure de l'alignement à mesure que ce pourcentage croît. La Figure 5-12 présente conjointement les résultats obtenus par le système d'alignement structurel « simple » et par le système hybride (la valeur reportée est la médiane des résultats obtenus sur cinq exécutions).

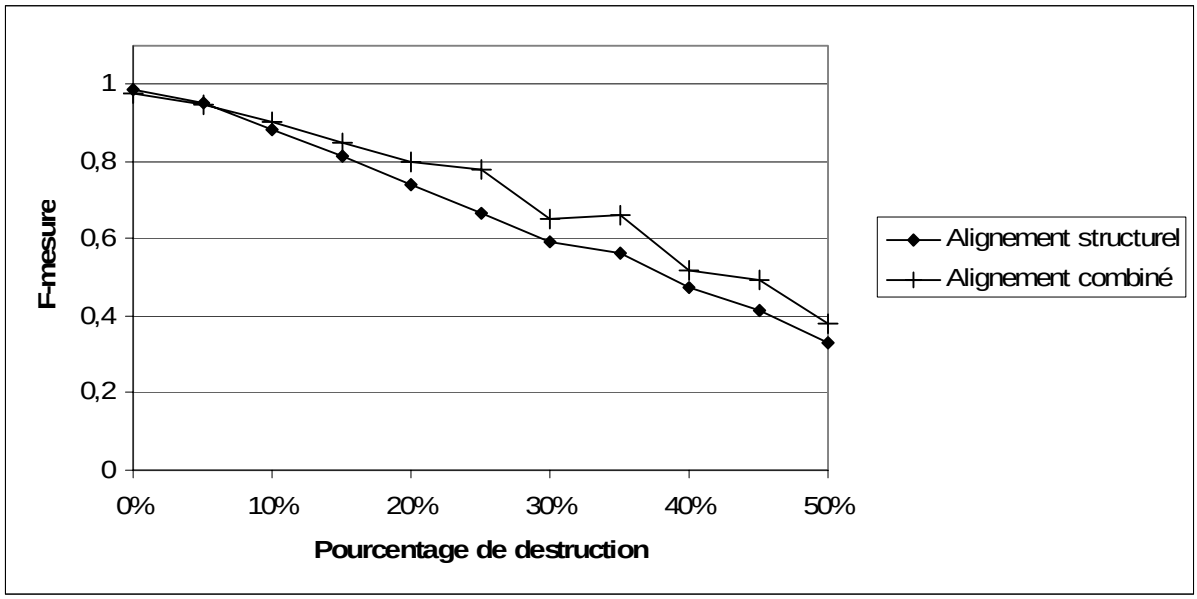


Figure 5-12 Résultats comparatifs de l'alignement structurel et combiné (*F*-mesure, en caractères)

Lorsque les deux textes présentent un parallélisme quasi parfait (comme c'est le cas du corpus JOC initial), la méthode structurelle pure obtient des résultats supérieurs à la méthode hybride. Cela peut se comprendre en considérant qu'en combinant deux techniques imparfaites, on combine généralement plus leurs défauts que leurs qualités. En revanche, il est intéressant de constater que, lorsque la dégradation atteint 10% du texte cible, l'« ancrage lexical » réalisé par le système hybride semble permettre de contrebalancer partiellement la perte de parallélisme structurel.

5.7. Participation à la campagne ARCADE II

La campagne d'évaluation ARCADE II a permis le développement de nouveaux corpus de référence pour l'évaluation de l'alignement :

- d'une part, le corpus JOC (Journal Officiel de la Communauté européenne) a été muni dans son intégralité d'un alignement de référence au français pour l'anglais, l'allemand, l'espagnol et l'italien ;
- d'autre part, le corpus MD a été nouvellement créé : il est constitué d'extraits d'archives du mensuel *Le Monde diplomatique* en français, arabe, persan, russe, grec, chinois et japonais. Tous les textes sont alignés au français au niveau des phrases, mais le contenu varie d'une langue à l'autre suivant les textes qu'il a été possible de collecter. La composition du corpus est détaillée par le Tableau 5-6 suivant.

	Nombre de documents	Nombre de mots ($x10^3$)		Nombre de phrases ($x10^3$)		Nombre d'alignements ($x10^3$)
français-arabe	150	517	403	14	11	11
français-persan	53	214	220	5.2	5.3	4.61
français-russe	50	173	158	4.2	4.2	4
français-grec	50	179	190	4.3	4.4	4.37
français-chinois	59	197	-	5.2	5.5	4.45
français-japonais	52	240	-	5.7	6.1	5.51

Tableau 5-6 : Composition du corpus MD de la campagne ARCADE II

Pour chacun des deux corpus décrits, une moitié des textes est fournie avec l'information de découpage en phrases (corpus « segmenté »), et l'autre moitié sans (corpus « brut » : l'information de découpage s'arrête au niveau des paragraphes). Nous avons participé à l'évaluation sur les corpus JOC brut et segmenté, en réalisant pour le corpus brut un découpage automatique simple afin de le renre compatible avec notre système, et sur le corpus MD segmenté seulement, faute de connaissances linguistiques suffisantes pour définir une heuristique de découpage automatique en phrases de textes en arabe ou chinois.

Le Tableau 5-7 ci-dessous présente les résultats obtenus sur le corpus JOC pour les quatre paires de langues considérées. Le programme employé est le système « hybride » décrit à la section précédente, et les textes ont été lemmatisé de manière totalement automatique, sans validation manuelle, en utilisant le logiciel TreeTagger avec ses paramètres par défaut. Nous donnons pour chaque test la précision, le rappel, la f-mesure, ainsi que le rang de notre système parmi les trois évalués.

	Brut				Segmenté			
	précision	rappel	f-mesure	rang	précision	rappel	f-mesure	rang
Anglais	0.9554	0.9818	0.9684	2	0.9733	0.9882	0.9806	1
Allemand	0.9504	0.9815	0.9657	1	0.9726	0.9867	0.9796	2
Espagnol	0.9598	0.9864	0.9729	1	0.9767	0.9881	0.9823	3
Italien	0.9629	0.9847	0.9732	2	0.9692	0.9825	0.9758	3
Global	0.9569	0.9837	0.9701	1	0.9729	0.9863	0.9795	3

Tableau 5-7 Résultat de l'évaluation de notre système par la campagne ARCADE II pour le corpus JOC

Un point particulièrement positif mis au jour par cette première évaluation est le bon comportement de notre système, en particulier relativement aux autres, dans le cas du corpus brut. Comme nous l'avons vu, ce cas est celui où, le découpage du texte étant totalement automatique, il existe le plus de « bruit » dans l'alignement. Cette constatation rejoint donc la conclusion à laquelle nous amenait notre expérience précédente d'alignement d'un texte de plus en plus dégradé.

Le Tableau 5-8 présente les résultats pour le corpus MD segmenté. Le système employé est cette fois le système d'alignement strucurel simple, étant donnée le manque d'outils permettant d'effectuer simplement la lemmatisation pour les langues considérées.

Deux chiffres sont donnés : le premier est le résultat de l'évaluation officielle, et présente des résultats très faibles pour le japonais, le grec et le persan. La cause de cette mauvaise qualité est une absence de correspondance entre les découpages en paragraphes avec la version française (la version japonaise, par exemple, indique un unique grand paragraphe par article), ajoutée au fait que l'implémentation du système que nous avons utilisée n'intégrait pas le mécanisme de détection des incohérences de structure décrit à la section 5.4.1.2. C'est pourquoi nous avons également indiqué les résultats d'une nouvelle évaluation effectuée depuis avec une version du système corrigée dans ce sens. Les chiffres indiqués en gras sont ceux qui ont évolué par rapport à la première évaluation. Il convient de noter, à propos des rangs indiqués, que seuls deux systèmes ont participé à l'évaluation sur ce corpus.

Les résultats obtenus sont encourageants à plus d'un titre : d'une part, ils montrent que notre système reste fonctionnel, avec un assez grand succès, pour les langues « exotiques » comme le chinois ou l'arabe, qui non seulement emploient des caractères non latins, mais peuvent également avoir sur certains points une « philosophie » radicalement differente de la représentation textuelle d'un texte (comme par exemple l'absence de marquage des séparations de mots en chinois). Il semble donc que nos efforts dans le sens d'une réelle indépendance vis-à-vis des languages considérés (notamment dans la définition de la fonction de coût d'alignement) montrent leur utilité.

	Résultat officiel				Système corrigé			
	précision	rappel	f-mesure	rang	précision	rappel	f-mesure	rang
Arabe	0.8847	0.9323	0.9079	1	0.8847	0.9323	0.9079	1
Persan	0.5920	0.6913	0.6378	2	**0.7475**	**0.8604**	**0.8000**	**2**
Russe	0.8080	0.8310	0.8194	2	0.8080	0.8310	0.8194	2
Grec	0.4561	0.4776	0.4666	2	**0.9554**	**0.9640**	**0.9597**	**2**
Chinois	0.7767	0.8635	0.8178	1	0.7767	0.8635	0.8178	1
Japonais	0.0323	0.1225	0.0511	2	**0.8713**	**0.9106**	**0.8905**	**1**
Global	0.7018	0.7910	0.7437	2	0.8737	0.8839	0.8788	1

Tableau 5-8 Résultat de l'évaluation de notre système pour le corpus MD segmenté

D'autre part, ils confirment la robustesse de notre approche pour l'alignement de textes présentant des différences de contenus assez importante. En effet, le corpus MD a également la caractéristique de présenter des différences beaucoup plus importantes entre textes que le corpus JOC : alors que ce dernier est traduit de la manière la plus littérale possible, la traduction des articles du *Monde diplomatique* suit plus l'esprit des textes que la lettre, et il peut ainsi arriver que certains paragraphes du texte d'origine soient supprimés dans le texte traduit et remplacés par une simple phrase de transition. Afin d'illustrer l'ampleur de ce phénomène, nous présentons sur les trois figures suivantes des diagrammes représentants les proportions de types d'alignement pour le corpus présentant les plus grandes disparités (chinois) et pour le corpus le plus proche de l'original (grec), ainsi qu'un diagramme de moyenne générale pour toutes les langues ; on pourra par exemple comparer les valeurs portées avec celles données pour les corpus JOC et *Petit prince* à la section 5.4.2.1.

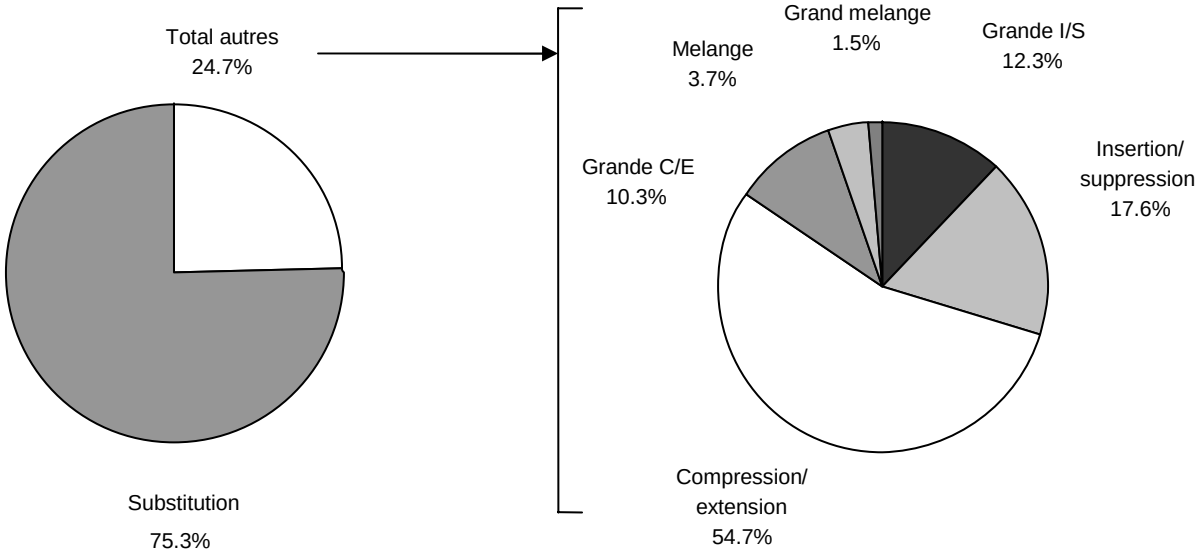


Figure 5-13 Proportions des types d'alignements rencontrés sur l'intégralité du corpus MD

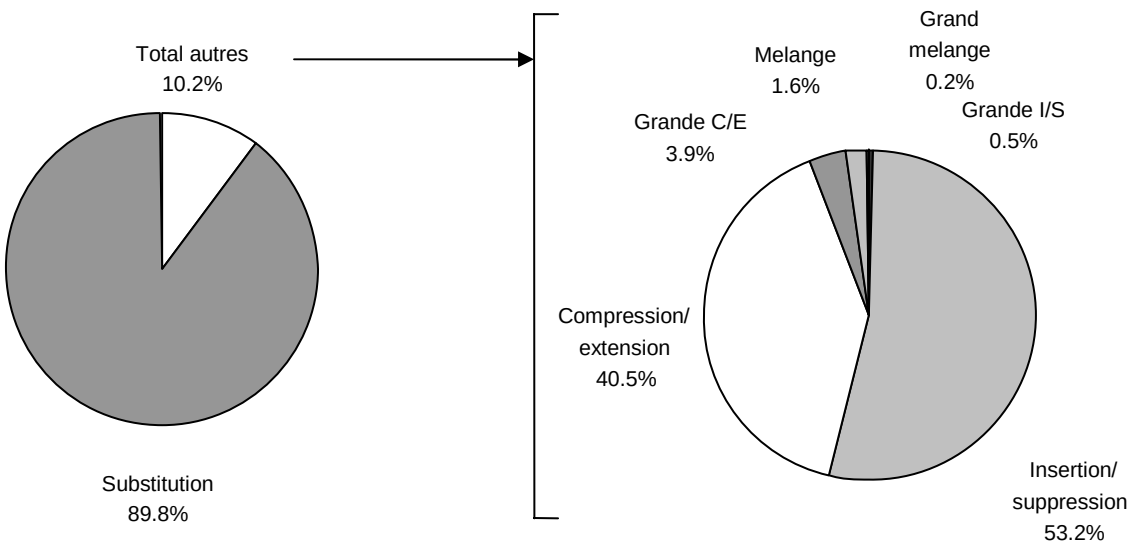


Figure 5-14 Proportions des types d'alignements rencontrés sur la version grecque du corpus MD

