Vers des moteurs de recherche "intelligents" : un outil de détection automatique de thèmes. Méthode basée sur l'identification automatique des chaînes de référence

Laurence Longo

UNIVERSITÉ DE STRASBOURG

ÉCOLE DOCTORALE des Humanités **(ED 520)**

UR 1339 LiLPa

THÈSE présentée par :

Laurence LONGO

soutenue le : **12 décembre 2013**

pour obtenir le grade de : **Docteur de l'université de Strasbourg**

Discipline/ Spécialité : Sciences du Langage/ Linguistique et informatique

Vers des moteurs de recherche « intelligents » : un outil de détection automatique de thèmes

Méthode basée sur l'identification automatique des chaînes de référence

THÈSE co-dirigée par :

Mme Catherine Schnedecker	Professeur, Université de Strasbourg
Mme Amalia Todirascu	MCF, Université de Strasbourg

RAPPORTEURS :

M. Yves Bestgen	Professeur, Université catholique de Louvain
M. Denis Maurel	Professeur, Université François Rabelais

AUTRES MEMBRES DU JURY :

Mme Agnès Tutin	Professeur, Université Stendhal Grenoble
M. Frédéric Landragin	Chargé de recherche, CNRS – UMR Lattice

UNIVERSITE DE STRABOURG
Ecole Doctorale des Humanités

LiLPa – Linguistique, Langues, Parole
Fonctionnements Discursifs & Traduction

Vers des moteurs de recherche « intelligents » : un outil de détection automatique de thèmes

Méthode basée sur l'identification automatique
des chaînes de référence

THESE DE DOCTORAT

Discipline : Sciences du Langage

Spécialité : Linguistique et Informatique

Soutenue le 12 décembre 2013
par

Laurence LONGO

JURY :

Mme Catherine SCHNEDECKER	Directrice	(Université de Strasbourg – LiLPa)
Mme Amalia TODIRASCU	Co-directrice	(Université de Strasbourg – LiLPa)
M. Yves BESTGEN	Rapporteur	(Université Catholique de Louvain – F.R.S.-FNRS)
M. Denis MAUREL	Rapporteur	(Université François Rabelais – Tours)
Mme Agnès TUTIN	Examinatrice	(Université Grenoble 3 – LIDILEM)
M. Frédéric LANDRAGIN	Examinateur	(CNRS – UMR Lattice)

Thèse réalisée dans le cadre d'une convention CIFRE (Convention Industrielle de Formation par la REcherche) dans la société RBS (Ready Business System), Strasbourg.

UNIVERSITE DE STRABOURG
Ecole doctorale des Humanités

LiLPa – Linguistique, Langues, Parole
Fonctionnements Discursifs & Traduction

Vers des moteurs de recherche « intelligents » : un outil de détection automatique de thèmes

Méthode basée sur l'identification automatique des chaînes de référence

THESE DE DOCTORAT

Discipline : Sciences du Langage

Spécialité : Linguistique et Informatique

Soutenue le 12 décembre 2013

par

Laurence LONGO

JURY :

Mme Catherine SCHNEDECKER	Directrice	(Université de Strasbourg – LiLPa)
Mme Amalia TODIRASCU	Co-directrice	(Université de Strasbourg – LiLPa)
M. Yves BESTGEN	Rapporteur	(Université Catholique de Louvain – F.R.S.-FNRS)
M. Denis MAUREL	Rapporteur	(Université François Rabelais – Tours)
Mme Agnès TUTIN	Examinatrice	(Université Grenoble 3 – LIDILEM)
M. Frédéric LANDRAGIN	Examinateur	(CNRS – UMR Lattice)

Thèse réalisée dans le cadre d'une convention CIFRE (Convention Industrielle de Formation par la REcherche) dans la société RBS (Ready Business System), Strasbourg.

Remerciements

La thèse est un travail solitaire, mais il n'aurait pu aboutir sans la présence et le soutien de nombreuses personnes que je tiens à remercier ici.

Je remercie chaleureusement Catherine Schnedecker pour avoir accepté, à mon grand honneur, de diriger « en cours de route » ma thèse (depuis juillet 2011). Je la remercie, entre autres, pour ses conseils, sa bienveillance, sa rigueur, son ouverture et son écoute sans faille, aussi bien en tant que directrice du laboratoire LiLPa qu'en tant que directrice de thèse. Les discussions et échanges que nous avons pu partager durant ces quelques années ont toujours été pour moi une énorme source de motivation et d'inspiration.

Je remercie Amalia Todirascu, ma co-directrice de thèse, pour m'avoir suivie régulièrement depuis le début de la thèse. A ses côtés, j'ai pu participer à de nombreuses conférences nationales et internationales, encadrer 3 stages de Master, effectuer de nombreuses heures de vacation en informatique à l'UFR, participer à des projets de recherche. Je la remercie pour toutes ces expériences enrichissantes et la confiance qui m'a été accordée.

Je remercie les membres du jury d'avoir accepté de donner de leur temps pour évaluer ce travail. Merci à Agnès Tutin, Yves Bestgen, Frédéric Landragin et Denis Maurel pour leurs conseils éclairés.

Je remercie vivement les membres de l'unité de recherche LiLPa pour avoir créé, au fil des séminaires, des réunions et des discussions pluridisciplinaires, un environnement propice à la curiosité et au désir de connaissances dans toutes les Sciences du Langage. Un grand merci à Beatrice Vaxelaire et Rudolph Sock qui m'ont soutenue dans les moments difficiles ainsi que pour leurs encouragements durant la dernière ligne droite de cette thèse.

Je remercie le président du directoire de la société RBS (Ready Business System) Daniel Romani et tout particulièrement mon responsable technique, Christian Dhinaut, pour m'avoir donné l'opportunité d'effectuer une thèse dans le cadre d'une convention CIFRE (Convention Industrielle de Formation par la REcherche) et ainsi bénéficier d'une expérience de trois ans en tant qu'ingénieur développement au sein de l'équipe R&D de RBS.

Je remercie Michel Charolles, Bernard Victorri et Frédéric Landragin pour m'avoir ouvert les portes du laboratoire Lattice (ENS-Paris 3) afin d'y suivre plusieurs séminaires et réunions. J'ai ainsi pu participer aux réflexions stimulantes et aux riches échanges autour de la coréférence lors des réunions mensuelles du groupe *Coref.* J'ai aussi pu prendre part au projet Peps *MC4* « Modélisation Contrastive et Computationnelle des Chaînes de Coréférence » qui a succédé au groupe *Coref* et qui a été le lieu, toujours dans une grande convivialité, de fructueuses réflexions autour de l'annotation et de la modélisation de la coréférence.

Je remercie les membres du Consortium Corpus écrit (IR-Corpus), groupe 8 « Annotations de plus haut niveau : syntaxe, sémantique, référence, annotations collaboratives », piloté par Amalia Todiraşcu et Agnès Tutin, avec qui nous avons pu échanger longuement sur les problèmes d'annotation de la coréférence et la constitution de corpus.

Je remercie les membres du projet Procope NHUMA, piloté par Catherine Schnedecker et Wiltrud Mihatch, avec qui nous avons partagé des journées dynamiques sur les noms d'humains. Toutes ces collaborations m'ont permis de bénéficier du savoir et des compétences généreusement offerts par chacun.

Je remercie mes anciens voisins de bureau chez RBS : Damien, Benjamin, Jonathan, Pascal. Merci pour votre soutien amical et votre curiosité scientifique.

Mes remerciements s'adressent bien évidemment aux autres doctorants, ex-doctorants et docteurs « LilPaliens » : Constanze, Angelina, Camille, Thomas, Lucie, Nourdine. La petite famille que nous avons constituée a donné naissance au premier colloque international jeunes chercheurs du LiLPa (CIJC 2012), en partenariat avec des doctorants de l'université de Bochum.

Je remercie amicalement les membres de l'association DoXtra (Association des doctorants et docteurs en Sciences Humaines de l'Université de Strasbourg) Stéphanie, Stéphane, Colette, Baba, pour les nombreux échanges transdisciplinaires autour de la thèse, pour leur motivation, leur dévouement et leur humanité. Je conserve des très bons souvenirs des déjeudis, des cours de relaxation, des pique-niques et des sorties culturelles que nous avons pu partager.

Je remercie Nathalie Hillenweck, directrice de l'UFR LSHA, Dominique Lauer et tout particulièrement Geneviève Hekpazo, une secrétaire hors pair dont l'aide administrative et l'écoute m'ont été précieuses durant mes deux années d'ATER au département d'informatique.

Je remercie Marie-Carmen Ramirez, secrétaire de l'école doctorale des Humanités, pour sa disponibilité et sa redoutable efficacité.

Je remercie Christophe, Daniéla, Julie, Delphine pour leurs conseils et leur soutien.

Je remercie enfin toutes les « heureuses rencontres » que j'ai pu faire au cours de mes participations à des conférences et autres manifestations scientifiques en France et à l'étranger : Baptiste, Lauréline, Matthias, Mathieu, Philippe, Hye Ran, Mai, Fanny, François, Marianne, Karen.

Ceux qui nous mènent à effectuer une thèse sont souvent des femmes et des hommes passionnés par la recherche. Je pense tout d'abord à mes professeurs que j'ai eu la chance de rencontrer durant mon cursus universitaire à Aix en Provence, qui m'ont initiée et « convertie » aux Sciences du Langage : Denis Autesserre, pour son cours fabuleux sur l'origine des langues, Claire Maury-Rouan pour son amour tourné vers la Langue des Signes Française, Véronique Rey pour ses cours de langues africaines et de neurolinguistique et enfin Christian Touratier qui m'a fait comprendre qu'il n'existait pas qu'un seul chemin menant à la thèse !

Je pense aussi à celles et ceux qui m'ont fait découvrir et apprécier l'informatique et le Traitement Automatique des Langues, que je n'ai plus pu quitter depuis : Corrinne Zaoui qui m'a donné mon premier cours de programmation, Nuria Gala qui m'a initiée à l'analyse syntaxique automatique et qui m'a accompagnée durant mes deux masters. Et une pensée émue à Jean Véronis, qui nous a malheureusement quittés le 8 septembre dernier, qui m'avait convaincue que « la linguistique ne pourrait bientôt plus être sans le TAL ».

Je remercie affectueusement mes parents qui m'ont toujours fait confiance et ont cru en moi, pour leur soutien indéfectible malgré l'éloignement géographique. Merci Maman d'avoir lu et relu mes chapitres, toujours avec autant d'intérêt et d'attention. Je remercie ma sœur Aurore, mon beau-frère Ludovic et mes deux neveux Clément et Quentin pour leur présence au bout du fil.

Je remercie tendrement mon compagnon Mathias qui a tout quitté en Provence pour me suivre dans cette merveilleuse aventure strasbourgeoise.

Je remercie enfin tous ceux avec qui j'ai pu partager cette expérience humaine entre Aix en Provence, Marseille, Grenoble, Montpellier, Toulouse, Paris, Amiens, Kehl et Strasbourg.

A ma mère.

« Il faut grouper les énonciations contenues dans chaque livre et les réduire à un certain nombre de chefs principaux, de façon à retrouver aisément toutes celles qui se rapportent au même objet. »

(Spinoza, 1670, 714-715, II, §5-10).

« Considérons un automate conçu pour lire un texte dans une langue naturelle donnée, l'interpréter, et enregistrer en quelque manière son contenu, par exemple pour être en mesure de répondre à des questions sur ce texte. Pour accomplir cette tâche, la machine devra remplir au minimum les exigences suivantes. Elle devra être en mesure de construire un fichier contenant la liste de toutes les entités, événements, objets etc... mentionnés dans le texte, et pour chaque entité enregistrer ce qui en est dit. »

(Karttunen, 1976), traduit par (Corblin, 1995a, 176).

« with the amount of textual data that is available and exponentially increasing there is a need to automatically process the same. One way of doing this is by topic identification, which is the process of assigning one or more labels to text. »

(Aery *et al.*, 2003 : 4).

Sommaire

Introduction générale

Enjeux

Même si l'arrivée des moyens informatiques a permis de résoudre le problème du stockage des données, le problème de l'exploitation d'un flot incessant d'informations demeure. En effet, les moyens actuels permettent de stocker des milliers de mégaoctets de données dans des espaces réduits (clés USB, disques durs, serveurs). Mais ces avancées technologiques provoquent d'autant plus de difficultés pour accéder rapidement à l'information pertinente (sur le Web ou sur l'intranet d'une organisation) qui réponde à un besoin précis, que les méthodes permettant d'organiser et de traiter les données entrantes n'ont pas suivi cette même évolution.