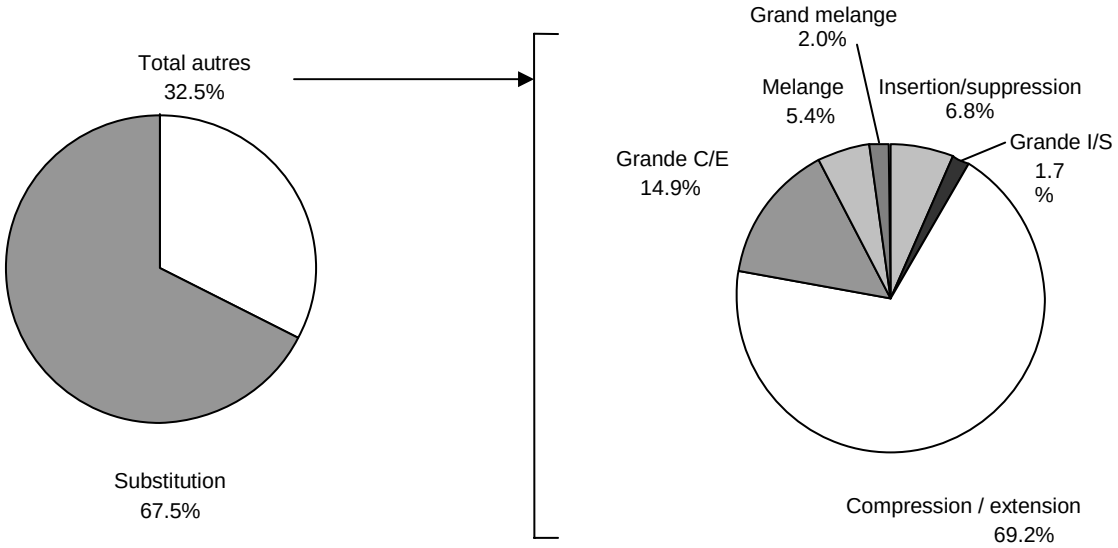


Figure 5-15 Proportions des types d'alignements rencontrés sur la version chinoise du corpus MD

5.8. Bilan et perspectives

Pour la tâche d'alignement multilingue, nous avons développé et évalué un système d'alignement hiérarchique applicable avec un assez grand succès non seulement sur les couples de langues proches comme l'anglais et le français, mais aussi sur les couples de langues distantes comme le français et le vietnamien. Notre aligneur itératif permet d'optimiser le résultat à chaque niveau d'alignement : division, paragraphe, et segment (phrase). Nous avons également étudié les possibilités de combinaison de cet algorithme avec une approche complémentaire réalisant l'alignement d'unités lexicales, selon une méthode originale qui permet d'obtenir des résultats intéressants sur des textes au parallélisme fortement dégradé.

Notre participation à la campagne d'évaluation ARCADE II nous a permis de confirmer ce bon fonctionnement du système pour un nombre important de langues européennes ou non, utilisant ou non l'alphabet latin, ce qui tend à prouver que nos efforts pour la mise en œuvre d'une méthode réellement indépendante des langues traitées ont été relativement fructueux.

Quoique nos travaux montrent la relativement bonne adaptation au vietnamien des algorithmes « classiques » existant pour l'alignement des langues occidentales, des problèmes spécifiques restent posés par certains particularismes du vietnamien, qui demandent, plus que l'adaptation de méthodes existantes, le développement d'heuristiques particulières. Le problème de l'identification des groupes de mots jouant le rôle d'entités lexicales, en particulier, semble une direction de recherche particulièrement prometteuse, et potentiellement riche d'enseignement également pour l'alignement de corpus en langues occidentales. Les performances de la partie lexicale de notre système d'alignement pourront également profiter de plusieurs améliorations existantes, comme le filtrage des paires de mots réalisant des « croisements » des occurrences alignées.

A moyen terme, le volume croissant de journaux bilingues publiés sur Internet nous donne l'espoir de pouvoir construire un corpus multilingue aligné de référence de grande taille, avec le vietnamien pour langue pivot, grâce à l'outil d'alignement structurel développé. Cela permettra d'une part une évaluation plus précise de systèmes futurs, sur une base plus « simple » et représentative qu'un texte littéraire, et d'autre part d'amorcer, grâce en particulier au système d'alignement lexical que nous avons implémenté, la constitution d'un lexique multilingue.

CONCLUSION

Comme nous avons pu le constater au Chapitre 1 de ce rapport, les ressources linguistiques jouent un rôle essentiel dans le domaine de l'ingénierie des langues. Des efforts très importants ont été investis dans la construction de lexiques, de corpus annotés et de grammaires électroniques pour le traitement des langues à plusieurs niveaux : morphosyntaxique, syntaxique, et sémantique, parallèlement au développement de techniques et outils pour automatiser cette construction. Ainsi, on trouve facilement des références sur le traitement des langues comme l'anglais, le français et beaucoup d'autres langues occidentales, ainsi que le japonais ou le chinois parmi les langues asiatiques. Les ressources linguistiques concernant ces langues sont amplement disponibles dans la communauté de TAL (accès libre ou avec licence). Ce n'est nullement le cas pour le vietnamien. Or toutes les applications de TAL, multilingues ou monolingues, requièrent plus ou moins de ces ressources. Dans ce contexte, partant d'un projet d'alignement multilingue français-vietnamien, nous nous sommes donné pour but d'établir un cadre de travail normalisé pour la construction des outils et ressources linguistiques en vue du traitement automatique de la langue vietnamienne. Si nous pouvons hériter des techniques et outils déjà développés pour d'autres langues, nous ne pouvons bien entendu pas réutiliser leurs ressources pour l'étude du vietnamien. Pour cette raison, nous nous sommes concentrée dans notre projet sur la création des ressources linguistiques spécifiques pour les traitements de base du vietnamien sur le plan monolingue, c'est-à-dire l'annotation morphosyntaxique et l'analyse syntaxique. Sur le plan multilingue, nous avons, d'une part, construit un corpus multilingue français, anglais et vietnamien de référence, et d'autre part, évalué notre outil d'alignement sur plusieurs corpus multilingues de différentes qualités.

Les problématiques abordées sont à première vue très ambitieuses, mais notre objectif n'a pas été de résoudre tous ces problèmes fondamentaux en TAL. Pour chaque tâche, nous avons tenté d'amorcer l'ensemble des démarches indispensables à sa réalisation, tout en assurant les possibilités de ré-exploitation et d'extension par la communauté de recherche en TAL au Vietnam des ressources et outils développés.

Nous résumons ci-dessous le travail effectué et en cours pour chaque type des ressources linguistiques dont l'état de l'art a éte présenté au Chapitre 1: *lexiques pour le TAL*, *grammaire électronique*, *corpus monolingues annotés* et *corpus multilingues parallèles*. Pour finir, nous abordons les projets de recherche auxquels nous participons, en insistant sur l'importance de la collaboration nationale et internationale pour le développement de l'ingénierie des langues au Vietnam.

Lexiques pour le TAL

Un lexique opérationnel pour le TAL doit contenir des informations linguistiques exploitables par des outils informatiques. Ces informations peuvent être morphologiques, syntaxiques et/ou sémantiques, servant à alimenter des outils d'annotation des corpus à chaque niveau correspondant ainsi que d'autres systèmes de TAL.

Durant notre thèse, nous avons travaillé principalement sur le lexique pour l'étiquetage lexical. Au Chapitre 3, nous avons proposé un ensemble de descripteurs lexicaux de référence pour les unités lexicales du vietnamien. Ces descripteurs sont fondés sur les descriptions linguistiques mises en avant par le Comité des Sciences Sociales du Vietnam [UYB 83], ainsi que d'autres ouvrages de référence sur la grammaire vietnamienne. Nous avons construit en collaboration avec le centre Vietlex un lexique contenant environ 35 000 mots, précisant chaque mot grâce aux descripteurs définis. Ce lexique joue un rôle essentiel pour notre système d'étiquetage lexical dans les deux étapes de segmentation et d'étiquetage des textes. Les descripteurs lexicaux proposés sont en considération dans le but de les rendre compatibles ou intégrés dans le répertoire de catégories de données défini par la norme ISO 12620.

Nous travaillons actuellement sur l'élaboration d'un lexique beaucoup plus riche en terme d'informations syntaxiques et sémantiques utiles pour les applications de TAL. Deux aspects doivent être étudiés : la définition des catégories de données convenables pour décrire les entrées lexicales du vietnamien, et le modèle du lexique. Les critères de construction du lexique mis en avant sont toujours l'extensibilité, la réutilisabité et la facilité d'échange des ressources. Le modèle LMF de la norme ISO 24613 et les catégories de données de la norme ISO 12620 sont nos principales références pour l'élaboration du lexique.

Enfin, toutes les ressources lexicales déjà construites ou en cours de construction font l'objet d'une distribution libre pour la recherche académique. Nous préparons également des outils pour l'accès et la construction coopérative du lexique sur Internet.

Grammaire

En ce qui concerne la tâche d'analyse syntaxique, nous avons étudié la modélisation de la grammaire vietnamienne par le formalisme TAG. Une première proposition de modélisation du groupe nominal a été définie, nous permettant de montrer l'adaptation du formalisme adopté aux mécanismes syntaxiques à l'œuvre en vietnamien, et de proposer un « modèle » pour le développement subséquent d'une grammaire à large couverture. Nous avons également recensé l'ensemble des phénomènes syntaxiques majeurs du vietnamien devant être pris en compte par une telle grammaire, en accompagnant ceux-ci d'exemples illustratifs, ce qui doit permettre de guider et d'évaluer le progrès des développements futurs. Cette étude nous permet également de définir une feuille de route détaillée pour le développement d'un lexique syntaxique, déjà mentionné dans le paragraphe précédent, ainsi que d'un corpus arboré du vietnamien.

La modélisation de grammaire en TAG est toujours en cours au sein de l'équipe Langue et Dialogue. Les ressources grammaticales construites sont disponibles publiquement sur la page des ressources du Loria. Nous avons également l'objectif de normaliser ces ressources afin de pouvoir les intégrer dans les archives de l'OLAC (www.language-archives.org).

Corpus monolingues annotés et outils d'annotation

Les techniques d'annotation morpho-syntaxique automatique des corpus s'appuient souvent sur un corpus d'entraînement. Nous avons donc construit dans un premier temps un corpus annoté manuellement avec l'aide des linguistes du centre Vietlex. Comme pour les autres ressources, la question de la normalisation du codage a de nouveau été mise en avant. Nous avons choisi un mode de représentation de l'annotation morpho-syntaxique conforme au modèle MAF, future norme ISO 24611.

Outre la construction du corpus de référence, nous avons développé deux outils pour l'automatisation de traitements de base : la segmentation des textes en mots et l'étiquetage lexical.

La première de ces deux tâches est réalisée grâce à un algorithme recherchant la séquence de mots du lexique de longueur minimale pouvant correspondre à une séquence de syllabes (*tiếng*) donnée. Cette méthode nous permet d'atteindre de manière totalement automatique une précision de l'ordre de 98 %. L'étiquetage est pour sa part effectué en deux étapes : dans un premier temps, le lexique construit nous permet d'associer à chaque mot d'un texte l'ensemble de ses étiquettes possibles ; les ambiguïtés apparues à la suite de cette opération sont ensuite levées par une méthode fondée sur l'utilisation conjointe de chaînes de Markov cachées et de « séquences figées » acquises sur le corpus d'apprentissage. Nous atteignons par ce moyen une précision de 95 % pour le jeu d'étiquettes correspondant aux parties du discours.

Nos recherches actuelle portent sur la construction d'un corpus arboré (*i.e.* annoté syntaxiquement) de référence pour le vietnamien. Nous souhaitons y intégrer deux types d'informations syntaxiques : les constituants syntaxiques et les rôles sémantiques. Des outils d'aide à l'annotation devront probablement être développés dans ce but.

Corpus multilingues parallèles et outils d'alignement

La construction des corpus multilingues de référence n'est pas une tâche évidente. La disponibilité des corpus parallèles pour les langues moins étudiées et dans les pays moins développés est beaucoup plus faible que pour les langues comme l'anglais ou le français. Nous avons, dans le cadre de la thèse, préparé des corpus français-vietnamien et anglais-vietnamien avec un codage normalisé des textes, afin de disposer de corpus alignés de référence (au niveau des phrases) pour les couples de langues français-vietnamien et anglais-vietnamien. L'alignement est dans un premier temps réalisé automatiquement grâce à un système d'alignement multilingue que nous avons developpé, basé sur la structure hiérarchique des documents. Il est ensuite contrôlé manuellement à l'aide d'une interface d'édition.

L'aligneur mis en œuvre opère par raffinements successifs d'alignements, niveau par niveau : division, paragraphe puis segment (phrase). Nous avons également étudié les possibilités de combinaison de cet algorithme avec une approche complémentaire réalisant l'alignement d'unités lexicales, selon une méthode originale qui permet d'obtenir des résultats intéressants sur des textes au parallélisme fortement dégradé. Nous avons évalué notre système sur des textes multilingues en français, anglais et vietnamien. Grâce au projet ARCADE II, nous avons pu évaluer notre aligneur sur des corpus de taille beaucoup plus importante en 10 langues différentes, y compris non européennes, cette évaluation confirmant la realtivement bonne qualité des résultats obtenus.

Notre outil d'alignement est accessible sur le site web du Loria. La validation des corpus alignés automatiquement est en cours, afin de rendre disponibles les corpus alignés de référence pour les couples de langues français-vietnamien et anglais-vietnamien à la communauté de TAL.

En conclusion, cette thèse est l'occasion pour nous de fonder des premières briques pour la construction des outils et des ressources linguistiques pour le traitement automatique du vietnamien. La recherche menée nous permet d'éclairer les pistes à suivre à court et moyen terme pour le développement de l'ingénierie des langues au Vietnam, dans un cadre monolingue ainsi que multilingue. Dans cette optique, nous souhaitons conclure ce document par une rapide présentation des projets actuellement en cours pour le développement du TAL pour le vietnamien.

Projets de recherche en traitement automatique du vietnamien

Les travaux présentés dans cette thèse ont été menés en collaboration entre l'équipe Langue et Dialogue du Loria, en France, et au Vietnam, l'Institut des technologies de l'information, le Centre de lexicographie (Vietlex) et l'Université des sciences de Hanoï. Il ne s'agit pas seulement de travaux scientifiques à strictement parler, mais aussi de tâches de motivation et de formation en vue de la définition d'un projet de recherche global en TAL au niveau national au Vietnam. Nous sommes ainsi parvenue, d'un point de vue plus « organisationnel », à élaborer un premier projet sur l'analyse automatique des textes du vietnamien dans le programme national de recherche et de développement de la technologie du Vietnam, et à représenter la Direction des Normes et de la Qualité du Vietnam dans le sous-comité TC 37 / SC 4 de l'ISO sur la normalisation de gestion des ressources linguistiques. Nous avons également créé une première équipe de recherche en TAL à l'Université Nationale de Hanoi, en collaboration avec le Centre de lexicographie du Vietnam à Hanoi.

Suite au projet national mentionné, en mars 2005, nous avons participé à l'organisation de la première rencontre nationale des chercheurs en TAL au Vietnam. La conférence a montré que les chercheurs dans le domaine du TAL au Vietnam sont actuellement d'une part, encore peu nombreux, et d'autre part assez peu coopératifs ; en particulier, les linguistes participent très rarement à la recherche en TAL. Il n'existe ni formation en informatique linguistique, ni formation en linguistique informatique dans les universités vietnamiennes. Nous avons également confirmé le manque important de ressources linguistiques vietnamiennes pour le TAL, ce qui ralentit considérablement la recherche. Cette conférence a fait naître un haut consensus sur le besoin de lancer un nouveau projet national de recherche et de développement des outils et ressources fondamentaux pour le traitement du vietnamien, en partant des expériences acquises sur notre projet, ainsi que ceux d'autres groupes. Ainsi, en 2006, nous avons réussi à créer un premier projet national avec la participation de chercheurs issus de 10 établissements publics du Vietnam. L'objectif principal de ce projet est de construire les outils et ressources linguistiques indispensables pour le traitement automatique du vietnamien. Les perspectives de travail que nous avons esquissées dans ce rapport trouvent donc dès maintenant un cadre pour se réaliser.

Un autre point important mis en avant, pour tous les projets à court, moyen ou long terme, est la formation. Nous avons passé un accord avec le Département de Linguistique (Université des Sciences Sociales et Humaines) afin de concevoir un premier cours en matière de Linguistique Informatique. Nous espérons donc qu'une nouvelle formation de linguistique informatique puisse être mise en place dans un proche futur.

Parallèlement à tous ces efforts nationaux, nous constatons également d'importants efforts régionaux pour la construction des outils et ressources linguistiques normalisées pour les langues asiatiques. L'organisation de l'école du TAL pour les langues asiatiques (ADD - *Asian Applied Natural Language Processing for Linguistics Diversity and Language Resource Development*, *cf. http://www.tcllab.org/*), ou le projet international sur la normalisation des ressources linguistiques pour les langues asiatiques (*cf.* http://cwn.ling.sinica.edu.tw/huang/publications/NEDO-brochure.doc) sont des exemples représentatifs de ce type d'initiatives, auxquelles nous prenons également part.

Toutes ces énergies convergentes, combinées avec l'expérience maintenant acquise en TAL grâce aux travaux consacrés aux langues européennes, nous laissent espérer que l'extension du domaine d'application du TAL aux langues encore considérées comme « minoritaires » devrait progresser à une vitesse exponentielle.

ANNEXES

Les annexes contiennent les parties suivantes :

- Descriptions lexicales du vietnamien
- Jeux d'étiquettes pour l'étiquetage lexical
- Codage TEI de dictionnaire papier du vietnamien
- Système de construction et de gestion de corpus vietnamiens annotés

Annexe A - Descriptions lexicales du vietnamien

Cette partie d'annexe précise les attributs et leurs valeurs (nommés en anglais) de chaque catégorie grammaticale du vietnamien (*cf.* section 3.3.2).

A.1. Noms

Attributs	Valeurs	Exemples
Type	proper	Việt Nam [Vietnam], Đạo Phật [Boudisme]
	common	sông [fleuve], núi [montagne], sông núi [pays]
Countability	non countable	cây cối [plantes], nhân dân [peuple], nước [eau]
	direct	cái [class.[72] des objets], mét [mètre]
	indirect	trâu [buffle], nhà [maison]
Unit	natural	con [class. des animaux], cái [class. des objets], bức [class. des panneaux]
	conventional	mét [mètre], cân [kg], giờ [heure], nắm [poignée], nhúm [pincée], hào [dix centimes], xu [centime]
	collective	tốp [groupe], đoàn [troupe], đội [équipe], đám [foule], đôi [couple], chục [dizaine]
	administrative	tỉnh [département], huyện [district], ngành[branche, division], môn [discipline, matière]
Meaning	recipient	cốc [verre], chén [tasse], thùng [caisse]
	object	cái [class. des objets], nhà [maison], ao [étang], xe [véhicule]
	plant	cây [plante, class. des plante], lúa [riz], hoa quả [fruits]
	animal	con [class. des animaux], mèo [chat], gà [poulet]
	social-family relations	thầy trò [maître et élève(s)], vợ chồng [femme et mari], cha con [père et fils/enfants], anh em [frères et sœurs]
	human	ngườI [class. des personnes], thợ [ouvrier], học sinh [élève]
	body part	tay [bras], chân [jambe], đầu [tête], sừng [corde], cọng [brin], lá [feuille], rễ [racine]

[72] classificateur, *cf.* section 2.4.1.1 - les classificateurs

	material	đá [pierre], đất [terre], sắt [fer], dầu [huile], khói [fumée]
	food	bánh [pain, gâteau], chè [compote liquide]
	disease	ho gà [coqueluche], hen [ashtme]
	sense	màu [couleur], sắc [couleur], tiếng [voix], giọng [ton, voix], mùi [odeur], vị [goût]
	location	chỗ [place, lieu, …], nơi [lieu, endroit, …], miền [région, zone], xứ [territoire, pays, …], vùng [région], phương [direction]
	time	dạo [moment], khi [fois, temps, moment], hồi [période, époque, …], chốc [instant, moment], lúc [moment, instant, temps], giây [seconde], phút [minute], buổi [séance, partie de la journée, temps], ngày [jour], tháng [mois]
	turn	lần [fois, reprise], lượt [fois, reprise, tour], phen [fois, reprise], đợt [étape]
	fact specificator	sự [fait, affaire, class. devant verbe/adjectif], việc [fait], cuộc [partie, class. devant verbe/adjectif], điều [fait, class. devant adjectif], vẻ [aspect, class. devant adjectif]
	abstract	thần [dieu, génie], ma [fantôme], hồn [esprit, âme], trí tuệ [intelligence, esprit], tình cảm [sentiment], lí thuyết [théorie], khoa học [science]
	other	units de mesure, *etc*.

A.2. Pronoms

Attributs	Valeurs	Exemples
Type	Personal	tôi [je], chúng tôi [nous]
	pronominal	mình [soi-même], tự [de soi-même, se]
	Indefinite	người ta [on], ai [quiconque]
	Time	bây giờ [maintenant], bao giờ [quand]
	Amount	cả [tout], tất thảy [tout], bao nhiêu [autant que]
	demonstrative	này [ce], kia [ce..là], nọ [ce..là], đấy [là]
	interrogative	ai [qui, quel], gì [quoi, que, quel], nào [quel], thế nào [comment]
	Predicative	thế [comme cela, ainsi, tel], vậy [comme cela, ainsi]

	Reflexive	nhau [l'un l'autre]
Person	First	tao, ta, tôi [je]
	Second	mày, cậu [tu]
	Third	nó, hắn, y [il]
Number	Singular	mày [je]
	Plural	họ [ils], chúng nó [ils], người ta [on]

A.3. Numéraux

Attributs	Valeurs	Exemples
Type	Cardinal	một [un], mười [dix], mười ba [treize]
	approximate	dăm [environ cinq, quelque], vài [deux, quelque], mươi [une dizaine]
	Fractional	nửa [un demi], rưỡi [(et) demi], rưởi [(cent, mille, … et) demi]
	Ordinal	(thứ) nhất [premier], (thứ) nhì [deuxième]

A.4. Verbes

Attributs	Valeurs	Exemples
Gradability	Gradable	thích [aimer], yêu [aimer], ghét [détester], giống [ressembler], muốn [vouloir]
	Non gradable	làm [faire], đi [aller], phảI [devoir], bắt đầu [commencer], kết thúc [finir]
Meaning	Copula	là [être], làm [se faire, jouer le rôle]
	existence	còn [(il) rester], có [avoir, il y a], mất [perdre], hết [n'avoir plus], hiện [paraître], xuất hiện [apparaître]
	transformation	hoá, biến, trở thành, nên, thành [devenir, changer en, …]
	process	bắt đầu [commencer], tiếp tục [continuer]
	comparison	giống [ressembler], khác [différer], ăn đứt [surpasser de beaucoup], bằng [égaler]
	Modal	toan [tenter], định [compter (faire)], có thể [pouvoir], nên [devoir]
	passivity	bị [subir], được [obtenir, gagner], phải [subir, contracter]

	feelings	mong [espérer], muốn [vouloir], ước [rêver], yêu [aimer], ghét [détester]
	utterance	nghĩ [réfléchir], nói [parler]
	imperative	khiến, sai , bảo, bắt [donner l'ordre (plus ou moins fort)]
	Dative	cho [donner], tặng [offrir], gửI [envoyer], lấy [recevoir]
	directive movement	ra [sortir], vào [entrer], lên [monter], xuống [descendre]
	non directive movement	đi [aller], chạy [courir], bò [ramper], lăn [rouler]
	moving	kéo [tirer], đẩy [pousser], xô [bousculer], buộc [lacer], cởi [enlever]
	Transitive action	đẽo [tailler], gọt [éplucher], vẽ [dessiner], viết [écrire]
	Intransitive action	ngồi [s'assoir], đứng [se tenir debout], nằm [s'allonger], cười [rire]

A.5. Adjectifs

Attributs	Valeurs	Exemples
Attribute	qualitative	tốt [bon], xấu [mauvais, laid], đẹp [beau]
	quantitative	cao [grand], thấp [petit], rộng [large]
Gradability	Gradable	tốt [bon], xấu [mauvais], cao [grand], thấp [petit]
	Non gradable	đen sì [trop noir]

A.6. Déterminants/Articles

Attributs	Valeurs	Exemples
Type	determined	một, mỗi, từng, mọi, cái
	undetermined	các, những, mấy
Number	singular	một, cái
	plural	các, những, mấy

A.7. Adverbes

Attributs	Valeurs	Exemples

Type	time	đã, sẽ, đang, vừa, mới, từng, xong, rồi
	degree	rất, hơi, khí, quá
	continuity	đều, cũng, vẫn, cứ, mãi, nữa, luôn luôn, thường, năng
	polarity	không, chưa, chẳng
	imperative	hãy, đừng, chớ
	effective	mất, được, ra, đi, cho
	verbal	vụt, bỗng, thình lình, quyết, nhất quyết, ắt là, chắc, quả, quả nhiên
Position	pre	rất
	post	mất, được, xong, rồi
	both	bỗng, chắc, quá

A.8. Prépositions

Attributs	Valeurs	Exemples
Type	locative	trên, dưới, trong, ngoài
	directive	từ, đến, qua, sang
	time	từ, cứ, độ
	objective	vì, cho
	target	vì, với, cho, đến, về
	relative	của, trừ, ngoài, khỏi, ở
	means	bằng, bởi, theo
	approximate	

A.9. Conjonctions

Attributs	Valeurs	Exemples
Type	coordinating	và, với, cùng, vì vậy, tuy nhiên, ngược lại
	subordinating	nếu ... thì

A.10. Interjections

Attributs	Valeurs	Exemples

Type	exclamation	ôi, chao, a ha
	onomatopoeia	ê, a, á, ối ...

A.11. Mots modaux

Attributs	Valeurs	Exemples
Type	global	à, a, á, ạ, ấy, chắc, chăng, cho, chứ…
	local	cả, cái, chẳng, chỉ, chính, có…
Meaning	opinion	chỉ, có, những, mỗi
	strengthening	thì, là, mà, đến, chính, ngay, cả
	exclamation	thôi, cho cam, chăng nữa, ư, nhỉ
	interrogation	à, ư, hử, hả, nhỉ
	call	ơi, hỡi, ạ, này
	imperative	đi, với, nhé, mà, nào, thôi...