Qu'il s'agisse d'une recherche effectuée dans un cadre scolaire, professionnel ou personnel, tout utilisateur a déjà été confronté au problème d'accès à l'information et amené à se poser les questions suivantes : quels sont, parmi tous les résultats proposés par mon moteur de recherche, ceux qui répondent précisément à mon besoin ? Comment trouver rapidement l'information que je recherche sans avoir à passer en revue tous les résultats proposés ?

Force est de constater que, parmi la masse de résultats renvoyés à l'issue d'une requête, rares sont ceux qui contiennent les informations attendues et, paradoxalement, que certains documents pertinents ne sont pas retrouvés par les moteurs de recherche. Ce manque de pertinence est dû, entre autres, à la méthode d'indexation par mots-clés utilisée par les moteurs de recherche, qui extrait tous les documents contenant le ou les mots de la requête. Les propriétés linguistiques des textes (syntaxe, contenu, genre textuel) ne sont malheureusement pas prises en compte. Pourtant, les textes respectent des règles de morphologie, de grammaire et, au-delà des frontières d'une simple phrase, les règles générales de cohérence et de cohésion. De plus, l'exploitation des informations liées aux genres textuels est nécessaire car ceux-ci sont contraints par la situation de communication.

Même si les moteurs de recherche proposent des classifications de documents dans un domaine spécifique (*Google books* pour les livres ; *Google scholar* pour les articles scientifiques), ou des pages similaires (documents traitant des mêmes sujets que la page consultée), ces options de « recherche avancée » exploitent les balises des pages Web (les métadonnées indiquant les mots-clés, titre, description, auteur, sujet du document) pour la plupart et n'apportent que des solutions partielles aux problèmes. Les outils de Traitement Automatique de Langues (TAL), intégrés à des moteurs de

recherche et utilisés pour l'indexation et le traitement des requêtes, peuvent apporter de réelles solutions à ces problèmes et ainsi améliorer considérablement les résultats des moteurs de recherche.

Afin d'interpréter le contenu des documents et améliorer l'indexation dans les moteurs de recherche, les outils de TAL proposent plusieurs niveaux d'analyse automatique : morphologique, syntaxique et sémantique peu profonde et robuste (identification des groupes nominaux et des groupes prépositionnels). L'indexation peut alors s'effectuer à l'aide des lemmes (Namer, 1994), d'une analyse syntaxique et sémantique (Qristal, Intuition), de termes spécifiques au domaine ou de concepts (définis dans une ontologie). Néanmoins, la plupart de ces outils se restreignent à analyser le texte phrase par phrase. Or, l'information complète et pertinente que l'utilisateur recherche se trouve disséminée dans l'ensemble du texte. Pour retrouver cette information, les moteurs de recherche doivent faire appel aux outils d'analyse de discours efficaces qui prennent en compte la structure thématique des documents. La forme sous laquelle l'information est présentée dépend également du genre textuel.

L'extraction de l'information pertinente par des systèmes de TAL constitue une opération de premier plan dans la recherche d'information par détection de thèmes. La détection automatique des thèmes consiste à identifier les termes d'un texte qui indiquent son sujet, ses acteurs ou ses thèmes, par exemple « le réchauffement climatique », « Barack Obama », « Etat-membre », « la satisfaction des clients ». Ces termes, considérés comme représentatifs du contenu du document (Nomoto et Matsumoto, 1996), constituent des descripteurs qui permettent de retrouver rapidement les documents pertinents parmi une collection de documents (Salton *et al.*, 1993). A la différence de la catégorisation des textes (Lewis, 1992) qui assigne un thème (parmi une liste de thèmes arbitrairement définie par des humains) à un document, l'identification automatique des thèmes que nous adoptons extrait les thèmes présents explicitement dans les documents[1].

Apport

Cette recherche propose de mettre en relation des hypothèses théoriques (sur la référence, les genres textuels, les thèmes) et d'appliquer des techniques issues du TAL pour fournir un système de détection automatique de thèmes permettant d'améliorer la classification des documents dans les moteurs de recherche. La question de la détection des thèmes est abordée de manière pluridisciplinaire puisque cohabitent la

[1] L'identification automatique des thèmes que nous proposons, à base de peu de ressources, ne nous permet pas d'établir des inférences à partir des éléments thématiques retrouvés. En ce sens, nous ne pouvons proposer que des thèmes présents explicitement dans le texte comme descripteurs de document.

linguistique, la psycholinguistique, le TAL et l'informatique. Nous souhaitons utiliser l'élasticité du cadre théorique défini par la linguistique afin d'adapter les méthodes existantes en statistique et en informatique pour servir notre approche en TAL.

La méthode que nous proposons reposerait sur la détection automatique des thèmes dans les documents. Le texte serait alors considéré comme composé de segments homogènes thématiquement du point de vue de leur contenu et dotés d'une cohésion interne forte. Ces segments seraient aussi reliés entre eux car ils rendraient compte de différentes facettes à propos d'un sujet, d'un acteur, d'un produit (succession de thèmes). Par exemple, dans un portrait littéraire, plusieurs facettes d'un même personnage sont traitées tour à tour : son enfance, sa carrière, sa famille, etc. Ainsi, c'est en exploitant la structure textuelle que nous proposons d'identifier les thèmes centraux des documents.

Dans notre approche, nous exploitons la structure du document à travers ses marqueurs linguistiques (cadres de discours de (Charolles, 1997), chaînes lexicales, anaphores (Kleiber, 1994) et chaînes de référence (Cornish, 1995 ; Corblin, 1995a, 1995b ; Schnedecker, 1997)) pour détecter automatiquement les thèmes des documents. Nous exploitons aussi les informations issues du genre textuel du document car elles sont liées à la situation de communication. Ainsi, à l'instar de (Bestgen, 2012), nous sommes convaincue que, lorsque notre objectif est applicatif, nous devons adopter une approche pluridisciplinaire du discours (d'un point de vue de la linguistique, mais aussi d'un point de vue de la psycholinguistique et du TAL).

L'objectif de ce travail et sa visée applicative consistent en l'amélioration substantielle d'un moteur de recherche global[2] par l'ajout d'un outil de détection automatique de thèmes ATDS-Fr (Automatic Topic Detection System for French). Cet outil permettra d'aider l'utilisateur à identifier les thèmes centraux d'un discours à des fins de documentation (archivage, classification).

ATDS-Fr adopte une approche hybride statistique-linguistique pour découper les documents en segments thématiquement homogènes et identifier, par le biais de marqueurs linguistiques, les thèmes des documents. Cette méthode mixte statistique-symbolique répond aux préoccupations actuelles en TAL, formulées notamment lors de l'atelier MIXEUR « Méthodes mixtes pour l'analyse syntaxique et sémantique du français » (Retoré *et al.*, 2013) de la dernière conférence TALN.

Parmi les marqueurs linguistiques que nous avons choisi d'utiliser dans notre système, les chaînes de référence – suite d'expressions référentielles référant à la même entité du discours, par exemple « Le nouvel iPad Air… l'iPad Air… il… il… » – font l'objet

[2] Le moteur de recherche global (ou « plein texte ») est celui de l'entreprise RBS où nous avons effectué notre thèse en convention CIFRE.

d'une attention particulière, étant donné leur forte implication dans la signalisation des thèmes des documents. Or, d'un point de vue linguistique, les chaînes de référence n'ont été étudiées que par quelques auteurs (Charolles, 1987 ; Corblin, 1995a, 1995b ; Schnedecker, 1997, 2005) qui ont essentiellement travaillé sur des textes narratifs monoréférentiels (portraits journalistiques, résumés de films, nouvelles, extraits de roman). Afin de répondre aux besoins textuels industriels, nous proposons d'étendre l'étude des chaînes de référence à divers genres textuels (informatifs, argumentatifs) portant sur des référents humains et non humains. Cela nous permettra, par la même occasion, de déterminer les contraintes conditionnant la composition des chaînes de référence afin de constituer une typologie des chaînes de référence suivant le genre textuel. Dans le domaine du TAL, la résolution de la référence a été traitée de manière parcellaire par les systèmes symboliques développés jusqu'à présent : l'identification automatique des relations de coréférence se réduit souvent à l'identification des anaphores pronominales. Pour le français, il n'existe pas à notre connaissance de modèle opérationnel permettant d'identifier automatiquement les chaînes de référence dans les documents. Notre contribution vise à combler en partie ces manques.

De ce fait, le module d'identification automatique des chaînes de référence *RefGen* que nous avons conçu utilise des méthodes classiques pour identifier les expressions référentielles (noms propres, pronoms, groupes nominaux, etc.) et il prend en compte, dans son calcul de la référence, d'autres paramètres tels que le genre textuel du document. Ce module est le module central du système de détection automatique des thèmes des documents (ATDS-Fr).

Notre thèse s'est inscrite dans le cadre de divers projets de recherche et groupes de réflexion. De 2009 à 2012, notre projet s'est intégré aux réflexions du groupe de travail « chaînes de coréférence » (dans l'opération « Identification des Référents et Transitions Référentielles ») dirigé par F. Landragin, Laboratoire Lattice, ENS (UMR 8094) puis au projet Peps *MC4* « Modélisation Contrastive et Computationnelle des Chaînes de Coréférence »[3] qui lui a succédé. Les objectifs de ces projets ont été d'étudier, de modéliser et d'annoter les relations de coréférence d'entités humaines dans des corpus variés (résumés de films, nouvelles, romans). Aussi, courant 2011, notre étude des chaînes de référence et le développement de *RefGen* se sont intégrés au projet de recherche de l'unité de recherche LiLPa « ExtractChain »[4] porté par Amalia Todirascu. Dans ce cadre, notre contribution à ces projets relève de l'étude et la modélisation des chaînes de référence portant sur des référents humains et non humains (*i.e.* organisations, entités abstraites) dans des textes non narratifs.

[3] http://www.cnrs.fr/inshs/recherche/reference.htm
[4] http://lilpa.unistra.fr/uploads/media/cqr_projet_global_29082011.pdf

Organisation de la thèse

La thèse est organisée en trois parties[5], allant de la linguistique au TAL, afin de mener au développement du module de détection automatique de thèmes visé.

Dans la partie I, sont abordés les divers aspects linguistiques utilisés dans notre travail, à savoir les thèmes, les chaînes de référence et les genres textuels, que nous mettons en relation au cours des trois chapitres de la partie. Le premier chapitre prend pour source de réflexion le flou demeurant autour de la définition de la notion de *thème*. A partir de ce constat seront présentés les deux niveaux phrastique et textuel de cette notion, permettant de positionner notre approche globale des thèmes. Pour identifier les thèmes dans les documents, nous présenterons divers types d'indices de cohésion textuelle de continuité et de rupture thématique tels que les cadres de discours ou les chaînes de référence. Ces dernières feront l'objet du chapitre 2, où nous en préciserons la définition suivie (celle de C. Schnedecker) et où nous émettrons l'hypothèse qu'elles représentent des éléments linguistiques fiables pour participer à la détection des thèmes textuels. Le chapitre 3 sera l'occasion de montrer, par le biais de deux études de corpus (portant sur des textes juridiques, des rapports publics, des articles de presse, un roman), l'impact du genre textuel sur la composition des chaînes de référence et permettra de dresser une typologie des chaînes de référence suivant le genre. Cette typologie sera utilisée pour configurer notre outil d'identification des chaînes de référence suivant le genre du document, afin de cibler les types d'expressions référentielles à privilégier selon le genre d'occurrence du document.

La partie II traite des aspects automatiques pour la détection de thèmes. Le chapitre 4 présentera les différentes méthodes (statistiques, linguistiques et hybrides) d'analyse thématique automatique disponibles pour déterminer la structure thématique d'un document. Nous verrons que la plupart des systèmes effectuent de la segmentation thématique essentiellement statistique plutôt que de la détection de thèmes à proprement parler et qu'ils ne proposent donc pas, pour une majorité des cas, de descripteurs thématiques explicites associés aux segments thématiques délimités. Les méthodes hybrides, alliant des techniques statistiques à l'identification de marqueurs linguistiques explicites, se révèleront les plus adéquates pour atteindre notre objectif. Focalisé sur l'identification automatique de la référence textuelle, le chapitre 5 dressera un panorama des divers systèmes symboliques ou par apprentissage statistique. Comme nous le verrons, ces systèmes, quelle que soit la méthode utilisée, présentent des lacunes que nous exposerons, notamment la dépendance à de larges ressources annotées pour les systèmes par apprentissage ou l'identification parcellaire

[5] Par souci de lisibilité, la numérotation des notes de bas de page recommence à chaque chapitre.

des relations de coréférence pour les systèmes symboliques. Ces limites seront l'occasion de positionner notre approche et justifieront le développement du module d'identification automatique des chaînes de référence *RefGen*.