A.12. Locutions

Attributs	Valeurs	Exemples
Category	noun	
	verb	
	adjective	
	other	

A.13. Éléments non autonomes

Attributs	Valeurs	Exemples
Location	pre	vô, phi, tiểu
	post	hoá, tặc

Annexe B - Jeux d'étiquettes utilisés pour l'étiquetage lexical

Premier jeu : jeu d'étiquettes « petit » comprenant 9 étiquettes (à l'exception des ponctuations)

1	N	Nom	6	C	Préposition et Conjonction
2	N	Verbe	7	I	Interjection
3	A	Adjectif	8	E	Mot modal
4	P	Pronom	9	X	Résidu
5	J	Adverbe			

Deuxième jeu : jeu d'étiquettes « moyen » comprenant 19 étiquettes (à l'exception des ponctuations)

1	N	Nom	11	P	Autres pronoms
2	Nn	Numéral ou déterminant	12	Vla	Verbe – être
3	Cm	Préposition	13	Va	Verbe « datif » (échange)
4	Cc	Conjonction	14	Vo	Verbe – orientation
5	Jt	Adverbe – temps	15	V	Verbes – autres
6	Jd	Adverbe – degré	16	A	Adjectif
7	Jr	Adverbe – comparatif	17	I	Interjection
8	Ja	Adverbe - affirmatif/négatif	18	E	Mot modal
9	Ji	Adverbe - impératif	19	X	Résidu
10	Pp	Pronom – personnel			

Troisième jeu : 56 étiquettes (à l'exception des ponctuations)

1	Np	Nom – propre	29	Vts	Verbe - transitif état
2	Nh	Nom – humain	30	Vtv	Verbe - transitif – volitif
3	Nna	Nom – nom administratif	31	Vito	Verbe - intransitif – orientation
4	Nc	Nom –comptable	32	Vto	Verbe - transitif – orientation
5	Nu	Nom – concrète (unité)	33	Vitt	Verbe - intransitif – transformation
6	Nn/M	Nom – numéral	34	Vtt	Verbe - transitif – transformation
7	Nt	Nom – classificateur	35	Vtc	Verbe - transitif – comparaison
8	Nl	Nom – location	36	Vta	Verbe - transitif – échange
9	No	Nom – objet	37	Vita	Verbe - intransitif – échange
10	Nai	Nom – animal	38	Vitn	Verbe - intransitif – standard
11	Npl	Nom – plantes	39	Vtn	Verbe - transitif – standard
12	Ne	Nom - plat (nourriture)	40	Aa	Adjectif – qualité
13	Nm	Nom – matériau	41	An	Adjectif – quantité
14	Nnp	Nom - phénomène naturel	42	Pp	Pronom – personnel
15	Nd	Nom – maladie	43	Pd	Pronom – démonstratif
16	Ng	Nom – collectif	44	Pn	Pronom – quantité
17	Na	Nom – abstrait	45	Pa	Pronom – qualité
18	Nx	Nom – résidu	46	Pi	Pronom – interrogatif
19	Vla	Verbe – être	47	Jt	Adverbe – temps
20	Vitf	Verbe - intransitif - émotion	48	Jd	Adverbe – degré
21	Vtf	Verbe - transitif - émotion	49	Jr	Adverbe – comparatif
22	Vitd	Verbe - intransitif - discours	50	Ja	Adverbe - affirmatif/négatif
23	Vtd	Verbe - transitif - discours	51	Ji	Adverbe – impératif
24	Vite	Verbe - intransitif - existence	52	Cm/O	Préposition
25	Vte	Verbe - transitif - existence	53	Cc/C	Conjonction
26	Vitm	Verbe - intransitif - mouvement	54	E	Mot modal
27	Vtm	Verbe - transitif - mouvement	55	I	Interjection
28	Vits	Verbe - intransitif – état	56	X	Résidu

Annexe C – Codage TEI de dictionnaire papier du vietnamien

La DTD (XML) pour le dictionnaire papier de Vietlex est basée sur le schéma de codage de dictionnaires de la TEI. La définition de DTD, ainsi qu'une interface de consultation du dictionnaire XML obtenu sur le Web sont mises en oeuvre par Nguyen Thanh Bon [NGU 04b].

```
<!-- <!DOCTYPE TEI.2 SYSTEM "xteivndict.dtd" [ -->
<!--  TEI DTD for Vietnamese Dictionary-->
<!--  Text Encoding Initiative: Guidelines for Electronic      -->
<!--  A TEI document is a text preceded by a TEI header.       -->
<!ENTITY % a.dictionaries '
    expand CDATA #IMPLIED
    norm CDATA #IMPLIED
    split CDATA #IMPLIED
    value CDATA #IMPLIED
    orig CDATA #IMPLIED
    location IDREF #IMPLIED
    mergedin IDREF #IMPLIED
    opt (y | n) "n"'>
<!--Next, we declare some specialized element classes, used in various content models in the dictionary tag set.-->
<!ENTITY % a.entries '
    type CDATA "main"
    key CDATA #IMPLIED'>
<!ENTITY % xml.global '
    xml:lang          CDATA #IMPLIED'>
<!ENTITY % a.global '
    %xml.global;
    id   ID   #IMPLIED
    n    CDATA #IMPLIED
    rend     CDATA #IMPLIED'>
<!ELEMENT TEI.2.dictionaries (teiHeader, body)>
<!ATTLIST TEI.2.dictionaries
    %a.global;
>
<!ELEMENT teiHeader (fileDesc, (encodingDesc)*)>
<!ATTLIST teiHeader
    %a.global;
    corresp IDREFS #IMPLIED
    next IDREF #IMPLIED
    prev IDREF #IMPLIED
    type CDATA "text"
```

```
      creator CDATA #IMPLIED
      status (new | update) "new"
      date.created CDATA #IMPLIED
      date.updated CDATA #IMPLIED
>
<!ELEMENT fileDesc (#PCDATA)>
<!ATTLIST fileDesc
      %a.global;
      corresp IDREFS #IMPLIED
      next IDREF #IMPLIED
      prev IDREF #IMPLIED
      TEIform CDATA "fileDesc"
>
<!--  5.3:  The encoding description                    -->
<!ELEMENT encodingDesc (#PCDATA)>
<!ATTLIST encodingDesc
      %a.global;
      ana IDREFS #IMPLIED
      corresp IDREFS #IMPLIED
      next IDREF #IMPLIED
      prev IDREF #IMPLIED
      TEIform CDATA "encodingDesc"
>
<!ELEMENT body (superEntry | entry)*>
<!ATTLIST body
      %a.global;
      corresp IDREFS #IMPLIED
      next IDREF #IMPLIED
      prev IDREF #IMPLIED
      decls IDREFS #IMPLIED
      TEIform CDATA "body"
>
<!--[declarations from 12.2.1: Dictionary entries and their structure inserted here ] -->
<!ELEMENT superEntry ((form)?, (entry)+)>
<!ATTLIST superEntry
      %a.global;
      %a.entries;
      TEIform CDATA "superEntry"
>
<!ELEMENT entry (hom | sense | def | eg | form | gramGrp | note | re | trans | usg | xr | ref)+>
<!ATTLIST entry
      %a.global;
      %a.entries;
      TEIform CDATA "entry"
```

```
>
<!ELEMENT hom (sense | def | eg | form | gramGrp | note | re | trans | usg | xr | ref)*>
<!ATTLIST hom
    %a.global;
    %a.entries;
    TEIform CDATA "hom"
>
<!ELEMENT sense (sense | def | eg | form | gramGrp | note | re | trans | usg | xr | ref)+>
<!ATTLIST sense
    %a.global;
    %a.dictionaries;
    level CDATA #IMPLIED
    TEIform CDATA "sense"
>
<!--[declarations from 12.3.1: The form group inserted here ]-->
<!ELEMENT form (#PCDATA | orth | pron | hyph | syll | gram)*>
<!ATTLIST form
    %a.global;
    %a.dictionaries;
    type CDATA #IMPLIED
    TEIform CDATA "form"
>
<!ELEMENT orth (#PCDATA)>
<!ATTLIST orth
    %a.global;
    %a.dictionaries;
    type CDATA #IMPLIED
    extent CDATA "full"
    TEIform CDATA "orth"
>
<!ELEMENT pron (#PCDATA)>
<!ATTLIST pron
    %a.global;
    %a.dictionaries;
    notation CDATA #IMPLIED
    extent CDATA "full"
    TEIform CDATA "pron"
>
<!ELEMENT hyph (#PCDATA)>
<!ATTLIST hyph
    %a.global;
    %a.dictionaries;
    TEIform CDATA "hyph"
>
```

```
<!ELEMENT syll (#PCDATA)>
<!ATTLIST syll
    %a.global;
    %a.dictionaries;
    TEIform CDATA "syll"
>
<!ELEMENT gram (#PCDATA)>
<!ATTLIST gram
    %a.global;
    %a.dictionaries;
    type CDATA #IMPLIED
    TEIform CDATA "gram"
>
<!--[declarations from 12.3.2: The grammatical group inserted here ]-->
<!ELEMENT gramGrp (#PCDATA | pos | subc | colloc)*>
<!ATTLIST gramGrp
    %a.global;
    %a.dictionaries;
    TEIform CDATA "gramGrp"
>
<!ELEMENT pos (#PCDATA)>
<!ATTLIST pos
    %a.global;
    %a.dictionaries;
    TEIform CDATA "pos"
>
<!ELEMENT subc (#PCDATA)>
<!ATTLIST subc
    %a.global;
    %a.dictionaries;
    TEIform CDATA "subc"
>
<!ELEMENT colloc (#PCDATA)>
<!ATTLIST colloc
    %a.global;
    %a.dictionaries;
    type CDATA #IMPLIED
    TEIform CDATA "colloc"
>
<!--[declarations from 12.3.3.1: Definition text inserted here ]-->
<!ELEMENT def (#PCDATA)>
<!ATTLIST def
    %a.global;
    %a.dictionaries;
```

```
    TEIform CDATA "def"
>
<!--[declarations from 12.3.3.2: Translation information inserted here ] -->
<!ELEMENT trans (#PCDATA)>
<!ATTLIST trans
    %a.global;
    %a.dictionaries;
    TEIform CDATA "trans"
>
<!--[declarations from 12.3.4: Etymologies inserted here ]-->
<!--[declarations from 12.3.5.1: Examples and citations inserted here ] -->
<!ELEMENT eg (#PCDATA)>
<!ATTLIST eg
    %a.global;
    %a.dictionaries;
    TEIform CDATA "eg"
>
<!--[declarations from 12.3.5.2: Usage information inserted here ]-->
<!ELEMENT usg (#PCDATA)>
<!ATTLIST usg
    %a.global;
    %a.dictionaries;
    type CDATA #IMPLIED
    TEIform CDATA "usg"
>
<!ELEMENT lbl (#PCDATA)>
<!ATTLIST lbl
    %a.global;
    %a.dictionaries;
    type CDATA #IMPLIED
    TEIform CDATA "lbl"
>
<!--[declarations from 12.3.5.3: Cross References inserted here ]-->
<!ELEMENT xr (#PCDATA | usg | lbl | ref)*>
<!ATTLIST xr
    %a.global;
    %a.dictionaries;
    type CDATA #IMPLIED
    TEIform CDATA "xr"
>
<!ELEMENT ref (#PCDATA)>
<!ATTLIST ref
    %a.global;
    %a.dictionaries;
```

```
    type CDATA #IMPLIED
    TEIform CDATA "ref"
>
<!--[declarations from 12.3.6: Related entries inserted here ]-->
<!ELEMENT re (#PCDATA | sense | def | eg | form | gramGrp | note | re | trans | usg | xr | ref)*>
<!ATTLIST re
    %a.global;
    %a.dictionaries;
    type CDATA #IMPLIED
    TEIform CDATA "re"
>
<!--[declarations from 12.4: Headword references inserted here ]-->
<!--  6.12: Elements available in all forms of the TEI main    -->
<!--  DTD                                                      -->
<!--  6.8.1:  Annotation                                       -->
<!ELEMENT note (#PCDATA)>
<!ATTLIST note
    %a.global;
    type CDATA #IMPLIED
    resp CDATA #IMPLIED
    place CDATA "unspecified"
    anchored (yes | no) "yes"
    target IDREFS #IMPLIED
    targetEnd IDREFS #IMPLIED
    TEIform CDATA "note"
>
<!--  14.2.1:  Extended pointers                               -->
<!ELEMENT xref (#PCDATA)>
<!ATTLIST xref
    %a.global;
    ana IDREFS #IMPLIED
    corresp IDREFS #IMPLIED
    next IDREF #IMPLIED
    prev IDREF #IMPLIED
    type CDATA #IMPLIED
    resp CDATA #IMPLIED
    crdate CDATA #IMPLIED
    targType CDATA #IMPLIED
    targOrder (Y | N | U) "U"
    evaluate (all | one | none) #IMPLIED
    doc ENTITY #IMPLIED
    from CDATA "ROOT"
    to CDATA "DITTO"
    TEIform CDATA "xref"
```

>

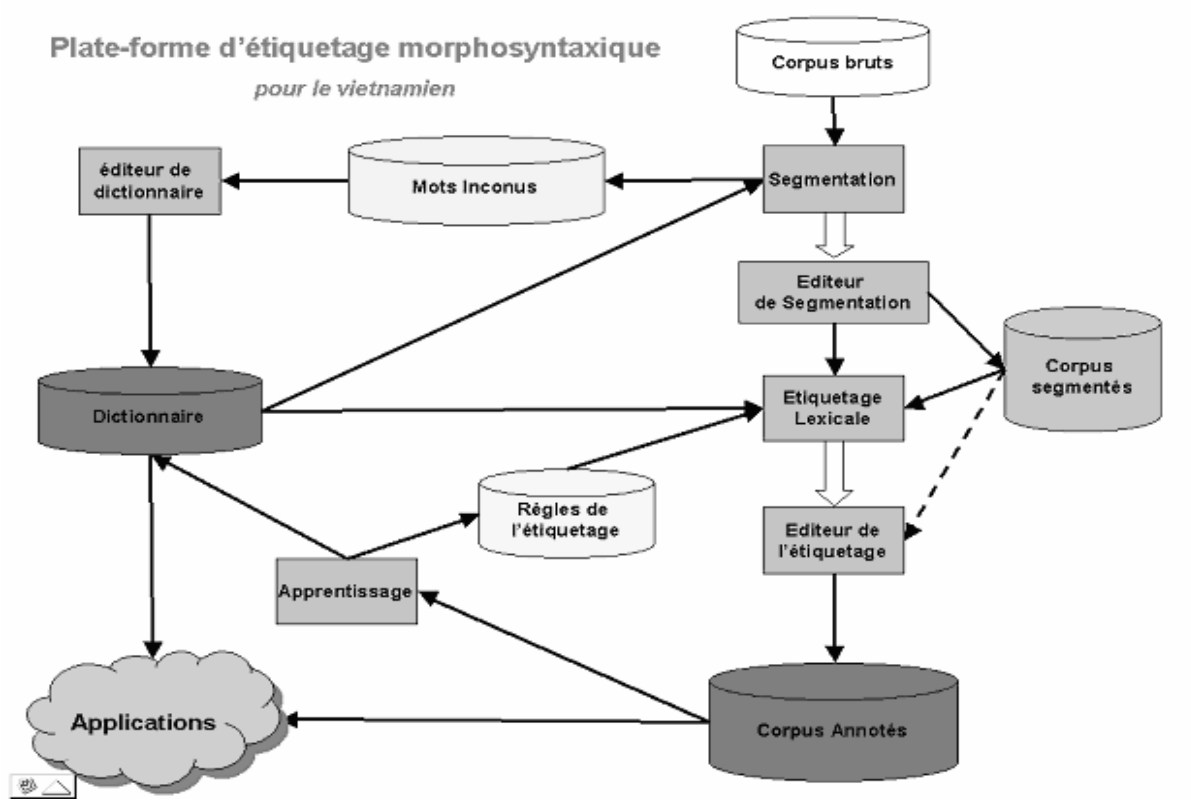
Annexe D - Système de construction et de gestion de corpus vietnamiens annotés

Le schéma ci-dessous présente le système de construction et de gestion de corpus vietnamiens annotés, qui intègre différents outils permettant :

- la segmentation automatique de texte en unités lexicales,
- l'étiquetage morphosyntaxique automatique,
- la révision manuelle de la segmentation et de l'étiquetage
- l'édition du lexique morphosyntaxique.

Le système gère toutes les ressources au format XML, suivant les schémas de codage standardisés. Il offre aux utilisateurs l'accès à ces données par des interfaces simples, ne requérant aucune connaissance de XML.

Deux autres applications sont également implémentées dans le cadre du stage de NGUYEN T. B. [NGU 04b] : une interface pour la consultation du dictionnaire vietnamien en ligne, et un concordancier permettant la recherche des contextes des mots et leur fréquences.



GLOSSAIRE

Ce glossaire est principalement destiné à éclaircir les notions mentionnées au chapitre 4. Il rassemble les définitions des termes spécifiques au domaine des grammaires formelles. Les définitions sont pour la plupart empruntées au glossaire de l'ouvrage d'Anne Abeillé [ABE 93], parfois légèrement modifiée pour des raisons de clarté.

Adresse d'un nœud dans un arbre : Identification de sa position. Par convention, l'adresse du nœud racine est 0 (ou ε). On calcule l'adresse d'un nœud quelconque (selon la convention de Gorn) en concaténant l'adresse de son père et le rang du nœud. Ainsi, l'adresse des nœuds immédiatement dominés par la racine est (de gauche à droite) 1, 2 *etc*., celle des nœuds immédiatement dominés par les précédents est (toujours de gauche à droite) 1.1, 1.2 *etc*. pour les descendants du premier nœud, 2.1, 2.2 *etc*. pour les descendants du deuxième nœud *etc*.

Arbre élémentaire (TAG) : Arbre de profondeur finie, unité de base d'une `grammaire` TAG.

Arbre dérivé (TAG) : Arbre résultant de la combinaison de plusieurs `arbres élémentaires`.

Arbre de dérivation : Représentation de la `dérivation` d'une phrase du `langage` pour une `grammaire régulière` ou `hors contexte`. Dans le modèle TAG, c'est la représentation arborescente d'une combinaison d'`arbres élémentaires` : les nœuds sont étiquetés par un couple (arbre élémentaire, `adresse`) et les branches ne sont pas ordonnées.

Attribut : Nom d'un `trait`.

Capacité générative « faible » d'une `grammaire` : `Langage` généré par la grammaire (ensemble des séquences de terminaux dérivables).

Capacité générative « forte » d'une `grammaire` : Ensemble des `dérivations terminées` générées par la grammaire (ou ensemble des descriptions syntagmatiques générées).

Catégorie : Symbole auxiliaire (non `terminal`) d'une `grammaire`. En GPSG et HPSG, les catégories sont définies comme ensembles de `traits`.

Coindexation : Partage de valeur entre deux `attributs`. Cela est noté, en termes de graphes, par le fait que deux arcs pointent sur le même nœud. S'il s'agit d'attributs à valeur non atomique, la coindexation garantit que les valeurs des deux attributs seront toujours les mêmes, et seront simultanément mises à jour si l'une ou l'autre est modifiée par `unification`.

Constituant immédiat : Un nœud *x* est un constituant immédiat d'un nœud *y* si *x* est immédiatement dominé par *y*. Un ensemble de nœuds forment un constituant immédiat d'un nœud si et seulement si ils sont complètement dominés par un nœud commun.

C-commande (c = constituant) : Le nœud A c-commande le nœud B *sii* (i) A ne domine pas B et B ne domine pas A, et (ii) le premier nœud branchant dominant A domine aussi B

Dérivation directe (noté $\Rightarrow$) : Si la règle $\psi \rightarrow \omega$ appartient à la grammaire, toute séquence (ou chaîne) $\alpha\omega\beta$, obtenue par concaténation d'ω avec α et β (qui sont des séquences d'éléments terminaux ou auxiliaires), est directement dérivable à partir de la séquence $\alpha\psi\beta$.

Dérivation (notée $\Rightarrow^*$) : Une séquence ω_n (obtenue par concaténation d'occurrences d'éléments terminaux ou auxiliaires) est dérivable à partir d'une séquence ω_1 *ssi* il existe un ensemble de séquences tel que $\omega_1 \Rightarrow \omega_2$, $\omega_2 \Rightarrow \omega_3$ *etc*. jusqu'à $\omega_{n-1} \Rightarrow \omega_n$. Une dérivation **terminée** ne porte dans sa partie droite que des éléments terminaux (mots du langage). Pour les grammaires régulières ou hors contexte, toute dérivation est représentable sous forme d'arbre, dessiné en attachant les éléments de la partie droite d'une règle utilisée comme descendants immédiats de l'élément de la partie gauche.

Dominance : Dans un arbre (ou graphe orienté), un nœud A domine un nœud B s'il existe un chemin de A vers B.

Dominance immédiate : Un nœud A domine immédiatement un nœud B si A domine B et qu'il n'existe aucun nœud C tel que A domine C et C domine B.

Étoile de Kleen : Symbole * qui permet de noter un nombre quelconque (de 0 à ∞) d'occurrences d'un élément donné. Par exemple, dans la règle suivante : $N' \rightarrow N\,(Adj)^*$, on note ainsi que N peut être suivi d'un nombre quelconque d'adjectifs et former N'.

Extension : Une structure de traits A est une extension d'une structure de traits B ($A \supset B$) *ssi* tous les traits atomiques présents dans B sont présents dans A avec la même valeur, et tous les traits complexes présents dans B ont pour valeur dans A une extension de leur valeur dans B. De façon intuitive, une structure plus spécifique est une extension d'une structure plus générale (ou moins spécifiée). L'extension permet de définir une relation d'ordre partiel entre structures de traits.

Fille : Une catégorie A est fille d'une catégorie B *ssi* B est la mère de A.

Généralisation : La généralisation de deux structures de traits A et B, notée A **G** B, est la structure de traits maximale qui subsume à la fois A et B. C'est-à-dire que A est une extension de A **G** B et B est une extension de A **G** B. Contrairement à l'unification, la généralisation ne peut pas échouer.

Grammaire (grammaire formelle) : Dans le cadre de ce glossaire, mot employé pour désigner une Grammaire de réécriture.

Grammaire ambiguë : Grammaire qui dérive au moins une séquence par plusieurs dérivations différentes. (Note : Dans le cas d'une grammaire de constituants, deux dérivations différentes correspondent à deux arbres différents.)

Grammaire arborescente : Grammaire de constituants où les règles de production sont remplacées par des arbres élémentaires. Différentes opérations peuvent être définies pour combiner entre eux les arbres élémentaires.

Grammaires équivalentes : Deux grammaires sont « faiblement » équivalentes si elles génèrent le même langage. Elles sont « fortement » équivalentes si elles génèrent le même langage par les mêmes dérivations (c'est-à-dire en associant les mêmes descriptions syntagmatiques aux mêmes phrases).

Grammaire de réécriture : Quadruplet comprenant un vocabulaire terminal noté Vt (mots du langage), un vocabulaire auxiliaire noté Va (non terminaux ou catégories ; $Vt \cap Va = \phi$), un symbole auxiliaire distingué (en général P pour Phrase, $P \in Va$), un ensemble de règles de production (ou règles de réécriture) de la forme : $\psi \rightarrow \omega$, où ψ et $\omega \in (Vt \times Va)^*$.

Grammaire de réécriture non contrainte (récursivement énumérables ou de type 0) : *cf.* Hiérarchie de Chomsky.

Grammaire contextuelle ou sensible au contexte (CSG - *Context Sensitive Grammar*), encore appelée grammaire de type 1 : *cf.* `Hiérarchie de Chomsky`.

Grammaire « légèrement » contextuelle (MCSG - *Midly Context Sensitive Grammar*) : `Grammaire` qui engendre un sous-ensemble des `langages` contextuels (comprenant des langages qui ne sont pas générés par une `grammaire hors contexte`).

Grammaire de constituants (ou **grammaire syntagmatique**) : `Grammaire de réécriture` de type 2 ou 3 (*cf.* `Hiérarchie de Chomsky`). Par extension, on appelle grammaire syntagmatique un modèle syntaxique qui se base sur une grammaire de type 2 (ou 3) même s'il est en fait équivalent à une grammaire plus puissante (HPSG).

Grammaire hors contexte ou indépendante du contexte (*CF*G - *Context Free Grammar*), encore appelée grammaire algébrique ou grammaire de type 2 : *cf.* `Hiérarchie de Chomsky`.

Grammaire lexicalisée : consiste (1) en un ensemble fini de structures associées chacune à un item lexical, à partir duquel se définit le domaine de localité dans lequel les contraintes s'appliquent ; (2) en opérations de composition de structures: chaque item lexical est nommé l'ancre de la structure correspondante => les contraintes sont locales par rapport à l'ancre.

Grammaire régulière (ou grammaire de Kleene), encore appelée de type 3 : *cf.* `Hiérarchie de Chomsky`.

Hiérarchie de Chomsky : [CHO 57] distingue quatre classes de `grammaires` et de `langages` selon la forme des règles de production utilisées.

- Les ***grammaires régulières*** (ou grammaires de Kleene - type 3) : $|\psi| = 1$, ψ contient un (seul) symbole auxiliaire, ω contient au plus un symbole auxiliaire et un nombre quelconque de symboles terminaux (qui doivent tous précéder ou suivre le symbole auxiliaire éventuel). On aura par exemple les règles : $A \rightarrow Aa$ ou $A \rightarrow abB$ mais pas $A \rightarrow AB$ ni $A \rightarrow bAa$;
- Les ***grammaires hors contexte*** (ou indépendantes du contexte - type 2) : $|\psi| = 1$, ψ contient un (seul) symbole auxiliaire, ω contient un nombre quelconque de symboles terminaux ou auxiliaires. On aura par exemple la règle : $A \rightarrow AB$;
- Les ***grammaires contextuelles*** (ou sensible au contexte - type 1) : ψ ne contient pas plus de symboles que ω. Les règles d'une grammaire de ce type peuvent se mettre sous la forme : $uxv \rightarrow uyv$, où x est un symbole auxiliaire, y une séquence non vide et u et v des séquences d'éléments de Vt ou Va. On dit que u..v forme le contexte du symbole x qui est réécrit. On aura par exemple la règle : $aAb \rightarrow aBAb$ qui réécrit A en BA dans le contexte a..b.
- Les ***grammaires non contraintes*** (ou récursivement énumérables - type 0) : $|\psi| > 0$ et ψ contient au moins un symbole auxiliaire. On aura par exemple la règle : $AB \rightarrow A$.

On a une relation d'inclusion stricte entre ces classes de grammaires, ainsi qu'entre les classes de langages correspondantes :

- Grécursiv-énum $\supset$ Gsensibles-contexte $\supset$ Ghors-contexte $\supset$ Grégulières
- Lrécursiv-énum $\supset$ Lsensibles-contexte $\supset$ Lhors-contexte $\supset$ Lrégulières

Instanciation : On instancie une variable en la remplaçant par une constante. On instancie une `structure de traits` quand on instancie toutes les variables qu'elle contient.