La partie III est consacrée au module de détection automatique de thèmes, ATDS-Fr. Dans le chapitre 6, nous décrirons les différentes composantes de l'architecture hybride d'ATDS-Fr et nous proposerons une application concrète de ce modèle afin de visualiser les thèmes et sous-thèmes obtenus par notre méthode. Dans le chapitre 7, nous nous focaliserons sur le développement du module central de notre méthode linguistique, le module *RefGen*. Nous détaillerons chacun de ses sous-modules : *RefAnnot*, pour l'annotation des diverses expressions référentielles et *CalcRef*, pour le calcul des chaînes de référence. Dans le chapitre 8, nous procèderons à l'évaluation des modules de *RefGen*, de manière manuelle et automatique, en utilisant les métriques d'évaluation actuelles de la coréférence. Nous montrerons qu'aucune de ces mesures ne permet de rendre compte des phénomènes que nous annotons et qu'une métrique spécifique serait à définir.

PARTIE I
Aspects linguistiques : thèmes, chaînes de référence et genres textuels

Pour pallier le constat que « jusqu'à présent, la linguistique n'a pas avancé scientifiquement au-delà de la phrase complexe » (Bakhtine, 1978 : 59), utiliser la structure du texte apparaît comme le moyen de fournir un meilleur accès au contenu des documents. En effet, la plupart des systèmes d'analyse sémantique se contentent d'effectuer des analyses locales du texte (proposition, phrase, phrase complexe) pour rechercher, extraire des informations ou constituer des ressources. Le texte est ainsi considéré uniquement comme une succession linéaire de propositions ou de phrases. Or, la structure du document constitue un indice essentiel qui permet de rendre compte du texte dans son ensemble : un tout organisé et cohérent (Charolles, 1995b).

Comme (Charolles, 1989 : 13) l'indique, « une fois franchies les limites de la phrase complexe, d'autres formes de structuration prennent le relais ». Le texte est caractérisé par un enchaînement en termes de thèmes (locaux ou globaux), rendant compte de son unité et il possède des spécificités liées à son genre, son registre, son type ou même son style (Biber et Conrad, 2009). Le texte est ainsi à considérer en tant qu'objet structuré dont l'organisation est signalée par des marques linguistiques (Péry-Woodley, 2001). La structure d'un document se révèle alors par la présence d'unités homogènes mises en relation par des marques de cohésion (connecteurs, cadres de discours, chaînes lexicales, anaphores, chaînes de référence).

L'identification des thèmes d'un texte demeure une tâche complexe aussi bien pour un humain que pour une machine (détection automatique). Pour (Grobet, 2002 : 10), la notion de thème « apparaît non comme quelque chose qui va toujours de soi, mais plutôt comme un processus impliquant un certain « travail » plus ou moins important, de la part des interprétants ».

Nous pensons que le thème n'est pas à considérer au singulier mais plutôt au pluriel car il n'existe vraisemblablement pas qu'un seul thème pour décrire une portion de texte ou un texte en son entier (*e.g.* subjectivité de l'à-propos). En effet, un même texte peut comporter un ou plusieurs thèmes suivant la granularité choisie par l'interprétant. Cette part de subjectivité dans l'identification des thèmes textuels suivant le point de vue de l'interprétant est notamment soulignée par (Brown et Yule, 1983). C'est peut-être cette part de subjectivité inhérente à la notion de thème qui serait la cause de la difficulté rencontrée pour définir et formaliser cette notion[1] (Guinaudeau, 2011).

Pour identifier les thèmes d'un texte, les interprétants utilisent des marques linguistiques et typo-dispositionnelles présentes dans le texte en plus de leur connaissance/expérience du monde. C'est en identifiant différents types de marqueurs linguistiques de cohésion et notamment les chaînes de référence qu'il est envisageable de détecter de manière automatique les thèmes des documents.

Dans cette première partie, nous présenterons la notion de *thème* et différentes théorisations dont elle a fait l'objet, du niveau phrastique au niveau textuel (chapitre 1). Comme nous le verrons dans le chapitre 2, les chaînes de référence représenteraient un type d'éléments linguistiques à prendre en compte pour détecter les thèmes des documents car elles constituent des indices d'introduction, de maintien et de rupture thématiques. Dans le chapitre 3, deux études des chaînes de référence dans des corpus variés montreront l'influence du genre textuel sur la composition des chaînes de référence.

[1] Comme nous le verrons dans le chapitre 1.

Chapitre 1

Thèmes

Dès l'Antiquité grecque, Platon et Aristote distinguent deux parties du discours : l'*onoma* (nom) et le *rhêma* (verbe ou prédicat). Dans ce cadre, le thème (*onoma*) constitue « ce dont on parle » tandis que le rhème (*rhêma*) correspond à « ce que l'on dit du thème ». Par exemple, pour la phrase :

> Dominique Perben plaide pour une liste unique de la majorité aux élections européennes. (corpus *Le Monde*)

on dit du thème « Dominique Perben » qu'il « plaide pour une liste unique aux élections européennes » (rhème).

La notion de *thème* se révèle être l'une des plus anciennes mais aussi l'une des plus complexes (Combettes, 1988) et ambiguës de la linguistique (Siouffi et Van Raemdonck, 1999). Son emploi quasi récurrent dans la réflexion linguistique en reflète l'importance. En plus des diversités terminologiques (*thème*, *topic*, *topique*, *saillance*, *focus* (ou *focus du discours* pour (Sidner, 1983)), *sujet*, *à propos*, *centre*, *base de l'énoncé* (Mathesius, 1942 ; Enkvist, 1978 : 170, note 2 ; Nølke, 1997 : 56-57)) auxquelles on doit se confronter[1], définir le thème est d'autant plus périlleux que la notion est à relier à une variété de niveaux d'analyse : de la phrase au texte, de la syntaxe à la pragmatique. De plus, ces différentes sensibilités théoriques se font souvent écho (Cadiot et Fradin, 1988). Ces inconsistances dans les définitions proposées ont pu être notamment relevées par (Goutsos, 1997).

Le thème est considéré comme le « point de départ » de l'énoncé (Daneš, 1974), l'élément le plus à gauche dans une phrase, l'à propos. La notion de *thème* n'est alors pas sans évoquer celle de sujet[2]. En effet, on trouve chez (Wilmet, 1986) une correspondance entre le sujet logique – ce dont on parle – et le thème. Chez ((Clark et Clark, 1977) cité par (Brown et Yule, 1983)), le terme *thème* est utilisé à la place de *sujet grammatical*. (Brown et Yule, 1983) soulignent que, même si l'élément le plus à gauche dans une phrase (*e.g.* le thème) coïncide souvent avec le sujet de la phrase, cela n'est pas le cas pour les phrases qui débutent par un adverbial (par exemple dans « *Sans hésiter*, Paul a pris le premier vol pour rejoindre Marie », l'élément le plus à gauche dans la phrase est un adverbial « sans hésiter » et pas le sujet « Paul »). Aussi, (Givón, 1983) pointe-t-il le problème généré par le fait d'assimiler le sujet au thème grammatical (Givón, 1976), qui obligerait à considérer plus d'un thème par proposition (on aurait par exemple deux thèmes « sans hésiter » et « Paul » dans l'exemple ci-dessus) qu'il faudrait alors hiérarchiser.

La notion de *thème* est traitée par des auteurs issus de divers cadres théoriques. Les définitions, variantes et nuances qui en découlent demeurent vastes. Il nous paraît donc utopique d'essayer d'en fournir un état de l'art exhaustif. Notre objectif est de cerner les principales conceptions du thème afin de parvenir à une définition stable pour notre étude. Dans la littérature, on trouve plusieurs comptes rendus et états de la question. Nous nous baserons entre autres sur les travaux de (Berthoud, 1996), (Brown et Yule, 1983), (Carter-Thomas, 2000,

[1] Pour (Garrod et Sandford, 1983) la notion de *thème* constitue un « champ de mines terminologique ».

[2] Au début du siècle dernier, les linguistes parlaient du *sujet psychologique* pour désigner le thème.

2009), (Combettes, 1977, 1983), (Galmiche, 1992), (Givón, 1979, 1983), (Goutsos, 1997), (Grobet, 2002), (Grobet et Montemayor-Borsinger, 2012), (Halliday, 1967a, 1967b), (Kintsch et van Dijk, 1978), (Marandin, 1988, 2003), (Porhiel, 2005a), (Prévost, 1998, 2001, 2003), (Touratier, 2010), (van Dijk, 1977a, 1977b).

Dans ce chapitre, après avoir fait état des problèmes définitoires touchant à la notion « multi-facettes » de *thème* (section 1), nous présenterons les différentes approches du thème de sa dimension phrastique à sa dimension textuelle (section 2). Nous aborderons aussi la notion de saillance, à la croisée de la phrase et du texte. En suivant notre objectif de modélisation informatique, nous préciserons l'approche des thèmes que nous adoptons dans notre projet. Dans une troisième section, nous ferons état de divers indices de cohésion disponibles pour identifier les ruptures et les continuités thématiques dans les documents, notamment les indices typodispositionnels, les chaînes de référence ainsi que les cadres de discours.

1 Problèmes définitoires

> « Que l'on prononce le mot « thème » au sein d'une assemblée de linguistes, il n'est pas sûr qu'il s'en trouvera deux pour s'accorder sur le sens de ce terme. » (Prévost, 1998 : 13)

D'après (Cadiot et Fradin, 1988), la notion de *thème* est une notion problématique car elle fonctionne de façon transverse. Elle conceptualise à la fois une position syntaxique, un rôle actanciel, une portion d'énoncés pourvue d'une fonction communicative faible (opposition *topic* vs *focus* de Daneš, 1968), un centre psychologique d'attention ou un lieu de cohérence, une condition de pertinence pour l'interprétation des énoncés. Sa définition reste en général intuitive (Rastier, 1996), vague et mystérieuse (Givón, 1983 : 5) : elle paraît résulter d'un savoir commun partagé permettant d'utiliser la notion sans être obligé de la définir explicitement. Ainsi, pour (Cadiot et Fradin, 1988), la notion de *thème* « n'est pas une notion conceptualisée : elle s'offre comme une donnée intuitive, primitive, antérieure à l'analyse. Il est symptomatique de ce point qu'elle est souvent introduite sans être définie ». (Kleiber, 1994 : 112) reprend ces remarques et parle du « caractère flou, intuitif, non conceptualisé de la notion de thème, topique ou encore discours ». De leur côté, (Brown et Yule, 1983 : 70) soulignent la situation paradoxale de cette notion : « Yet the basis for the identification of « topic » is rarely made explicit. In fact, « topic » could be described as the most frequently used, unexplained term in the analysis of discourse ». (Galmiche, 1992) en vient même à soulever un certain malaise lié à l'utilisation de cette notion :

> « on ne peut s'empêcher d'évoquer l'extrême inconfort que l'on éprouve devant une notion (un concept ?) aux dénominations multiples, aux caractérisations variées, souvent équivoques, voire contradictoires – car on peut difficilement parler de définitions – sans oublier quelques rares critères d'identification, peu fiables ou à la limite de la circularité. » ((Galmiche, 1992 : 3), cité par (Grobet, 2002 : 52)).

Nombreux sont, dans la littérature, les exemples fournis en substitut de définition claire et explicite et nombreuses sont les méthodes exposées sans avoir précisé au préalable la définition du thème suivie. Par exemple, (Beaver, 2004) définit le thème en terme de contraintes plutôt qu'en fournissant une réelle définition. (Nomoto et Matsumoto, 1996) considèrent que le titre du document ou bien seulement un nom du titre constitue le thème textuel. Certains comme (Reinhart,

1982) considèrent qu'est thème l'élément qui répond à une question ou un test comme « à propos de quoi parle le texte ? », « sur quoi porte le texte ? ».

(Kleiber, 1992) qualifie aussi de réductrice la notion de *thème*, qui n'est pas redéfinie suivant l'unité étudiée (phrase, discours). (Azzam *et al.*, 1998 : 1) soulignent enfin le problème de l'évaluation :

> « The term *focus*, along with its many relations such as theme, topic, center, etc., reflects an intuitive notion that utterances in discourse are 'about' something. This notion has been in accounts of numerous linguistic phenomena, but it has rarely been given a firm enough definition to allow its use to be evaluated. »

Face à une telle diversité dans l'emploi de cette « notion-caméléon » (Kleiber, 1992 : 15), certains linguistes se sont essayés au dur exercice consistant à établir un panorama de la situation. C'est ainsi que l'on peut trouver des articles dont le titre est relativement explicite : « Présentation, Une crise en thème ? » (Cadiot et Fradin, 1988), « la notion de thème : flou terminologique et conceptuel » (Prévost, 1998), « Au carrefour des malentendus : le thème » (Galmiche, 1992). D'ailleurs, ce dernier article souligne le problème central de l'hétérogénéité de la notion de *thème*, parce que « la plupart des analyses utilisant cette notion accordent une place trop grande à l'intuition » (Bosredon et Galmiche, 1992 : 2).

Pourtant, à l'issue de ces différentes typologies établies à propos de la notion de thème, la difficulté persiste encore pour définir cette notion (Siblot, 2000 : 33) :

> « Les auteurs qui ont cherché à caractériser la notion de *thème* (Cadiot, Fradin, 1988 ; Galmiche, 1992 ; Prévost, 1998) aboutissent au même constat paradoxal : le thème s'avère indispensable tout autant qu'indéfinissable. L'extrême variété des points de vue et les statuts différents que les diverses problématiques lui assignent paraissent même exclure la possibilité d'une appréhension cohérente. »

Il devient alors essentiel d'expliciter le sens dans lequel le terme est utilisé dans notre perspective de traitement automatique du langage. En ce sens, nous rejoignons (Prévost, 1998 : 34) :

> « Que les désaccords persistent ne nous paraît pas le plus grave : ils contribuent même à "fertiliser" ce concept. Il nous semble en revanche indispensable, lorsque l'on évoque le Thème, d'expliciter le sens que l'on donne à ce terme, sous peine de le voir se transformer en notion "fourre-tout" ».

Après ce bref état des problèmes définitoires touchant la notion de *thème*, passons aux différentes approches des thèmes, de la phrase au texte, avant de spécifier l'approche des thèmes que nous adoptons.

2 Du thème phrastique au thème textuel

Selon (van Dijk, 1977a ; Reinhart, 1981 ; Brown et Yule, 1983 ; Berthoud, 1996 ; Goutsos, 1997[3] ; Porhiel, 2005a) il est nécessaire de distinguer deux types de thèmes : les *thèmes phrastiques* pour référer à un segment particulier dans la phrase (niveau local) et les *thèmes textuels* pour marquer l'unité sémantique d'un texte (niveau global). (Tyrkkö, 2011 : 162) résume cette opposition de la manière suivante :

> « Sentence and discourse topic are thus two theoretically opposite notions: where sentence topic is essentially the answer to the question "what is this sentence about ? " and answerable in many cases with a simple word, a question about the topic of a discourse requires an answer which considers the whole information hierarchy of the text. »

Néanmoins, même si la notion de *thème* est utilisée dans deux domaines distincts, ces derniers ne demeurent pas étanches pour autant[4].

2.1 Thème phrastique

2.1.1 Fonctionnement

Le thème phrastique semble fonctionner systématiquement en couple dichotomique[5]. ((Galmiche, 1992 : 3), cité par (Flament, 2006 : 1), (Elalouf, 2006 : 8)) liste plusieurs couples de dénominations qui correspondent à des points de vue différents, allant de la sémantique à la logique, en passant par la grammaire, la psychologie, la pragmatique, etc. :

- thème — propos
- thème — prédicat
- thème — rhème

[3] (Goutsos, 1997) oppose le « what » (*i.e.* le thème phrastique) au « how » (*i.e.* le thème textuel).

[4] Par exemple, l'à propos (ou *aboutness*) est une conception qui touche à la fois le thème au niveau phrastique et au niveau textuel.

[5] Nous verrons dans la section 2.1.2 qu'il s'agira plutôt d'un continuum pour le thème considéré comme base du dynamisme communicatif.

- topic	comment
- (topique)	(commentaire)
- topic	focus
- présupposition	focus
- sujet psychologique	prédicat psychologique
- thème	commentaire[6]

Bien que lié à d'autres termes (*focus*, *rhème*, etc.), le thème n'en est pas mieux défini car les termes de *focus*, *commentaire*, *rhème*, sont eux-mêmes sujets à confusion (Porhiel, 1998). Pour décrire ces couples, (Galmiche, 1992 : 4) fait état d'une série de mots-clés relevés dans la littérature (la colonne de gauche regroupe les mots-clés du thème, la colonne de droite ceux du rhème) :

- ce dont on parle	ce qu'on en dit
- ancien	nouveau
- donné	non-donné
- connu	non-connu
- présupposé	focalisé
-	emphatique
- point de départ	but
- base	aboutissement
- statique	dynamique
- moins informatif	plus informatif
- support	apport
- fond	figure
- récupérable	non-récupérable
- prévisible	non-prévisible
- activé	
- saillant	
- centre d'intérêt	
- lié à la conscience immédiate	
- lié au contexte	indépendant du contexte
- lié aux circonstances	
- notoire	

Galmiche remarque que parfois les définitions des éléments dans la colonne de droite ne sont que des négations des éléments de la colonne de gauche (*e.g.* « prévisible » *vs* « non prévisible »). Néanmoins, les éléments situés dans la colonne de gauche peuvent correspondre avec ceux de la colonne de droite (*e.g.*

[6] Ce dernier couple que nous ajoutons à la liste proposée par Galmiche est donné par (Porhiel, 1998 : 14).

« plus informatif » et « saillant »), ce qui n'aide pas mieux la compréhension et la clarté des notions.

Considérés généralement comme mutuellement exclusifs (Bilhaut, 2006), thème et rhème (topique et focus) peuvent parfois se confondre chez (Dik et *al.*, 1981) ou bien jouer un rôle (plus ou moins important) dans la progression de l'information *via* une échelle du dynamisme communicatif[7] chez (Firbas, 1964).

2.1.2 Les principales conceptions du thème phrastique

(Porhiel, 1998) puis (Grobet, 2002), reprenant la classification proposée par (Schlobinski et Schütze-Coburn, 1992), dégagent quatre conceptions du thème : le thème comme *point de départ psychologique et/ou positionnel*, le thème comme « *ce dont il est question* », le thème comme *élément « connu »* et le thème comme *base du dynamisme communicatif*[8]. Nous reprenons chacune de ces conceptions ci-dessous.

– **Le thème comme *point de départ psychologique et/ou positionnel*** : cette conception à base psychologique est issue de la Perspective Fonctionnelle de la Phrase élaborée par l'Ecole de Prague. Dans cette approche, la structure informative de la phrase est vue selon une dualité thème/ rhème où le thème est le point de départ et la première partie, le reste de la phrase constitue la seconde partie :

> « Un examen […] minutieux des phrases du point de vue de l'assertion montre qu'une majorité écrasante de phrases contient deux éléments fondamentaux de contenu : une affirmation et un élément au sujet duquel l'affirmation est faite. L'élément au sujet duquel quelque chose est affirmé peut être appelé la base de l'énoncé ou le thème et ce qui est afffirmé au sujet de la base est le noyau de l'énoncé ou le rhème ». (Mathesius, 1975 : 81)

Le thème occupe explicitement le début de la phrase. Ce critère positionnel est définitoire de la notion de thème dans cette vision (par exemple chez (Trávnícek, 1962)).

La conception fonctionnelle de (Halliday, 1985) est similaire à celle de l'école de Prague, puisque le thème phrastique est le point de départ choisi

[7] Nous reparlerons de cette vision dans la section suivante.

[8] Notons que le rattachement à l'une ou l'autre des conceptions pour un auteur ou une école peut évoluer.

par le locuteur pour énoncer son message : « The Theme is a function in the CLAUSE AS A MESSAGE. It is what the message is concerned with: the point of departure for what the speaker is going to say. » (Halliday, 1985 : 36). Cependant, pour Halliday, la position initiale du thème n'est pas un critère définitoire mais un des moyens par lequel le thème se réalise dans certaines langues[9] : « As a general guide, the Theme can be identified as the element which comes in first position in the clause » (Halliday, 1985 : 39).

- **Le thème comme *« ce dont il est question »* :** cette conception du thème défini comme « ce dont on parle » est la plus fréquente. Elle correspond à la notion d'*aboutness* qui peut être considérée soit du point de vue pragmatique (Daneš, 1974 ; Lambrecht, 1994 ; Dik, 1978), soit du point de vue syntaxique (Givón, 1983). Vue du côté de la pragmatique, la notion d'*à-propos* correspond à la dualité *topic-comment* de Daneš ou *topic-focus* de Dik. (Lyons, 1970 : 257) a une conception similaire et appelle thème « la personne ou la chose dont on dit quelque chose ». Le thème s'oppose au commentaire qui correspond à « ce que l'on dit de ce dont on parle, que l'on affirme ou que l'on nie du topique ». Pour préciser cette définition en termes d'à propos, (Lambrecht, 1994) distingue le thème de l'expression thématique (*topic expression*) qui y renvoie :

  > « TOPIC: A referent is interpreted as the topic of a proposition if *in a given situation* the proposition is construed as being *about* this referent, *i.e.* as expressing information which is *relevant to* and which increases the addressee's *knowledge of* this referent. […] Topic is *a pragmatically construed sentence relation*.
  >
  > TOPIC EXPRESSION: A constituent is a topic expression if the proposition expressed by the clause with which it is associated is pragmatically construed as being about the referent of this constituent. » (Lambrecht, 1994 : 131)

  La notion de thème exprime ici l'information pertinente pour l'auditeur qui va lui permettre d'enrichir ses connaissances à propos de ce référent ((Lambrecht et Michaelis, 1998 : 494), cité par (Brunetti *et al.*, 2012)).

  Vu du côté de la syntaxe, l'accent est mis sur le lien explicite établi entre thème et sujet tel que défini chez (Givón, 1976, 1983), où les sujets non marqués sont des thèmes par défaut. Ces deux points de vue (pragmatique

[9] Selon (Prévost, 1998) et (Charolles et Prévost, 2003), la position initiale du thème tend pourtant à constituer un critère définitionnel.

et syntaxique) ne sont pas étanches puisqu'ont lieu des déplacements fréquents entre pragmatique et syntaxe pour mettre en évidence les points communs du thème pragmatique et du sujet syntaxique (Prévost, 1998, 2001). Dans les phrases suivantes :

a) *Paul* a acheté des sushis mais *il* ne les aime pas vraiment.

b) *Paul* a acheté des sushis mais *Ø* ne les aime pas vraiment.

Les constituants sujets « Paul » et « il » en a) sont deux expressions thématiques (suivant la terminologie de Lambrecht). Il est possible en b) d'omettre le pronom personnel sujet « il » sans rendre la phrase irrecevable. Cela s'explique par le fait que « Paul » est le sujet mais aussi le thème de la phrase.

- **Le thème comme élément « connu »**

Dans cette approche, le thème est ce qui est donné dans le texte ou ce qui est connu (ou supposé connu) par le destinataire alors que le rhème est ce qui est inconnu, non récupérable ou nouveau. Certains auteurs comme (Chafe, 1970, 1976) parlent d'« information ancienne/information nouvelle » : « La distinction entre information ancienne et nouvelle est le principal phénomène qui sous-tend les discussions sur ce qui a été appelé topique et commentaire, ou thème et rhème » ((Chafe, 1970 : 211) cité par (Touratier, 2010)). L'information ancienne est récupérable à partir de la situation (ou par anaphore) alors que l'information nouvelle est focale car elle ne peut pas être récupérée par les informations antérieures (Halliday, 1967a, 1967b). Le thème (*i.e.* l'élément connu) est ici considéré comme « saillant » car il est présent dans la mémoire à court terme de l'interlocuteur (Combettes, 1992).

A cette relation bipartite connu-nouveau, (Chafe, 1987, 1994) ajoute le degré *accessible* qui vient s'insérer entre les deux premiers[10]. La définition du thème en terme d'information supposée donnée (*givenness*) peut être alors évaluée selon ((Prince, 1981) citée par (Grobet, 2002)) de trois manières différentes : en fonction de l'aspect prévisible ou récupérable de l'information (*predictability/recoverability*) (Halliday, 1967a, 1967b ; Halliday et Hasan, 1976), suivant la saillance de l'information stockée dans la mémoire de l'interlocuteur (*saliency*) (Chafe, 1976) et selon le degré de connaissance partagée (*Shared Knowledge*) (Clark et Haviland, 1977). A

[10] A cette première échelle de saillance ou d'accessibilité cognitive des référents, d'autres échelles vont suivre, telles que celles de (Givón, 1983), de (Levelt, 1989) ou d'(Ariel, 1990).

partir de ce dernier point, Prince distingue plusieurs degrés de familiarité présumée : des informations totalement nouvelles et inconnues de l'interlocuteur (*brand-new*), des informations connues mais non utilisées (*unused*), des informations inférées par ce qui est dit ou par la situation (*inferrable*), des situations déjà évoquées par leur présence dans la situation ou par le texte (*evoked*) (voir Figure 1[11]). Par exemple, dans :

I got on *a bus* yesterday and *the driver* was drunk. (Prince, 1981 : 233)

« a bus » est un élément nouveau (*brand-new*) tandis que « the driver » est inférable par rapport à « the bus » car nous savons que les autobus possèdent un chauffeur.