Langage : Ensemble des séquences de `terminaux` dérivables (*cf.* `Dérivation`) par les `règles de production` à partir du symbole distingué de la `grammaire`. C'est un sous-ensemble de l'ensemble de toutes les combinaisons possibles d'occurrences d'éléments du Vocabulaire terminal correspondant.

Mère : Une `catégorie` A est mère d'une catégorie B s'il existe une `règle de réécriture` où A apparaît en partie gauche et B en partie droite.

Monotonie : L'opération d'`unification` est dite monotone en vertu des propriétés suivantes (voir aussi `Extension`) :

- $A \cup B \supset A$
- $A \cup B \supset B$
- Si $A \supset B$ alors $\forall C$, $A \cup C \supset B \cup C$,

c'est-à-dire que les relations d'`extension` sont conservées par l'unification. On dit, de façon informelle, que l'unification « ajoute » de l'information, sans en ôter. Un formalisme est dit monotone s'il n'utilise que des opérations monotones.

Polarité : Valeur (positive ou négative) d'un `trait booléen`.

Précédence : Le nœud A précède le nœud B *ssi* A se trouve à gauche de B et A ne domine pas B et B ne domine pas A.

Préterminal : Catégorie mère d'un élément terminal.

Principe de projection (Chomsky) : L'information lexicale doit être représentée à tous les niveaux de la grammaire.

Quasi-arbre : Cas particulier d'arbre non standard, utilisé en grammaire d'arbres adjoints (TAG), où l'on peut ne spécifier que partiellement les relations de dominance. En particulier, la relation de dominance est définie comme réflexive et les distances entre nœuds peuvent être sous-spécifiées (*cf.* chapitre 4, section 5.3 de l'ouvrage [ABE 93]).

Règle lexicale : Il s'agit d'un terme ambigu, utilisé dans au moins trois sens distincts. En `grammaire` formelle, il s'agit d'une `règle de production` dont la partie droite est réduite à un élément `terminal`, c'est-à-dire un mot du langage. En GPSG, on appelle règles lexicales les règles DI dont la catégorie Tête (en partie droite) est un symbole auxiliaire correspondant à un `préterminal`, par exemple : $SV \rightarrow T$, SN (avec T = V). En LFG, en HPSG et en TAG, les règles lexicales expriment des relations régulières entre différentes descriptions d'entrées lexicales, c'est-à-dire des relations entre équations fonctionnelles (LFG), entre structures de traits typées (HPSG) ou entre arbres élémentaires (TAG).

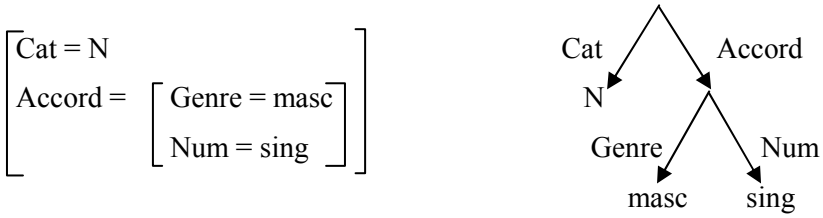
Règles de production (ou **règle de réécriture**) : Règle de la forme : $\psi \rightarrow \omega$, où ψ et ω sont des séquences d'éléments du vocabulaire `terminal` ou auxiliaire. On appelle ψ la partie gauche de la règle et ω sa partie droite. La flèche se lit « se réécrit ». Une contrainte générale est que ψ soit non vide et comporte au moins un symbole auxiliaire. Une autre contrainte, qui rend le `langage` (généré par la `grammaire`) **décidable**, est que la longueur de ψ ne dépasse pas celle de ω. Une règle de production **non branchante** a sa partie droite réduite à un seul symbole, par exemple : $SV \rightarrow V$.

Saturation : Un prédicat (ou une fonction) est saturé quand tous ses arguments sont instanciés. On parle en HPSG de syntagmes saturés quand leur trait Sous-cat a la valeur vide.

Sous-catégorisation : Faculté pour un élément lexical de nécessiter un certain nombre de ou un certain type d'éléments dans son entourage. Par exemple, on dit qu'un verbe sous-catégorise un syntagme quand on ne peut pas le construire sans ce type de syntagme.

Structure de traits (d'autres termes : complexe de traits, structure attribut-valeur SAV) : Un ensemble de `traits`, généralement noté entre crochets, qui respecte la contrainte qu'un `attribut` ne peut apparaître au même niveau plusieurs fois avec une valeur différente[73]. Une structure de traits ***complexe*** comporte des `traits complexes`. Une structure de traits ***vide*** ne comprend aucun trait. On peut représenter une structure de trait par un graphe orienté (*cf*. l'exemple). Une structure de traits ***cyclique*** comprend au moins un cycle dans la représentation graphique.

Exemple : Voici deux représentations (entre crochets et graphe) d'une même structure de traits :



Structure de traits « réentrante » (ou à réentrée) : Structure de traits dont au moins deux attributs sont conindexés (partagent la même valeur, *cf*. `Coindexation`).

Structure profonde : Lieu de représentation de la descriptison structurale et produite par les règles de la syntaxe.

Structure de surface : Entrée de la représentation phonétique produite par des règles de transformation appliquées à la structure profonde. Structure de surface contient des traces et un ensemble d'informations interprétables en forme logique et forme phonologique.

Subsomption : Relation inverse de la relation d'`extension`. De façon intuitive, une `structure de traits` plus générale **subsume** une structure de traits comprenant des informations plus spécifiques. Toute structure se subsume elle-même.

Terminal : Élément du vocabulaire d'une `grammaire` correspondant à un mot du `langage`. Les éléments non terminaux sont les symboles auxiliaires de la grammaire (par exemple N, V *etc*.).

Trait : Un couple formé d'un attribut (qui est le nom du trait) auquel est associée une valeur, parmi un ensemble de valeurs définies *a priori*. Cette valeur peut être également une variable. Notations : `[Attribut = valeur]` ou `<Attribut> = valeur`.

Trait atomique (**à valeur atomique**) : Trait dont la valeur est un symbole non décomposable, par ex. `[Genre = masc]`[74].

Trait binaire (**booléen**) : Trait atomique dont la valeur est égale à `+` ou `-`, et peut être calculée à l'aide de connecteurs logiques (ET, OU), par ex. `[Humain = +]`[75].

Trait complexe (**non atomique**) : Trait dont la valeur est une liste ou un ensemble de traits, par ex. `[Accord = [Num = sing, Genre = masc]]`[76].

Type : Ensemble de `traits` (ou de propriétés) prédéfini auquel on donne un nom. Par exemple, le type *nom* ou *syntagme* en HPSG. Une `structure de traits` est **bien typée** si elle comprend tous les traits prévus par son type et que ceux-ci sont instanciés (*cf*. `Instanciation`). On définit une hiérarchie entre types, selon la relation d'ordre (`extension`) définie sur les structures de traits.

---

[73] On peut ajouter certains opérateurs aux structures de traits comme la négation ou la disjonction des valeurs de trait.

[74] i.e. le genre est masculin.

[75] i.e. l'attribut « Humain » porte la valeur « positive ».

[76] i.e. les valeurs d'accord sont : l'attribut « nombre » porte la valeur « singulier », l'attribut « genre » porte la valeur « masculin ».

Unification : En logique, c'est une substitution (remplacement de toutes les occurrences d'une variable par une constante) qui rend deux clauses identiques. En grammaire d'unification, on appelle unification de deux `structures de traits` A et B (notée $A \cup B$) la structure minimale qui est à la fois une `extension` de A et de B, si elle existe. Si elle n'existe pas, on dit que l'unification « échoue ». L'opération d'unification est idempotente ($A \cup A = A$), commutative ($A \cup B = B \cup A$) et associative ($(A \cup B) \cup C = A \cup (B \cup C)$).

L'unification de deux structures de traits typées $[A]_t$ et $[B]_{t'}$ est définie comme suit :

$[A]_t \cup [B]_{t'} = [A \cup B]_{t''}$

avec $t'' \leq t$ et $t'' \leq t'$ et t'' est maximal parmi les sous-types communs à t et t'

(i.e. pour tout type $t°$, on a l'implication : $(t° \leq t \text{ et } t° \leq t') \Rightarrow t° \leq t''$).

Valence (d'un verbe) : Nombre d'actants qu'il est susceptible de recevoir.

BIBLIOGRAPHIE

[ABE 93] Abeillé A., « Les nouvelles syntaxes », Armand Colin Editeur, Paris, FR, 1993.

[ABE 00] Abeillé A., Blache P., « Grammaires et analyseurs syntaxiques », in Pierrel J-M. (ed.) *Ingénierie des langues,* Hermes Science Europe, p. 51 - 76, 2000.

[ABE 02] Abeillé A., « Une grammaire d'arbres adjoints pour le français », Editions du CNRS, Paris, FR, 2002.

[ABE 03a] Abeillé A., « Treebanks - Building and Using Parsed Corpora », Dordrecht: Kluwer Academic Publishers, 2003.

[ABE 03b] Abeillé A., Clément L., Toussenel F., « Building a treebank for French », in Abeillé A. (ed.) *Treebank - Building and Using Parsed Corpora*, Dordrecht, Kluwer Academic Publishers, p. 165-188, 2003.

[ADD 99] Adda G., Mariani J.,, Paroubek P., Rajman M., Lecomte J., « Métrique et premiers résultats de l'évaluation GRACE des étiqueteurs morphosyntaxiques pour le français », Actes de la *Conférence francophone internationale sur le Traitement Automatique des Langues Naturelles (TALN 99)*, Cargèse, FR, 1999.

[ANT 93] Antoni-Lay M.-H., Francopoulo G., Zaysser L., « A generic model for re-usable lexicons: Genelex Project », *Literary and Linguistic Computing vol. 8 n. 4*, 1993.

[BAK 03] Baker C. F., Fillmore C. J., Cronin B., « The structure of the FrameNet database », *International Journal of Lexicography*, 16(3), 2003.

[BAL 94] Balkan L., « Test Suites: some issues in their use and design », in Cranfield International Conference on Machine Translation: *Machine Translation: Ten Years on*, Cranfield University, 12-14 November 1994.

[BEL 92] Belaïd A., Belaïd Y., « Reconnaissance des formes - Méthodes et applications », InterEditions, Paris, 1992.

[BLO 33] Bloomfield L., "Language", Allen & Unwin, New York, 1933.

[BOI 01] Boitet C., « Méthodes d'acquisition lexicale en TAO : des dictionnaires spécialisés propriétaires aux bases lexicales généralistes et ouvertes », Actes de la *Conférence francophone internationale sur le Traitement Automatique des Langues Naturelles (TALN 01)*, Tours, FR, 2001.

[BON 00a] Bonhomme P., « Codage et normalisation de ressources textuelles », in Pierrel J-M. (ed.) *Ingénierie des langues,* Hermes Science Europe, p.173-191, 2000.

[BON 00b] Bonhomme P., Lopez P., « TAGML : codage XML et ressources pour les grammaires d'arbres adjoints lexicalisés », in *Proceedings of 2th International Conference on Language Ressources and Evaluation* (LREC 2000), Athènes, GR, 2000.

[BOU 94] Bourigault D., « LEXTER, un Logiciel d'EXtraction de TERminologie. Application à l'acquisition des connaissances à partir de textes », *Thèse de doctorat en informatique linguistique*, École des Hautes Études en Sciences Sociales, Paris, FR, 1994.

[BRA 98] Bray T., Paoli J., Sperberg-McQueen C.M., « Extensible Markup Language (XML) 1.0 », W3C Recommandation, http://www.w3.org/TR/REC-xml, 1998.

[BRA 03] Brants T., Skut W., Uszkoreit H., « Syntactic annotation of a German newspaper corpus », in Abeillé A. (ed.) *Treebank - Building and Using Parsed Corpora*, Dordrecht, Kluwer Academic Publishers, p. 73-88, 2003.

[BRE 82] Bresnan J., « The mental representation of grammatical relations », MIT Press, 1982.

[BRI 95] Brill E., « Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part of Speech Tagging », *Computational Linguistics*, 21(4), p.543-565, 1995.

[BRO 91] Brown P.F., Lai J.C., Mercer R.L., « Aligning sentences in parallel corpora », in *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, Berkeley, p. 169-176, 1991.

[BUR 90] Burnage G., « CELEX: A Guide for Users », Center for Lexical Information, University of Nijmegen, 1990.

[BUR 95] Burnard L. (ed.), « User's reference guide for the British National Corpus version 1.0 », Oxford, Oxford University Computing Services, 1995.

[CAL 98] De Calmès M., Pérennou G., « BDLEX : a Lexicon for Spoken and Written French », in *Proceedings of 1st International Conference on Langage Resources & Evaluation* (LREC1998), Grenade, SP, 1998.

[CAO 00] Cao Xuân Hạo, « Tiếng Việt - mấy vấn đề ngữ âm, ngữ pháp, ngữ nghĩa (Vietnamese - Some Questions on Phonetics, Syntax and Semantics) », NXB Giáo dục, Hanoi, 2000.

[CAR 03] Carroll J., Minnen G., Briscoe T., « Parser Evaluation Using a Grammatical Relation Annotation Scheme », in Abeillé A. (ed.) *Treebank - Building and Using Parsed Corpora*, Dordrecht, Kluwer Academic Publishers, 2003.

[CHA 95] Chanod J.-P., Tapanainen P., « Creating a tagset, lexicon and guesser for a French tagger », ACL SIGDAT workshop on "From texts to tags: Issues in multilingual language analysis", p. 58-64, University College, Dublin, Ireland, 1995.

[CHA 04] Charoenporn T., Sornlertlamvanich V., Kasuriya S., Hansakunbuntheung C., Isahara H., « Open Collaborative Development of the Thai Linguistics Resources », in *Proceedings of 4th International Conference on Language Ressources and Evaluation* (LREC04), Lisbon, PT, 2004.

[CHE 03] Chen K-J., Luo C-C., Chang M-C., Chen F-Y., Chen C-J., Huang C-R., « SINICA Treebank. Design criteria, representational issues and implementation », in Abeillé A. (ed.) *Treebank - Building and Using Parsed Corpora*, Dordrecht, Kluwer Academic Publishers, p. 231-248, 2003.

[CHI 06] Chiao Y.-C., Kraif O., Laurent D., Nguyen T.M.H., Semmar N., Stuck F., Véronis J., Zaghouani W., « Evaluation of multilingual text alignment systems : the ARCADE II project », in *Proceedings of the 5th international Conference on Language Resources and Evaluation* (LREC06), Genoa, IT, 2006.

[CHO 57] Chomsky N., « Syntactic Structures », Mouton, La Haye, 1957.

[CHO 00] Choueka Y., Conley E.S., Dagan I., « A comprehensive bilingual word alignment system. Application to disparate languages: Hebrew and English », in Véronis J. (ed.) *Parallel Text Processing*, Dordrecht, Kluwer Academic Publishers, p. 69-96, 2000.

[CLE 04] Clément L., Sagot B., Lang B., « Morphology based automatic acquisition of large-coverage lexica », in *Proceedings of 4th International Conference on Language Ressources and Evaluation* (LREC04), Lisbon, PT, 2004.

[CME 04] Čmejrek M., Cuřín J., Havelka J., « Prague Czech-English Dependency Treebank. Any hopes for a common annotation scheme », in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, Boston, Massachusetts, USA, 2004.

[CRA 03] Crabbé B., Gaiffe B. et Roussanaly A., « Une plate-forme de conception et d'exploitation d'une grammaire d'arbres adjoints lexicalisés », in *Actes de la Conférence francophone internationale sur le Traitement Automatique des Langues Naturelles (TALN 03)*, Batz-sur-mer, FR, 2003.

[CKIP 93] CKIP, « Analysis of Syntactic Categories for Chinese », CKIP Technical Report #93-05, Institute of Information Science, Taipei, 1993.

[DAI 94] Daille B., « Approche mixte pour l'extraction de terminologie : statistique lexicale et filtres linguistiques », *Thèse de doctorat en informatique*, Université de Paris VII, Paris, FR, 1994.

[DIE 99] Diệp Quang Ban, Hoàng Văn Thung, « Ngữ pháp tiếng Việt (Vietnamese Grammar, vol. 1-2) », NXB Giáo dục, Hanoi, 1999.

[DIN 01] Dinh D., Hoang K., Nguyen V. T., « Vietnamese Word Segmentation », in *Proceedings of the Sixth Natural Language Processing Pacific Rim Symposium* (NLPRS2001), Tokyo, JP, 2001.

[ELB 95] El-Bèze M, Spriet T., « Etiquetage probabiliste et contraintes syntaxiques », in *Actes de la Conférence francophone internationale sur le Traitement Automatique des Langues Naturelles* (TALN95), Marseille, FR, 1995.

[ERJ 96] Erjavec T., Ide N., Petkevic V., Véronis J., « Multext-East: Multilingual Text, Tools and Corpora for Central and Eastern European Languages », in *Proceedings of the First TELRI European Seminar*, p. 87-98, 1996.

[FAR 02] Farrar S., Lewis W. D., Langendoen D. T., « An Ontology for Linguistic Annotation », AAAI '02 Workshop: *Semantic Web Meets Language Resources*, Menlo Park, CA, 2002.

[FEN 04] Feng J., Hui L., Yuquan C., Ruzhan L., « An Enhanced Model for Chinese Word Segmentation and Part-Of-Speech Tagging », *SIGHAN Workshop*, ACL2004, Barcelona, SP, 2004.

[FIL 04] Fillmore C. J., Baker C. F., Sato H., « FrameNet as a "Net" », in *Proceedings of 4th International Conference on Language Ressources and Evaluation* (LREC04), Lisbon, PT, 2004.

[FRA 03] Francopoulo G., « Proposition de norme des Lexiques pour le traitement automatique du langage », CN RNIL N 7, 25 novembre 2003. http://pauillac.inria.fr/atoll/RNIL

[FUN 97] Fung P., McKeown K. R., « A technical word and term translation aid using noisy parallel corpora across language groups ». *Machine translation*, 12 (1/2), 1997.

[GAL 91] Gale W.A., Church K.W., « A program for aligning sentences in bilingual corpora », in *Proceedings ACL 1991*, Berkeley, 1991.

[GAL 93] Gale W.A., Church K.W., « A program for aligning sentences in bilingual corpora », *Computational Linguistics,* 19(3), p. 75-102, 1993.

[GAL 95] Gale W.A., Sampson, G., « Good-Turing frequency estimation without tears », *Journal of Quantitative Linguistics*, 2(3), p. 217-237, 1995.

[GAR 94] Garside R., Hutchinson J., Leech G.N., McEner A.M., Oakes M.P., « The exploitation of parallel corpora in projects ET 10/63 and CRATER », in Jones D. (ed.), *New Methods in Language Processing*, p. 108-115, UMIST, 1994.

[GAZ 85] Gazdar G., Klein E., Pullum G., Sag I., « Generalized Phrase Structure Grammar », Havard University Press, 1985.

[GOL 91] Goldfarb C., « The SGML Handbook », Oxford University Press, 1991.

[GRO 75] Gross M., « Méthode en syntaxe », Paris, Hermann, 1975.

[HAL 03] Ha L. A., « A method for word segmentation in Vietnamese », in *Proceedings of Corpus Linguistics*, Lancaster, UK, 2003.

[HAR 51] Harris Z., « Structural Linguistics », Chicago University Press, Chicago, 1951.

[HAU 53] Haudricourt A.G., « La place du vietnamien dans les langues austroasiatiques », *Bulletin de la Société de Linguistique de Paris*, 49, 1, 1953.

[HAU 54] Haudricourt A.G., « De l'origine des tons en vietnamiens », *Journal Asiatique*, 242,1, 1954.

[HOA 02] Hoàng Phê, « Từ điển tiếng Việt (Vietnamese Dictionary) », Vietnam Lexicography Centre, NXB Đà Nẵng, 2002.

[HUU 98] Hữu Đạt, Trần Trí Dõi, Đào Thanh Lan, « Cơ sở tiếng Việt (Basis of Vietnamese) », NXB Giáo dục, Hanoi, 1998.

[IDE 93] Ide N., Le Maitre J., Véronis J., « Outline of a model or lexical databases », *Information Processing and Management,* 1993.

[IDE 94] Ide N., Véronis J., « MULTEXT: Multilingual Text Tools and Corpora », in *Proceedings of the 15th International Conference on Computational Linguistics*, COLING'94, Kyoto, JP, 1994.

[IDE 95a] Ide N., Sperberg-McQueen C.M., « The TEI: History, Goals, and Future », in Ide N., Véronis J. (ed.) *Text Encoding Inititative: Background and context*, Kluwer Academic Publishers, Dordrecht, 1995.

[IDE 95b] Ide N., Véronis J., « Encoding dictionaries », in Ide N., Véronis J. (ed.) *Text Encoding Inititative: Background and context*, Kluwer Academic Publishers, Dordrecht, 1995.

[IDE 00a] Ide N., Kilgarriff A., Romary L., « A Formal Model of Dictionary Structure and Content », in *Proceedings of Euralex 2000,* Stuttgart, 113-126, 2000.

[IDE 00b] Ide, N., Bonhomme, P., Romary, L.. « XCES: An XML-based Standard for Linguistic Corpora », in *Proceedings of the Second Language Resources and Evaluation Conference (LREC),* Athens, GR, 2000.

[IDE 01a] Ide N., Macleod C., « The American National Corpus: A Standardized Resource of American English », in *Proceedings of Corpus Linguistics 2001,* Lancaster UK, 2001.

[IDE 01b] Ide N., Romary L., « Standards for Language Resources », *Proceedings of the IRCS Workshop on Linguistic Databases*, Philadelphia, p. 141-149, 2001.

[IDE 03] Ide N., Romary L., « Encoding Syntactic Annotation », in Abeillé A. (ed.) *Treebank - Building and Using Parsed Corpora*, Dordrecht, Kluwer Academic Publishers, p. 281-296, 2003.

[IDE 04] Ide N., Romary L., « A registry of Standard Data Categories for Linguistic Annotation », in *Proceedings of 4th International Conference on Language Ressources and Evaluation* (LREC04), Lisbon, PT, 2004.

[ISA 93] Isabelle P., Dymetman M., Foster G., Justras J-M., Macklovitch E., Perrault F., Ren X., Simard M., « Translation analysis and translation automation », in *Proceedings of the Fifth International Conference on Theoretical and Methodological Issues in Machine Translation* (TMI'93), Kyoto, Japan, 1993.

[ISO 05a] ISO/CD 24611, « Language Resource Management - Morpho-syntactic Annotation Framework », ISO TC 37 / SC 4 N225, 25th Oct 2005. *http://tc37sc4.org/documents*

[ISO 05b] ISO 24613, « Language resource management - Lexical markup framework (LMF) », ISO Geneva 2005.

[ISO 06] ISO 24610-1, « Language Resource Management - Feature Structures - Part 1: Feature Structure Representation », ISO Geneva 2006.

[JAR 03] Järniven T., « Bank of English and beyond », in Abeillé A. (ed.) *Treebank - Building and Using Parsed Corpora*, Dordrecht, Kluwer Academic Publishers, p. 43-59, 2003.

[JOS 87] Joshi A., « Introduction to Tree Adjoining Grammar », in Manaster-Ramer A. (ed.), *The Mathematics of Language*, J. Benjamins, 1987.

[KAW 95] Kawtrakul A. et.al, « A Lexibase Model for Writing Production Assistant System », *The 2nd Symposium on Natural Language Processing*, Bankok, TH, August 2-4, 1995.

[KAW 02] Kawtrakul A., Suktarachan M., Varasai P., Chanlekha H., « A state of the art of Thai language resources and Thai language behavior analysis and modeling », in *Proceedings of the ACL-02 - Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, University of Pennsylvania, USA, 2002.

[KAY 79] Kay M., « Functional Grammars », in *Proceedings of 5th annual meeting of the Berkeeley Linguistics Society*, p. 142-158, Berkeley, 1979.

[KAY 85] Kay M., « Parsing in functional unification grammar », in Dowty D.R., Karttunen L., Zwicky A.M. (eds.), *Natural Language Parsing*, Cambridge University Press, 1985.

[KAY 93] Kay M., Röscheisen M., « Text-translation alignment », *Computational Linguistics,* 19(1), p.121-142, 1993.

[KUC 67] Kucera H., Francis W.N., « Computational Analysis of Present-Day American English », Providence, Brown University Press, 1967.

[KUI 92] Kuipec J., « Robust Part-of-Speech Tagging Using a Hidden Markov Model », *Computer Speech and Language,* vol. 6, p. 225-242, 1992.

[LAP 00] Laporte E., « Mots et niveau lexical », in Pierrel J-M. (ed.) *Ingénierie des Langues,* Hermes Science Europe, 2000.

[LAU 94] Lauriston A., « Automatic Recognition of Complex terms: Problems and the TERMINO Solution », *Terminology*, 1(1):147–170, 1994.

[LEE 94] Leech G., Garside R., Bryant M., « CLAWS4: The tagging of the British National Corpus», in *Proceedings of The International Conference on Computational Linguistics COLING 94*, p. 622-628, 1994.

[LEH 96] Lehmann S., Estival D., Oepen S., « TSNLP – Des jeux de phrases-test pour l'évaluation d'applications dans le domaine du TALN », in *Actes de la Conférence francophone internationale sur le Traitement Automatique des Langues Naturelles (TALN 96)*, Marseilles, FR, 1996.

[LEH 05] Le H. P., « Vers une grammaire électronique pour le vietnamien », Rapport pour l'obtention du DEPA, Institut pour la francophonie de l'Informatique à Hanoi (IFI), 2005.

[LEH 06] Le H. P., Nguyen T. M. H., Romary L., Roussanaly A., « A Lexicalized Tree Adjoining Grammar for Vietnamese », in *Proceedings of the 5th international Conference on Language Resources and Evaluation* (LREC06), Genoa, IT, 2006.

[LEV 95] Levinger M., Ornan U., Itai A., « Learning morpho-lexical probabilities from an untagged corpus with an application to Hebrew », *Computational Linguistics,* 21(3), p. 383-404, 1995.

[LIT 76] Li Charles N., Thompson Sandra A., « Subject and Topic: A new Typology of Language », in Charles N. Li (ed.) *Subject and Topic,* London/New York: Academic Press, p. 457-489, 1976.

[MAC 93] MacMahon J.G., Smith F.J., « Improving statistical language model performance with automatically generated word hierarchies », *Computational Linguistics*, 19(2), p. 313-330, 1993.

[MAN 99] Manning C.D., Schütze H., « Foundations of Statistical Natural Language Processing », MIT Press, Cambridge MA, 1999.

[MAN 03] Mangeot M., Sérasset G., Lafourcade M., « Construction collaborative de données lexicales multilingues, le projet Papillon », in Zock M. et Carroll J. (ed.), *Revue TAL (Traitement Automatique des Langues), édition spéciale, Les dictionnaires électroniques : pour les personnes, les machines ou pour les deux ?*, Vol. 44, pp. 151-176, 2003.

[MAR 93] Marcus M.P., Santorini B., Marcinkievicz M.A., « Building a large annotated corpus of English: the Penn TreeBank », *Computational Linguistics*, 19(2), p. 313-330, 1993.

[MAR 03] Marciniak M., Mykowiecka A., Przepiosrkowski A., « An HPSG-annotated test suite for Polish », in Abeillé A. (ed.) *Treebank - Building and Using Parsed Corpora*, Dordrecht, Kluwer Academic Publishers, 2003.