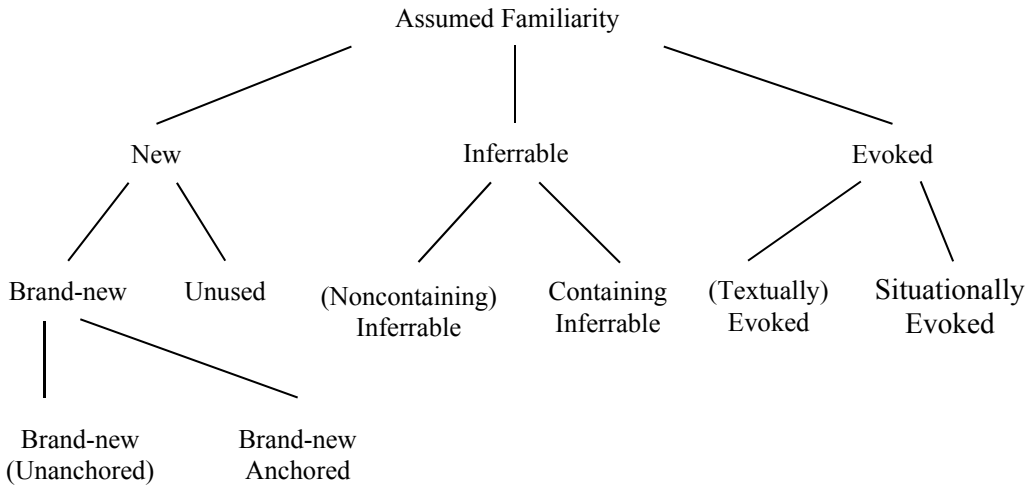


Figure 1 - Familiarité présumée d'après (Prince, 1981 : 237)

(Lambrecht, 1994) nuance cette notion d'information ancienne pour le thème : « affirmer comme cela est souvent fait dans les discussions pour le topique, que le topique est « l'information ancienne » est, pour le moins, une erreur » ((Lambrecht, 1994 : 164-165), cité par (Touratier, 2010)). Dans son approche, il utilise plutôt le terme de *présupposition* pour l'information « ancienne » et *assertion* pour l'information « nouvelle », qui prend la forme d'une proposition abstraite. Par exemple, dans « C'est la clé qui est coincée », la présupposition est « quelque chose est coincé » et l'assertion est « le quelque chose qui est coincé est la clé ».

[11] Voir aussi (Combettes, 1992 : 12) qui propose une progression des connaissances supposées partagées, allant de la « plus familière » à la « moins familière » : Elément évoqué > Non utilisé > Inférence > Représentation partielle > Nouveau rattaché au contexte > Entièrement nouveau (non rattaché au contexte).

Pour (Prévost, 1998), la majorité des approches (à part celle de Lambrecht) qui considèrent le thème comme élément connu ne constituent pas des analyses de la structure informative de l'énoncé, car ce sont souvent les référents que l'on qualifie de « connu/nouveau » :

> « on a tendance à assimiler les analyses formulées en terme de *connu/nouveau* à des approches traitant de l'information. Or c'est rarement le cas. En effet, on assiste fréquemment à une confusion quant à l'objet que l'on qualifie de connu ou de nouveau, qui, le plus souvent, n'est pas l'information elle-même. » (Prévost, 1998 : 17).

Pour Prévost, l'information est « un concept relationnel » (Prévost, 1998 : 18 ; 2001 : 34) : c'est la mise en relation des référents qui crée l'information, mais les référents ne sont pas informatifs par eux-mêmes. Lambrecht a d'ailleurs formulé explicitement la différence entre la représentation mentale du référent et les relations dans lesquelles le référent est impliqué :

> « The definition of a topic as a referent which stands in a certain RELATION to a proposition makes the topic concept intrinsically different from the concept of identificability and activation, which have to do with the PROPERTIES of (the representations of) discourse referents in the interlocutors' minds at given points in a conversation. The distinction between the mental representations of referents and the pragmatic relations which these referents enter into as elements of propositions is related to the distinction between "old/new referents" and "old/new information" discussed in Section 2.2. And like that distinction, it has often been neglected in the discourse-pragmatic literature. » (Lambrecht, 1994 : 160)

- **Le thème comme base du Dynamisme Communicatif :** cette conception a surtout été développée par Firbas à la suite de Mathesius. Dans son approche, (Firbas, 1966, 1972) gradue le couple thème-rhème en termes de « degrés de Dynamisme Communicatif » (CD en anglais), c'est-à-dire la force avec laquelle le thème ou le rhème contribuent au développement de la communication. Dans cette vision, le thème constitue l'élément peu informatif alors que le rhème est l'élément qui possède le degré le plus élevé :

> « By the degree of CD carried by a sentence element we understand the extent to which the sentence element contributes to the development of the communication, to which, as it were, it 'pushes' the communication forward. The elements carrying the

> lowest degree of CD constitute the theme, those carrying the highest degree, the rheme. » (Firbas, 1966 : 240).

En règle générale, les phrases progressent depuis l'élément ayant le plus bas degré de Dynamisme Communicatif (le thème) vers celui possédant le taux le plus élevé (le rhème). En plus du *thème* et du *rhème*, Firbas ajoute la *transition*, permettant la création d'une sorte de continuum informatif. La transition possède un degré de Dynamisme Communicatif plus important que le thème mais moins élevé que le rhème. On obtient alors une suite thème – transition – rhème pour la répartition de la force communicative, comme par exemple « *M. Brown* [thème] *s'est révélé* [transition] *un excellent professeur* [rhème]» (exemple issu de (Firbas, 1966 : 240) cité par (Touratier, 2010)).

Le degré de Dynamisme Communicatif est le résultat de l'interaction entre plusieurs facteurs issus de la Perspective Fonctionnelle de la Phrase : l'ordre linéaire des mots, la dépendance par rapport au contexte, la structure sémantique[12].

> « in assessing degrees of CD, the analyses of the written texts have taken into consideration (i) linear modification, (ii) the character of the semantic content of the linguistic element as well as the character of the semantic relation involved, and (iii) the retrievability of the information from the immediately relevant preceding context. An interplay of these three factors determines the distribution of degrees of CD over the written sentence. It determines the perspective in which a semantic and grammatical sentence structure is to function in the act of communication; that is, it determines its functional sentence perspective ». (Firbas, 1992 : 10-11).

Le thème n'est pas tenu à une position spécifique dans la phrase, il est lié au contexte précédent (Porhiel, 2005a). La position d'un élément dans l'organisation syntaxique de la phrase suggère (mais ne détermine pas) l'interprétation qui va en être faite en termes de thème et rhème.

Lorsque les facteurs de dépendance par rapport au contexte et de structure sémantique ne peuvent pas s'appliquer, le degré de Dynamisme Communicatif augmente par degrés entre le début et la fin de la phrase. Néanmoins, il paraît difficile d'identifier un thème s'il n'est pas connu de l'interlocuteur (même si l'interlocuteur pourrait identifier le thème (ou une partie du thème) grâce au contexte).

[12] Un dernier critère est l'intonation pour les langues parlées.

Même si le modèle de Firbas décrit la dynamique de la phrase de manière plus détaillée qu'une simple dichotomie thème/rhème et que son approche n'oblige pas le thème à occuper la position initiale dans la phrase, l'analyse en termes de Perspective Fonctionnelle de la Phrase semble être une tâche difficile à reproduire et à vérifier (Goutsos, 1997 ; Grobet, 2002). De plus, l'évaluation du degré de Dynamisme Communicatif apparaît complexe vu que la notion même n'est pas donnée (elle doit être calculée suivant les trois facteurs), ce qui n'aide pas l'identification des thèmes et des rhèmes.

2.1.3 Bilan

(Demol, 2010 : 153) fournit une synthèse des différentes conceptions du thème phrastique que nous reprenons ci-dessous :

définition	affinement des critères	sources
Information donnée (ou connue), en termes de…	*prédictibilité*	Kuno (1978), Halliday (1967a, 1967b), Halliday et Hasan (1976)
	saillance	Chafe (1976, 1980)
	connaissance partagée > la familiarité présumée (assumed familiarity)	Clark et Haviland (1977), Kuno (1978, 1979), Prince (1981), Copeland et Davies (1983)
point de départ de l'énoncé	*interprétation fonctionnelle-psychologique dominante*	Trávnicek (FSP), Halliday (1994), Magretta (1977)
	interprétation formelle-psychologique dominante	Chomsky (1965), Foley et van Valin (1984), Gundel (1985), van Oosten (1986)
élément porteur du plus bas degré de dynamisme communicatif		Firbas (1964, 1986)
ce dont on parle *(aboutness relation)*	*interprétation pragmatique dominante*	Bally (1965), Daneš (1974), Hawkinson et Hyman (1974), Dik (1978), Reinhart (1982), Gundel (1985, 1993), van Oosten (1986), Lambrecht (1994)
	Interprétation syntaxique dominante (topique=sujet)	Keenan (1976), Givón (1984)

Tableau 1 - Les différentes conceptions du thème phrastique (Demol, 2010 : 153)

(Grobet, 2002 : 22) remarque que ces quatre conceptions du thème « ne sont pas nécessairement exclusives : elles permettent au contraire des recouvrements et des croisements ». Néanmoins, il semble bien que plusieurs termes soient utilisés pour définir des objets linguistiques différents du terme unique « thème », car le *point de départ*, l'élément *connu*, la *base* ou *ce dont il est question* ne coïncident pas

nécessairement. De plus, si les deux premières conceptions du thème (*i.e.* le thème comme point de départ et comme ce dont il est question) peuvent être intraphrastiques, les deux dernières (*i.e.* le thème comme élément connu et comme base du Dynamisme Communicatif) nécessitent la prise en compte d'un contexte plus large et prennent alors une dimension plutôt textuelle (Prévost, 1998, 2001).

Si l'on admet une dichotomie thème/rhème au niveau de la phrase, le passage au niveau textuel nécessite l'étude des transitions entre phrases (*i.e.* les progressions thématiques).

2.2 Les progressions thématiques

La structure d'un document se lit à partir de sa « construction thématique » (Tomassonne, 2002). Les phrases d'un document suivent un enchaînement spécifique, une organisation thématique, que (Daneš, 1974) a nommé *progression thématique*[13] et qu'il définit comme :

> « the choice and ordering of utterance themes, their mutual concatenation and hierarchy, as well as their relationships to the hyperthemes of the superior text units (such as paragraph, chapter…), to the whole of the text, and to the situation. » ((Daneš, 1974 : 114) cité par (Goutsos, 1997 : 15))[14].

Daneš présente trois types canoniques de progression thématique : la progression à thème *linéaire*, la progression à thème *constant* et la progression à thèmes *dérivés*. Il est à noter qu'un texte présente rarement un seul type de progression (Riegel *et al.,* 2002 : 609). De plus, ces types abstraits peuvent se combiner ou se réaliser avec des variantes (omissions, insertions).

– **Progression à thème linéaire :**

Dans la progression à thème linéaire (le cas considéré comme classique), le rhème de chaque phrase est « l'origine » du thème de la phrase suivante, tel que schématisé[15] ci-dessous :

[13] « Our basic assumption is that text connexity is represented, *inter alia*, by thematic progression (TP). » (Daneš, 1974 : 114), cité par (Carter-Thomas, 2000 : 89)).

[14] (Carter-Thomas, 2000 : 89) reformule cette définition de la manière suivante : « le terme de progression thématique désigne l'ensemble des relations thématiques dans le texte : la concaténation et la connexion des thèmes, leur ordre et la hiérarchie qui les unit, dans leurs relations aux paragraphes et à l'ensemble du texte ainsi qu'à la situation de communication. C'est le cadre qui permet à l'ensemble de prendre forme. ».

[15] Dans les diagrammes, *Ph* représente la phrase, *Th* le thème, *Rh* le rhème et *HTh* l'hyperthème.

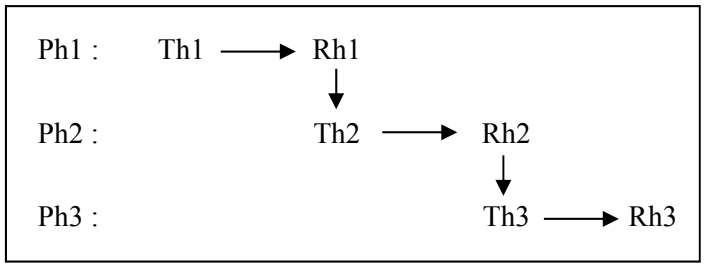


Figure 2 - Progression à thème linéaire

On peut illustrer ce premier type de progression par l'exemple suivant (les thèmes sont en italique et les rhèmes sont en gras) :

> *Il* entra, gratta ses pieds sur une grille luisante aux lames acérées et suivit **un couloir bas bordé par des lampes à lumière pulsée**. Tout au bout *du couloir*, il y avait **une porte**. *Elle* portait le numéro indiqué dans le journal et il entra sans frapper comme le recommandait *l'annonce*. (Boris Vian, *L'Écume des jours*)

Ce premier type de progression est privilégié dans l'argumentation. Qu'il s'agisse de la progression à thème constant ou des deux autres types de progression thématique, le constituant rhématique n'est pas nécessairement repris à l'identique dans l'énoncé qui suit (on pourra avoir un hyponyme, un méronyme, un hyperonyme, une description définie ou une anaphore).