[MAS 12] Maspero H., « Étude sur la phonétique historique de la langue anamite », in *Bulletin de l'Ecole Française d'Extrême-Orient*, 12, p. 1-123, 1912.

[MAS 98] Mason O., Tufiş D., « Tagging Romanian Texts: a Case Study for QTAG, a Language Independent Probabilistic Tagger », in *Proceedings of the first International Conference on Language Ressources and Evaluation* (LREC98), Granada, SP, p. 589-596, 1998.

[MEL 84] Mel'cuk I., Arbatchewsky-Jumarie N., Eltnisky L., Iordanskaja L., Lessard A., « DEC : Dictionnaire explicatif et combinatoire du français contemporain, recherches lexico-sémantiques I », Presses de l'université de Montréal, Montréal (Quebec), CA, 1984.

[MEL 88] Mel'cuk I., Arbatchewsky-Jumarie N., Dagenais L., Elnitsky L., Iordanskaja L., Lefebvre M.-N., Mantha S., « DEC : Dictionnaire explicatif et combinatoire du français contemporain, recherches lexico-sémantiques II », Presses de l'université de Montréal, Montréal (Quebec), CA, 1988.

[MER 94] Merialdo B., « Tagging English text with a probabilistic model », *Computational Linguistics*, 20(2), p. 155-172, 1994.

[MIL 90a] Miller P., Torris T. (ed.), « Formalismes syntaxiques pour le traitement automatique du langage naturel », Hermes Science Europe , 1990.

[MIL 90b] Miller G., Backwith R., Fellbaum C., Gross D., Miller K., « Five papers on WordNet », Technical report, Cognitive science laboratory, Princeton University, July, 1990.

[NGU 98a] Nguyễn Minh Thuyết, Nguyễn Văn Hiệp, "Thành phần câu tiếng Việt (The components of sentence in Vietnamese)", NXB Đại học Quốc gia, Hanoi, 1998.

[NGU 98b] Nguyễn Tài Cẩn, « Ngữ pháp tiếng Việt (Vietnamese Grammar) », NXB Đại học Quốc gia, Hanoi, 1998.

[NGU 99] Nguyen T. M. H., « Alignement étendu de textes parallèles », Mémoire DEA Informatique, Université Henri Poincaré - Nancy I, 1999.

[NGU 03a] Nguyen Thi Minh Huyen, Le Hong Phuong, Vu Xuan Luong, « A case study of the probabilistic tagger QTAG for Tagging Vietnamese Texts », in *Proceedings of ICT.rda'03* (The First National Symposium on Research, Development and Application of Information and Communication Technology), Hanoi, VN, 2003.

[NGU 03b] Nguyen T. M. H., Romary L., Vu X. L., « Une étude de cas pour l'étiquetage morphosyntaxique de textes vietnamiens », in *Actes de la Conférence francophone internationale sur le Traitement Automatique des Langues Naturelles* (TALN 03), Batz-sur-mer, FR, 2003.

[NGU 04a] Nguyen T. B., Nguyen T. M. H., Romary L., Vu X. L., « Lexical descriptions for Vietnamese language processing », in *Proceedings of the Asian Language Resources Workshop*, IJC-NLP 2004, Hainan, CN, 2004.

[NGU 04b] Nguyen T. B., « Plate-forme d'étiquetage morphosyntaxique pour le vietnamien », Rapport pour l'obtention du DEPA, Institut pour la francophonie de l'Informatique à Hanoi (IFI), 2004.

[PAR 00] Paroubek P., Rajman M., « Etiquetage morphosyntaxique », in Pierrel J-M. (ed.) *Ingénierie des Langues,* Hermes Science Europe , 2000.

[POL 87] Pollard C., Sag I., « Information-based Syntax and Semantics », CSLI series, University of Chicago Press, 1987.

[POL 94] Pollard C., Sag I., « Head-driven Phrase Structure Grammar », CSLI series, University of Chicago Press, 1994.

[PRZ 03] Przepiórkowski A., Woliński M., « The Unbearable Lightness of Tagging - A Case Study in Morphosyntactic Tagging of Polish », in *Proceedings of 4th International Workshop on Linguistically Interpreted Corpora* (LINC-03), Budapest, HG, 2003.

[ROM 00a] Romary L., Bonhomme P., « Parallel alignment of structured documents », in Véronis J. (ed.) *Parallel Text Processing,* Dordrecht, Kluwer, p. 201-217, 2000.

[ROM 00b] Romary L., « Outils d'accès à des ressources linguistiques », in Pierrel J-M. (ed.) *Ingénierie des Langues,* Hermes Science Europe, 2000.

[ROM 04] Romary L. Salmon-Alt S., Francopoulo G., « Standards going concrete: from LMF to Morphalou », Workshop *Enhancing and using electronic dictionaries*, The 20th International Conference on Computational Linguistics (COLING), Geneva, CH, 2004.

[SAL 04] Salmon-Alt S., Bick E., Romary L., Pierrel J.-M., « La FReeBank : vers une base libre de corpus annotés », in *Actes de la Conférence francophone internationale sur le Traitement Automatique des Langues Naturelles* (TALN 04), Fez, Maroc, 2004.

[SAL 05] Salmon-Alt S., Romary L., « Tutoriel : Lexical Markup Framework : Principes fondateurs et application aux lexiques syntaxiques », Journée ATALA « *Interface lexique-grammaire et lexiques syntaxiques et sémantiques* », 12 mars 2005.

[SAM 97] Samuelsson C., Voutilainen A., « Comparing a linguistic and a stochastic tagger », ACL-EACL'97, Madrid, SP, 1997.

[SCH 94a] Schmid H., « Part-of-Speech Tagging with Neural networks », *International Conference on Computational Linguistics*, Kyoto, JP, p. 172-176, 1994.

[SCH 94b] Schmid H., « Probabilistic part-of-speech tagging using decision trees », *International Conference on New Methods in Language Processing*, Manchester, UK, 1994.

[SER 93] Sérasset G., « Recent Trends of Electronic Dictionary Research and Development in Europe », *Technical Memorandum Electronic Dictionary Research* (EDR), Tokyo, JP, 1993.

[SHA 88] Shabes Y., Abeillé A., Joshi A., « Parsing strategies with lexicalized grammars : Tree Adjoining Grammars », in *Proceedings of 12th International Conference on Computational Linguistics* (COLING 1988), Budapest, 1988.

[SHI 86] Shieber S., « An introduction to unification-based theories of grammar », CSLI, University o Chicago Press, 1986. [Trad. Fr. in [MIL 90a], chapitre 1]

[SHI 90] Shieber S., Schabes Y., « Synchronous Tree Adjoining Grammars », in *Proceedings of 13th International Conference on Computational Linguistics* (COLING 1990), Helsinki, FI, 1990.

[SIM 98] Simard M., « The BAF : a corpus of English-French bitext », *Proceedings of First International Conference on Language Resources and Evaluation* (LREC), Granada, SP, 1998.

[SIN 00] Singh S., McEnery T., Baker P., « Building a parallel corpus of English/Panjabi », in Véronis J. (ed.) *Parallel Text Processing,* Dordrecht, Kluwer, p. 335-346, 2000.

[SOR 00] Sornlertlamvanich V., Potipiti T., Charoenporn T., « Automatic Corpus-Based Thai Word Extraction with the C4.5 Learning Algorithm », in *Proceedings of International Conference on Computational Linguistics* (COLING 2000), Saarbrücken, DE, 2000.

[SPE 94] Sperberg-McQueen C.M., Burnard L., « Guidelines or Electronic Text Encoding and Interchange », Chicago and Oxford: *Text Encoding Initative*, 1994.

[SPR 96] Sproat R., Shi C., Gale W., Chang N., « A stochastic finite-state word-segmentation algorithm for Chinese », *Computational Linguistics,* 22(3), p. 377-404, 1996.

[SUN 98] Sun M., Shen D., Tsou B. K., « Chinese word segmentation without using lexicon and hand-crafted training data », in *Proceedings of COLING-ACL 98*, Montreal, Quebec, CA, 1998.

[TAK 06] Takenobu T., Sornlertlamvanich V., Charoenporn T., Calzolari N., Monachini M., Soria C., Huang C.-R., YingJu X., Hao Y., Prévot L., Kiyoaki S., « Infrastructure for standardization of Asian language resources », in *Proceedings of COLING-ACL 2006*, Sydney, AU, 2006 (to appear).

[TUF 99] Tufiş D., « Tiered Tagging and combined classifier », in Jelineck F. and Nörth E. (ed.) *Text, Speech and Dialogue, Lecture Notes in Artificial Intelligence 1692,* Springer, 1999.

[UYB 83] Uỷ ban Khoa học Xã hội Việt Nam (Comité de Sciences Sociales du Vietnam), « Ngữ pháp tiếng Việt (Vietnamese Grammar) », NXB Khoa học Xã hội, Hanoi, 1983.

[VAI 03] Vaillant P., « Une grammaire formelle du créole martiniquais pour la génération automatique », in *Actes de la Conférence francophone internationale sur le Traitement Automatique des Langues Naturelles* (TALN 03), Batz-sur-mer, FR, 2003.

[VER 92] Véronis J., Ide N., « A feature-based model for lexical databases », in *Proceedings of the 14th International Conference on Computational Linguistics, COLING'92,* Nantes, FR, p. 588-594, 1992.

[VER 98] Vergnes J., Giguet E., « Regards théoriques sur le tagging », in *Actes de la Conférence francophone internationale sur le Traitement Automatique des Langues Naturelles* (TALN98), Paris*,* 1998.

[VER 00] Véronis J. (ed.), « Parallel Text Processing », Dordrecht, Kluwer, 2000.

[VER 00a] Véronis J., Langlais Ph., « Evaluation of parallel text alignment systems: ARCADE », in Véronis J. (ed.) *Parallel Text Processing,* Dordrecht, Kluwer, p. 369-388, 2000.

[VER 00b] Véronis J., « Alignement de corpus multilingues », in Pierrel J-M. (ed.) *Ingénierie des langues,*Hermes Science Europe, p.151-171, 2000.

[VIL 03] Villemonte de la Clergerie E., Rajman M., « Petit panorama des approches en analyse syntaxique », in Villemonte de la Clergerie E., Rajman M. (ed.), *Revue TAL (Traitement Automatique des Langues), Evolution en analyse syntaxique*, Vol. 44, p. 7-14, 2003.

[VIL 04] Vilnat A., Monceaux L., Paroubek P., Robba I., Gendner V., Illouz G., Jardino M., « Annoter en constituants pour évaluer des analyseurs syntaxiques », in *Actes de la Conférence francophone internationale sur le Traitement Automatique des Langues Naturelles* (TALN 04), Fez, Maroc, 2004.

[WAN 03] Wang H., Yu S., « The Semantic Knowledge-base of contemporary Chinese and its Applications in WSD », *the Second SIGHAN Workshop on Chinese Language Processing*, ACL03, Sapporo, JP, 2003.

[WON 96] Wong P., Chan C., « Chinese Word Segmentation based on Maximum Matching and Word Binding Force », in *Proceedings of the 16th conference on Computational linguistics*, Copenhagen, DK, 1996.

[YOS 03] Yoshinaga N., Miyao Y., Torisawa K., Tsujii J., « Parsing comparison across grammar formalisms using strongly equivalent grammars. Comparison of LTAG and HPSG parsers: A case study », in Villemonte de la Clergerie E., Rajman M. (ed.), *Revue TAL (Traitement Automatique des Langues), Evolution en analyse syntaxique*, Vol. 44, p. 15-39, 2003.

Résumé

Le travail présenté dans ce mémoire porte sur la construction des outils et ressources linguistiques pour les tâches fondamentales de traitement automatique de la langue vietnamienne, dans un contexte monolingue ainsi que multilingue. Nous présentons pour cette langue encore peu étudiée des solutions possibles aux problèmes d'annotation morpho-syntaxique (définition de descripteurs lexicaux « de référence », construction d'un lexique avec ces descriptions, des outils de segmentation et d'étiquetage lexical), d'analyse syntaxique (première tentative de modélisation de la grammaire vietnamienne en employant le formalisme TAG, cadre de construction de ressources pour l'analyse syntaxique) et d'alignement multilingue (constitution d'un corpus multilingue, développement d'un système d'alignement multilingue). Afin d'assurer la réutilisabilité des travaux réalisés, et dans l'espoir de les voir stimuler le développement du TAL au Vietnam, nous avons apporté une attention particulière aux questions de normalisation de la gestion des ressources linguistiques.

Mots clés

alignement multilingue, analyse syntaxique, annotation linguistique, corpus annotés, étiquetage lexical / morphosyntaxique, grammaire d'arbres adjoints, lexique, normalisation, partie du discours, ressources linguistiques, segmentation, traitement automatique des langues, vietnamien.

Abstract

The work presented in this document deals with the constitution of linguistic resources and tools for the fundamental tasks of automatic processing of the Vietnamese language, both in monolingual and multilingual contexts. We present possible solutions to the problems of morpho-syntactic annotation (definition of "standardized" lexical descriptors, development of a lexicon with these descriptors, and the tools for word segmentation and part-of-speech tagging), syntactic analysis (first tentative to model the Vietnamese grammar using the TAG formalism, framework to build the language resources needed for parsing), and multilingual alignment (constitution of a multilingual corpus, development of a system for the alignment of multilingual texts). In order to ensure the reusability and extendibility of the built linguistic resources, we have paid a particular attention to the questions of standardization of language resource management.

Keywords

annotated corpus, language resources, lexicon, linguistic annotation, multilingual alignment, natural language processing, parsing, part-of-speech, POS tagging, segmentation, standardization, tree adjoining grammar, Vietnamese.