– **Progression à thème constant**

La progression à thème constant, de son côté, conserve le thème dans des phrases successives, alors que les rhèmes sont différents d'une phrase à l'autre. Ce type de progression, privilégié dans les narrations (souvent pour le personnage principal), peut être modélisé de la manière suivante :

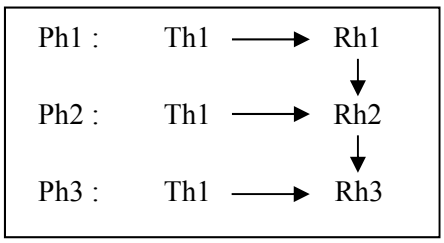


Figure 3 - Progression à thème constant

Voici un exemple de progression où le thème (en italique) demeure constant dans les trois phrases suivantes et fait l'objet de reprises *via* le pronom « il » :

> *Le groupe d'enquête* a sélectionné plusieurs expériences afin de couvrir le plus largement les champs d'intervention des services publics. *Il* a ainsi retenu des organismes publics où la relation avec les usagers s'intègre

> dans un contexte concurrentiel ou potentiellement concurrentiel (La Poste, une école d'ingénieurs, certaines activités du ministère de l'équipement). *Il* a également rencontré des responsables d'organismes œuvrant dans les secteurs sanitaire et social (caisses d'assurance-maladie ou de retraite, hôpitaux) et d'administrations à caractère plus régalien, où le service est par nature imposé aux usagers (impôts, préfectures et police nationale). (corpus *La Documentation Française*)

– **Progression à thèmes dérivés**

Enfin, dans la progression à thèmes dérivés (ou *éclatés* ou *en éventail*), les thèmes sont issus d'un « hyperthème » (HT) qui peut se trouver au début du passage ou dans un passage précédent (ou bien il peut ne pas être textualisé du tout). Ainsi, les relations thèmes-rhèmes ne lient plus seulement les thèmes et rhèmes de deux énoncés consécutifs, mais les thèmes et rhèmes de plusieurs phrases à un hyperthème dont les thèmes des différentes phrases sont alors appelés des « sous-thèmes » (voir Figure 4) :

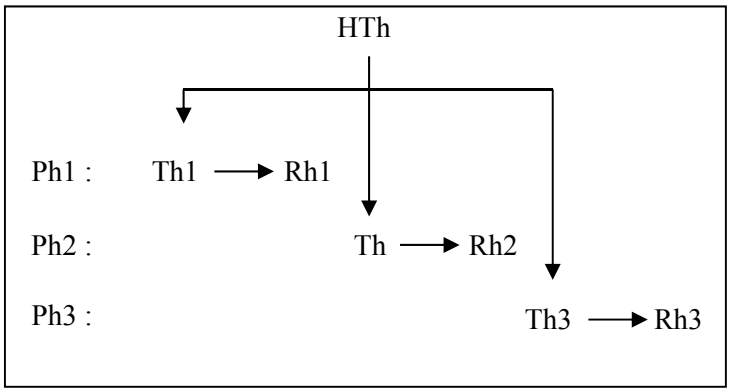


Figure 4 - Progression à thèmes dérivés

Ce type de progression apparaît de manière plus ou moins évidente selon que l'hyperthème est mentionné explicitement. Ce type est privilégié dans les descriptions, comme par exemple (l'hyperthème « un jeune homme » est mentionné en début de paragraphe) :

> **Un jeune homme**... - **Visage** long et brun ; la **pommette** des joues saillante, signe d'astuce ; **les muscles** maxillaires énormément développés, indice infaillible auquel on reconnaît le Gascon. (corpus *les Trois Mousquetaires*)

(Combettes, 1977, 1988) souligne l'intérêt des progressions thématiques de Daneš qui permet d'aller plus loin que le simple stade de l'analyse phrastique. Néanmoins, comme il a déjà été signalé, un texte possède souvent plusieurs types de progressions qui peuvent se combiner et rendre, de ce fait, le processus plus

complexe (Adam, 1977). De plus, Daneš signale que de manière indépendante aux progressions thématiques, il arrive qu'aucun élément thématique ne fasse l'objet de reprise (on parle alors de *rupture*), ces ruptures étant très fréquentes pour des textes longs d'après (Combettes, 1978). De ce fait, cette approche ne suffit pas à elle seule à rendre compte de la cohérence globale du texte. En effet, la cohérence est construite par l'interlocuteur qui va, au fil du texte, mémoriser puis synthétiser des informations situées au-delà du cadre interphrastique dans une même unité sémantique correspondant au thème textuel.

2.3 Thème textuel

(Grobet, 2002 : 25) remarque que le thème textuel « est souvent exclu des discussions terminologiques ». La notion de thème textuel[16] serait introduite par ((Keenan et Schieffelin, 1976 : 338), cités par (Grobet, 2002)) et réfèrerait à : « the PROPOSITION (or set of propositions) about which the speaker is either providing or requesting new information ». Le thème textuel n'est pas la somme des thèmes phrastiques d'un texte. Il n'est pas non plus une extension de la notion de thème phrastique (Goutsos, 1997).

De manière analogue au thème phrastique, le thème textuel fait l'objet de plusieurs conceptions, allant d'une conception commune (le thème textuel comme *idée centrale)* à une idée de groupement (le thème textuel comme *agrégat*) en passant pas une conception sémantico-pragmatique (le thème textuel comme *à propos*). Néanmoins, à la différence du thème phrastique, le thème textuel n'est pas le premier élément d'un couple dichotomique thème/rhème (ou inclus dans un continuum) et il ne possède pas de rôle fonctionnel dans la distribution linéaire de l'information (van Dijk et Kintsch, 1983). Nous proposons ci-dessous différentes conceptions du thème pris dans sa dimension textuelle.

2.3.1 Le thème textuel comme *idée centrale*

Pour des auteurs tels que (Wilson, 1998 : 68), le thème textuel permet d'accéder à l'idée générale du texte : « it is widely accepted that the function of a discourse topic is to provide access to contextual information required for the comprehension of the associated text or discourse ». Plutôt que d'être une simple

[16] On retrouve une nouvelle fois dans la littérature une diversité terminologique à propos de cette notion (thème textuel, thème de discours, thème discursif) ainsi que des variantes (topique textuel, topique discursif, topique de discours, *topic* discursif, etc.). Nous retiendrons dans notre approche la première dénomination (thème textuel) pour conserver l'opposition phrase (local)/texte (global).

addition ou une valeur moyenne des thèmes phrastiques, le thème textuel est vu comme le résultat d'un traitement cognitif des phrases prises dans leur ensemble (*i.e.* l'idée principale) :

> « The ideas about theme developed in this study have their roots in the rather intuitive understanding of theme that most of us had in primary and secondary school – that is, that theme is "main idea" in a text. The theme-line of a text is its "central thread". Theme also may be described as a "minimum generalization" of a text: a statement broad enough to represent the entire text, yet specific enough to represent its uniqueness. » (Jones, 1977 : v)

La première conception du thème textuel comme idée principale reflète l'intuition générale que l'on peut avoir de cette notion. Néanmoins, l'accent est mis sur l'unité du texte, sa cohérence, représentée par le thème textuel.

2.3.2 Le thème textuel comme *à propos*

Comme le thème phrastique, le thème textuel peut être formulé en termes d'à propos (*aboutness*). Dans cette conception, le thème textuel constitue ce à propos de quoi portent les entités principales du texte :

> « A discourse, taken in the wide sense of any kind of coherent text (a story, a monologue, a dialogue, a lecture, etc.) is "about" certain entities. » (Dik, 1997 : 313)

> « 'topic' is a relevant functional notion only at the discourse level, minimally at the chain or paragraph level. Put plainly and in operational terms, the topic is only 'talked about' or 'important', if it remains 'talked about' or 'important' during a number of successive clauses. » (Givón, 1990 : 902)

Dans son approche, (Dik, 1997) concilie le thème phrastique et le thème textuel. Pour lui, le texte contient plusieurs thèmes ordonnés de manière linéaire ou de manière hiérarchique. Chaque thème est d'abord considéré comme nouveau lorsqu'il sera introduit la première fois en discours (*new topic*). Puis, il sera donné (*given topic*) lorsqu'il sera repris. Un thème sera considéré comme sous-thème (*sub-topic*) lorsqu'il n'est pas énoncé de manière explicite dans le texte mais qu'une inférence à partir d'un autre thème explicite le rend disponible. Le dernier type de thème est le thème réactivé (*resumed topic*) qui correspond à un thème qui n'a pas été mentionné récemment dans le discours et qui doit être réintroduit pour retrouver le statut de thème donné.

Dik définit des stratégies (*topicality strategy*) permettant l'introduction, le maintien et la réactivation des thèmes dans le discours (Dik, 1997 : 315-326) :

- Parmi les stratégies permettant d'**introduire** un nouveau thème textuel, on trouve des énoncés qui vont définir de manière explicite ce qui va être le thème textuel : des énoncés métalinguistiques (*e.g.* « Je vais vous raconter une histoire à propos de M. Martin »), des constructions existentielles (*e.g.* « Il était une fois un prince »), des thèmes en position objet (*e.g.* « Dans le hall nous vîmes arriver une femme avec un grand chapeau. »).
- Concernant les stratégies utilisées pour **maintenir** un thème textuel, on trouve principalement ce qui est appelé des *chaînes thématiques* (*topic chains*) et qui incluent les relations anaphoriques (pronoms, anaphores zéro, etc.) ainsi que les parallélismes syntaxiques (*i.e.* les thèmes apparaissant dans des positions syntaxiques similaires). Un thème donné peut aussi être maintenu *via* un sous-thème, c'est-à-dire une entité qui peut être inférée à partir de relations d'hyperonymie, de méronymie, « on the basis of our knowledge of what is normally the case in the world » (Dik, 1997 : 323). Par exemple, dans « Paul a organisé *une réception* samedi dernier, mais *la nourriture* n'était pas bonne », nos connaissances nous permettent de savoir que l'on a coutume de servir de la nourriture lors d'une réception (c'est-à-dire que « nourriture » est une partie de « réception »). De ce fait, il est possible de parler à propos de la nourriture comme si elle avait été introduite dans le discours.
- Pour ce qui est des stratégies visant à **réactiver** le thème, le locuteur peut indiquer explicitement qu'un changement thématique est en cours. Il peut utiliser des références anaphoriques, il peut aussi référer au thème réactivé comme à une entité déjà mentionnée précédemment (de manière à ce qu'il soit en quelque sorte accessible à l'auditeur) ou enfin utiliser une combinaison de ces trois moyens.

Ces stratégies conçues dans une approche discursives sont considérées comme une « extension » d'une théorie plutôt centrée sur la phrase, objet d'étude principal du courant fonctionnaliste auquel Dik appartient (Bilhaut, 2006). De ce fait, son approche diffère peu des modèles pragmatiques tels que celui de Lambrecht.

De son côté, (Chafe, 1994, 2001, 2008) définit le thème textuel d'un point de vue conversationnel comme une collection d'idées introduites et développées par les participants du dialogue :

> « A topic in this sense is a coherent collection of ideas, introduced by some participant in a conversation and typically

> developed mainly by that participant, although by several participants jointly. A topic may then be explicitly closed, or it may just peter out. » (Chafe, 2008 : 27)

De ce point de vue, le thème textuel ne correspond pas forcément à un thème particulier mais il se constitue à partir d'un ensemble d'éléments du contexte considérés comme saillants ou activés[17] (Brown et Yule, 1983). De plus, le thème textuel est négocié par chaque locuteur au cours de la conversation :

> « if there is an entity identifiable as "the topic of conversation", the analyst should consider what evidence from each individual speaker's contributions he is using to make that identification. He should also remain aware of the fact that conversation is a process and that each contribution should be treated as part of the negociation of "what is being talked about". » ((Brown et Yule, 1983 : 94) cité par (Grobet, 2002 : 26-27)).

Chafe conçoit le thème textuel comme l'ensemble des informations semi-actives situées dans la conscience du locuteur :

> «We can think of each such topic as an aggregate of coherently related events, states and referents that are held together in some form in the speaker's semiactive consciousness. A topic is available for scanning by the focus of consciousness, which can play across the semiactive material, activating first one part and then another until the speaker decides that the topic has been adequately covered for whatever purpose the speaker may have in mind. » (Chafe, 1994 : 112)

Néanmoins, les éléments constitutifs du thème peuvent avoir une accessibilité variable, ce qui signifie que le thème peut se manifester à différents degrés. Suivant cette perspective purement cognitive, l'identification du thème textuel est pour le moins subjective (Brown et Yule, 1983), car elle est fonction du point de vue des participants au dialogue.

2.3.3 Le thème textuel comme *macrostructure*

Dans ses travaux, (van Dijk, 1977b, 1983) oppose de manière explicite le thème textuel au thème phrastique : le thème textuel reflète l'organisation (hiérarchique) globale du texte alors que le thème phrastique réfère à l'organisation linéaire de la phrase :

[17] Ces éléments saillants forment le cadre thématique (*topic framework*) dans lequel le thème est constitué pour (Brown et Yule, 1983) (cf. section 2.3.3).

> « At the level of the sentence, a topic is a specific function assigned to some part of a (possibly compound) proposition and indicates the way information is *linearly distributed*, whereas a textual topic indicates how information is *globally organized*. In the first case, the topic is the link, between given and new information, for each sentence in the discourse, whereas the textual topic is the hierarchical organization of the whole of information of all sentences, taken 'at the same time'. » (van Dijk, 1977b : 59)

Pour traiter la question de la compréhension du discours, (van Dijk, 1977b), (Kintsch et van Dijk, 1978) ont proposé de distinguer le niveau sémantique microstructurel du niveau macrostructurel :

- **niveau microstructurel** : situé au niveau de la phrase, il concerne l'organisation structurelle et les relations entre les différentes propositions du discours. En suivant la cohérence référentielle établie entre les différentes mentions d'un référent, il est ainsi possible d'identifier la microstructure d'un texte :

  > « We can establish a linear or hierarchical sequence of propositions in which coreferential expressions occur. The first (or superordinate) of these propositions often appear to have a specific cognitive status in such a sequence, being recalled two or three times more often than other propositions. » (Kintsch et van Dijk, 1978 : 365)

- **niveau macrostructurel** : il identifie et décrit la structure sémantique du texte dans sa globalité. C'est sur cette base que repose la notion de thème textuel (*i.e.* un « résumé » du texte) :

  > « the propositions of a text base must be connected relative to what is intuitively called a topic of discourse (or topic of conversation), that is, the theme of the discourse or a fragment thereof. Relating propositions in a local manner is not sufficient. There must be a global constraint that establishes a meaningful whole characterized in terms of a discourse topic ». (Kintsch et van Dijk, 1978 : 366)

La macrostructure du discours est obtenue à partir de l'interprétation de propositions issues de chaque phrase du texte : « that is, a macro-structure of a sequence of sentences is a SEMANTIC REPRESENTATION of some kind, viz a proposition entailed by the sequence of propositions underlying the discourse (or

part of it). » (van Dijk, 1977b : 137). Cette proposition complexe[18] (Grobet, 2002) se compose d'une série de propositions organisées de manière hiérarchique. Les éléments constitutifs de la macrostructure peuvent être les titres, les sous-titres et le thème de la première phrase de chaque paragraphe. D'autres facteurs importants pour (van Dijk, 1977a) interviennent dans la construction de la macrostructure : les connaissances générales de l'interlocuteur, ses expériences, le contexte. De ce fait, pour un même texte, les macrostructures établies par des interlocuteurs ont de fortes chances d'être différentes. Pour l'exemple suivant issu de (van Dijk, 1977b : 49), la macrostructure pourrait être « Eva a pris le train pour Prague et a commencé son nouveau travail » (« Eva took the train to Prague and started her new job ») :

> "Eva awoke at five o'clock that morning. Today she had to start with her new job in Prague. She hurriedly took a shower and had some breakfast. The train would leave at 6:15 and she did not want to come late the first day. She was too nervous to read the newspaper in the train. Just before eight the train finally arrived in Prague. The office where she had found the job was only a five minutes walk from the station (…)."

Pour relier le thème global du texte (la macrostructure) aux différentes propositions du texte (la microstructure), (van Dijk, 1977b ; Kintsch et van Dijk, 1978) proposent trois règles principales de correspondance (ou *macrorules*)[19] :

- la règle de **suppression** (*deletion*) élimine les informations non essentielles ainsi que celles qui ne seront plus pertinentes dans la suite du texte ;
- la règle de **généralisation** (*generalization*) condense les diverses propriétés des référents ;
- la règle d'**intégration** (*construction*) construit des propositions plus générales qui résument des faits, comme par exemple « payer » peut sous-entendre à la fois « faire du shopping » et « manger au restaurant », etc.

Ces règles ont donc pour objectif de simuler cognitivement les types de réduction d'information (sélection des informations importantes, reformulation, etc.) qui s'opèrent pour mener au résumé d'un texte (van Dijk, 1995)[20].

[18] Ou « macroproposition » chez van Dijk.

[19] « In order to show how a discourse topic is related to the respective propositions of a text base, we thus need semantic mapping rules with microstructural information as input and macrostructural information as output. » (Kintsch et van Dijk, 1978 : 366).

[20] « Macrostructures were related to their (local) *microstructures*, that is, to the propositions expressed by the sentences of the text, by mapping rules (e.g., those of deletion, generalization, and construction) that theoretically simulate the types of information reduction that characterizes the process of abstracting or summarizing a text. » (van Dijk, 1995 : 385).

Les règles de correspondance fonctionnent de manière récursive, ce qui signifie qu'il est possible d'avoir plusieurs niveaux de macrostructure tant que les contraintes des règles sont respectées (*i.e.* qu'elles ne sont pas violées). On obtient ainsi une organisation hiérarchique qui représente la structure thématique du texte. Au sommet de cette structure, on retrouve souvent, suivant le genre textuel, une proposition exprimée dans le titre, tel que « TORNADO KILLS 500 PEOPLE » pour un article de presse (van Dijk, 1985 : 117). Cette caractéristique liée au genre textuel se retrouve dans la forme même de la structure hiérarchique ou *superstructure*. La superstructure attribue un rôle prédéfini à différentes portions textuelles suivant le schéma correspondant au genre textuel d'occurrence. Par exemple, l'article de presse impose qu'en première position apparaisse l'information principale puis les causes et le contexte des événements (van Dijk, 1985 : 122).

Il apparaît dans l'approche sémantique formelle de van Dijk que le thème textuel ne nécessite pas d'être mentionné de manière explicite dans le texte, surtout s'il est issu de la règle de généralisation ou d'intégration (van Dijk, 1980). D'après (Wolters, 2001), l'approche de van Dijk souffre aussi en plusieurs points dégagés par (Gülich et Raible, 1977), notamment le manque de détails fournis sur la méthode permettant d'extraire les propositions et le manque de précisions données pour connaître dans quelle situation utiliser une règle de correspondance plutôt qu'une autre. (Brown et Yule, 1983) soulignent le caractère subjectif du thème de discours chez van Dijk qui fait essentiellement appel à l'interprétation : la macrostructure d'un texte consiste uniquement en une proposition logico-formelle constituée à partir d'une phrase résumant le texte.

2.3.4 Le thème textuel comme *cadre thématique*

Dans leur approche axée sur l'analyse conversationnelle, (Brown et Yule, 1983 : 70) qualifient la notion de *thème* comme « an intuitively satisfactory way of describing the unifying principle which makes one stretch of discourse 'about' something and the next stretch 'about' something else. ». Le texte serait ainsi découpé en fragments. Chaque fragment constituerait une unité cohérente car elle porterait sur un thème spécifique et l'on passerait d'un thème à un autre au cours de la conversation (les auteurs fournissent l'exemple suivant : « the conversationnalists stop talking about 'money' and move on to 'sex' (Brown et Yule, 1983 : 69-70)). Brown et Yule distinguent le « theme » du « topic » : le *theme* correspond au premier constituant de la phrase (et il définit le contexte initial) et le *topic* correspond au thème textuel.

Pour (Brown et Yule, 1983) un thème textuel n'a pas d'existence en lui-même. Ce sont les locuteurs qui vont viser un thème en construisant leur discours. De leur côté, les interlocuteurs devront « choisir » parmi une variété de thèmes potentiels. Il existe ainsi de nombreuses manières d'exprimer le thème textuel. Afin de délimiter les différentes interprétations possibles du thème textuel, Brown et Yule ont défini la notion de *cadre thématique* (ou *topic framework*) :

> « What is required is a characterisation of 'topic', which would allow each of the possible expressions, including titles, to be considered (partially) correct, thus incorporating all reasonable judgements of 'what is being talked about'. We suggest that such a characterization can be developed in terms of a **topic framework**. […] Those aspects of the context which are directly reflected in the text, and which need to be called upon to interpret the text, we shall refer to as *activated features of context* and suggest that they constitute the contextual framework within which the topic is constituted, that is, the *topic framework*. (Brown et Yule, 1983 : 74-75)

Le cadre thématique est un type de représentation du thème textuel qui consiste à activer, à un moment donné, des zones du contexte dans lesquelles les référents (objets, lieux, événements) sont situés[21]. C'est une reformulation de la conception du thème textuel comme à propos (cf. 2.3.2). Les divers éléments susceptibles de constituer des paraphrases du thème textuel sont les *entités thématiques*. Ainsi, lorsqu'un locuteur développe un des éléments du cadre thématique, il « parle » d'une entité thématique. Au sein du cadre thématique, les éléments activés correspondent aux entités thématiques « dont parle le locuteur ». (Brown et Yule, 1983 : 76, 78) proposent le cadre thématique suivant à partir de l'exemple conversationnel ci-dessous :

> R : in those days + when we were young + there was no local fire engine here + it was just a two-wheeled trolley which was kept in the borough + in the borough eh store down on James Street + and whenever a fire broke out + it was just a question of whoever saw the fire first yelling 'Fire' + and the nearest people ran for the trolley and how they got on with it goodness knows + nobody was trained in its use + anyway everybody knew to go for the trolley + well + when we were children + we used to use this taw [t>:] + it smouldered furiously + black thick smoke came from it and we used to get it burning + and then go to a letterbox and just keep blowing + open the letterbox + and just keep blowing the smoke in + you see + till you'd fill up the lower part of the

[21] « The topic framework, as we have described it, represents the area of overlap in the knowledge which has been activated and is shared by the participants at a particular point in a discourse. » (Brown et Yule, 1983 : 83).

> house with nothing but smoke + just to put the breeze up + just as a joke + and then of course + when somebody would open a window or a door the smoke would come pouring out + and then + everybody was away then for the trolley + we just stood and watched all of them ++
>
> S : so that's what 'smoke the houses' is ?
>
> R : probably + probably + we called it 'the taw' +

Cadre thématique (présenté sous forme de liste) :

> Conversation between Participant R (50+ years, Scottish, male, …) and Participant S (20+ years, American, female, …) in location p (Stornoway, …) at time t (late 1970s, …)
>
> A joke - the taw – smoke – into houses – out of houses – people get trolley – the use of the trolley.

Pour Brown et Yule, les thèmes textuels ne peuvent pas être réduits à de simples entités du discours. Le thème textuel n'est pas seulement ce à propos de quoi porte le segment de discours mais inclut aussi des concepts et des éléments du contexte. Leur approche est plus interactive (Demol, 2010) que celle de van Dijk par exemple, car elle permet de suivre la création des différents thèmes au cours du texte. Néanmoins, (Grobet, 2002) met en évidence le manque de critères objectifs et directement applicables permettant de déterminer quelles entités thématiques doivent être retenues dans le cadre thématique. Il semble que seul l'interlocuteur soit à même de décider cela, grâce à son expérience du monde et au contexte, donc avec ici aussi une certaine part donnée à la subjectivité. (Goutsos, 1997) note enfin que, dans leur vision, (Brown et Yule, 1983) ont tendance à confondre le thème textuel avec la totalité du contexte, ce qui n'aide pas à déterminer le thème.

2.3.5 Le thème textuel comme agrégat de thèmes phrastiques

Le thème textuel peut être vu comme un agrégat, une compilation (Porhiel, 2005a) ou une composition de thèmes phrastiques (Rimmon-Kenan, 1985 ; Marandin, 1988 ; Goutsos, 1997 : 17 ; Downing *et al.*, 1998 : 268). (Rimmon-Kenan, 1985 : 399) définit le thème textuel comme un assemblage d'éléments disséminés dans le texte : « Un thème n'est pas une entité *dans* : ce n'est pas un segment inclus *dans* le continuum du texte, mais une construction dont l'assemblage s'effectue à partir des éléments discontinus du texte. ». Dans cette approche, les éléments constitutifs du thème textuel sont explicitement présents dans le texte. Néanmoins, (Goutsos, 1997 : 11) souligne qu'il ne suffit pas

seulement de mettre des éléments les uns à la suite des autres pour former le thème textuel, il est nécessaire de regrouper ces éléments : « Discourse topic cannot simply stem from the concatenation of initial constituents in succession, without any larger-scale grouping of elements. »

Travaillant sur des textes non narratifs (*e.g.* des textes expositifs en anglais et en grec), (Goutsos, 1997) a mis au point un modèle qui considère le thème textuel comme un cadre structurant (*topic as a structuring frame*). Il rejoint (Schiffrin, 1992) :

> « topics – regardless of the type or level at which they are defined – are ultimately created through discourse. It is the interactions between speakers and respondents that create the structures and meanings of talk and, thus, that create the framework in which a message is understood to be 'about' something ». (Schiffrin, 1992 : 174), cité par (Goutsos, 1997))

Le modèle de Goutsos se focalise sur la séquentialité du discours : le thème textuel est considéré comme une entité qui n'est pas préétablie et dont la compréhension s'établit au cours du discours (« we progressively expand our understanding of what someone is talking about (Schiffrin, 1992 : 195) as the discourse unfolds » (Goutsos, 1997 : 30)). Ce sont les marques de paragraphe et les marqueurs linguistiques de continuité et de discontinuité utilisés par les locuteurs (*i.e.* des « traces » (Floor, 2004), des « indices ») qui permettent d'identifier la structure du discours et par là-même les thèmes : « linguistic elements are considered to be employed by writers and perceived by readers as signals of discourse patterns of organization. » (Goutsos, 1997 : 41). Ce sont les locuteurs et les interlocuteurs qui possèdent des représentations mentales des thèmes, pas les textes. Pour signaler la structure du texte, les locuteurs vont mettre en séquence des zones de continuation et des zones de transition. Les marques de ruptures et de continuités thématiques sont plus facilement identifiables que le contenu thématique lui-même. C'est donc la manière dont se forme le thème qui permet d'apporter un éclairage sur le thème.

La phrase est considérée comme une unité minimale de discours. Ce découpage permet au locuteur de marquer des ruptures et des continuités dans le texte. La rupture thématique est marquée par des indices linguistiques : utilisation d'indéfinis à l'initiale (« A new body, »), redénomination du nom propre. La continuité thématique entre deux phrases successives est aussi renforcée par des indices linguistiques explicites : des marqueurs de cohésion (« this »), l'utilisation du même temps grammatical (« was – was »), le parallélisme dans l'utilisation

d'adverbiaux à l'initiale (« For the newly industrialized world… – For this group, »), l'utilisation de marqueurs référentiels, la répétition. Ces marqueurs de continuité, utilisés seuls ou combinés à d'autres marqueurs, créent une redondance dans le texte permettant à l'interlocuteur de réduire ses efforts de compréhension. Ainsi, la séquentialité dans un texte, vue du côté du locuteur, consiste à mettre en évidence les zones de continuité (*continuation spans*) et de discontinuité (*transition spans*) qui seront perçues par l'interlocuteur pour pouvoir interpréter le texte :

> « Therefore, an equally important task for the writer is to indicate discontinuity within the larger presupposed continuity of the text. […] In short, the writer is faced with the tasks to manage the interaction through discourse in sequential terms and to segment discourse into chunks and indicate their boundaries; that is, the discontinuity between one and another. » (Goutsos, 1997 : 43).

Pour ce faire, plusieurs stratégies thématiques (*topic strategies*) de séquentialité sont proposées dans le modèle de Goutsos (voir Figure 5) : la continuité thématique (*topic continuity*) et le changement thématique (*topic shift*). Ces stratégies thématiques sont réalisées par des techniques séquentielles :

- la continuité thématique est réalisée par la technique de continuation thématique (*topic continuation*), comme, par exemple, entre les phrases 3.2 et 3.3 en italique (exemple issu de (Goutsos, 1997 : 43)) :

  > 3.1 But that is where the common perception of the south ended. 3.2 For the newly industrialized world, which wanted the right to follow the same path as the north, it was not only a matter of a green dividend. 3.3 For this group, including Mexico, Brazil, India, China and Malaysia, the unfettered ability to industrialise was essential, pollution an unavoidable consequence.

- le changement thématique est réalisé par les techniques de fermeture thématique (*topic closure*), de cadrage thématique (*topic framing*) et d'introduction thématique (*topic introduction*). Par exemple, le changement thématique est réalisé par le cadrage thématique « for the ecological movement » dans « *For the ecological movement, on the other hand*, nuclear power – centralised, polluting, expensive high technology – represented everything it hated. » (exemple issu de (Goutsos, 1997 : 52)).

On trouve aussi des techniques séquentielles de second ordre, plus complexes, telles que la digression thématique (*topic digression*) et la dérivation thématique (*topic drift).* Les techniques séquentielles peuvent être optionnelles (comme la

fermeture thématique et le cadrage thématique) ou obligatoires (comme l'introduction thématique et la continuité thématique) et elles suivent un ordre précis. De ce fait, sans introduction thématique par exemple, il ne peut y avoir de continuité.

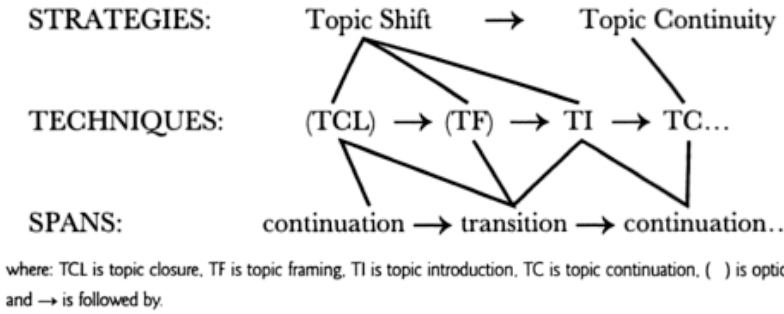


Figure 5 - Modèle de la structure thématique (Goutsos, 1997 : 75)

Les techniques séquentielles sont signalées par des indices linguistiques (ou signaux thématiques (*topic signals*)) explicites qui constituent à la fois des indices pour le locuteur et pour l'interlocuteur pour identifier les relations séquentielles du texte. Par exemple, le cadrage thématique est la technique séquentielle utilisée pour indiquer de manière explicite les limites séquentielles. Elle indique à la fois la fin d'une zone de continuité et le début d'une zone de discontinuité (*i.e.* transition) thématique. Elle se caractérise notamment par l'emploi de marqueurs du discours (*accordingly, of course, similarly, now, then*) utilisés en combinaison avec les marques de paragraphe, les adverbiaux cadratifs (*on the other hand*, *for the ecological movement*) et les marqueurs référentiels (les démonstratifs (*this*, *that*), les groupes nominaux anaphoriques). De son côté, l'introduction thématique signale la fermeture du cadre précédent et l'ouverture d'un nouveau cadre et elle est indiquée par une série d'indices linguistiques : des groupes nominaux indéfinis, la redénomination du nom propre, le changement de temps grammatical. L'introduction thématique implique que, dans la suite, le texte se caractérise par des marques de continuation thématique, où les phénomènes de répétition, de reprise anaphoriques pronominales et la constitution de chaînes de référence vont être utilisés pour maintenir le cadre thématique. Dans l'exemple suivant, l'introduction thématique intervient dans la phrase 4.5 et donne lieu à une continuation thématique (phrases 4.6 à 4.9) puis d'une fermeture thématique (phrase 4.10) :

> 4.5 Second, there is the question of the initial state of the universe. 4.6 Some people feel that science should be concerned with only the first part; they regard the question of the initial situation as a matter for metaphysics or religion. 4.7 They would say that God, being omnipotent, could have started the universe off any way he wanted. 4.8 That may be

> so, but in that case he also could have made it develop in a completely arbitrary way. 4.9 Yet it appears that he chose to make it evolve in a very regular way according to certain laws. 4.10 It therefore seems equally reasonable to suppose that there are also laws governing the initial state. (exemple issu de (Goutsos, 1997 : 95-96)).

Dans cet exemple, l'introduction thématique est signalée par le marqueur d'énumération présent à l'initiale « second » et renforcée par le présentatif « there ». La continuation entre 4.5 et 4.6 est notamment assurée par la répétition de « the question of ». La continuation entre les phrases 4.7 et 4.9 s'établit *via* l'utilisation de pronoms (« he, it »), de démonstratifs (« that »), de conjonctions (« yet »). La fermeture thématique est annoncée par le marqueur du discours « therefore ».

Goutsos hiérarchise les indices linguistiques selon leur importance dans l'établissement des techniques séquentielles (du plus explicite au moins explicite) :

> « the following tentative hierarchy of topic signal can be established: orthographic markers > metadiscourse items > prediction pairs > discourse markers > cohesive devices > time framing > sentence-structure patterns » (Goutsos, 1997 : 82).

Par exemple, les marques de paragraphe[22] sont plus explicites (*i.e.* visibles plus rapidement) que le changement de temps grammatical. Cette hiérarchie permet de résoudre les conflits éventuels entre plusieurs types de marqueurs. Néanmoins, Goutsos note que l'ordre de cette hiérarchie peut être modifié suivant le genre textuel. Par exemple, les textes narratifs privilégieront les marqueurs référentiels.

Pour (Carretero, 1998), les stratégies proposées dans le modèle de Goutsos sont redondantes : les changements thématiques et les continuations thématiques coïncident avec les zones de continuité et de rupture thématiques. De ce fait, le modèle pourrait être simplifié en supprimant les stratégies. (Carretero, 1998) souligne aussi que la linéarité du modèle de Goutsos ne prend pas en compte le rôle thématique joué par des marqueurs de cohésion dans des séquences non contiguës. En effet, pour l'exemple ci-dessous, Goutsos explique que son modèle n'autorise pas que « the first part » (phrase 4.6) ait un lien avec « parts » (phrase 4.2), « first » (phrase 4.3) et « second » (phrase 4.5) :

> 4.2 However, the approach most scientists follow is to separate the problem into *two parts*. 4.3 First, there are the laws that tell us how the universe changes with time. 4.4 (If we know what the universe is like at any one time, these physical laws tell us how it will look at any later

[22] Les marques de paragraphe opèrent souvent en synergie avec des marqueurs linguistiques pour identifier les relations thématiques.

> time.) 4.5 *Second*, there is the question of the initial state of the universe. 4.6 Some people feel that science should be concerned with only *the first part*; they regard the question of the initial situation as a matter for metaphysics or religion. (exemple issu de (Goutsos, 1997 : 94-95))

Nous pensons, avec Carretero, qu'il est indispensable de prendre en compte ces marqueurs de cohésion à « longue portée » qui peuvent s'étendre sur plusieurs phrases pour identifier la continuité thématique.

Cette continuité thématique, pouvant se dérouler sur plusieurs phrases ou paragraphes, est souvent liée à une notion issue de la psychologie cognitive : la saillance.

2.4 Saillance

Bien que la notion de *saillance* ne soit pas une notion linguistique en soi (Schnedecker, 2009 ; Landragin, 2012), elle apparaît comme une « notion sous-jacente à toute discussion sur la structuration thématique, aussi bien au niveau de la phrase qu'au niveau de la proposition ou qu'au niveau du texte entier » (Carter-Thomas, 2000 : 90). En effet, les diverses approches du thème, qu'elles l'abordent au niveau phrastique ou textuel, font ressortir l'idée qu'un élément est mis en relief dans la phrase, le paragraphe ou le texte. Il y aurait donc une saillance locale (au niveau de la phrase) et une saillance plus globale (au niveau du texte). Au niveau local, suivant les approches, l'élément jugé saillant constituerait soit le thème (*e.g.* un élément connu, déjà activé), soit le rhème (*e.g.* un élément nouveau, récent, le focus), ce qui n'est pas sans poser de problèmes (Carter-Thomas, 2000 ; Landragin, 2003, 2007, 2012). Au niveau global, l'élément saillant constituerait le référent central du discours :

> « la notion de *saillance* est employée en sémantique du discours pour décrire le statut de centralité de certains référents dans la conscience de l'énonciateur. Un référent est saillant s'il s'impose à l'attention. Certaines entités représentées dans le discours sont pensées comme plus centrales ou plus pertinentes que d'autres » ((Neveu, 2000 : 100), cité par (Schnedecker, 2009 : 5))

En analyse du discours, la saillance est une des propriétés d'une entité considérée comme privilégiée pour une reprise. Dans l'énoncé suivant,

> Finalement, *M. Giscard d'Estaing* a proposé une série d'économies. *Il* s'est demandé s'il était " urgent " de créer une chaîne éducative, *il* s'est interrogé sur le financement des interventions militaires ou humanitaires de la France dans le monde. (Le Monde)

« M. Giscard d'Estaing » possède les caractéristiques suivantes. Il s'agit :

- d'une entité située en tête de phrase, qui occupe la position de sujet grammatical,
- d'une entité en position de thème de paragraphe (introduite par le cadre de discours « finalement »),
- d'un référent dénommé par un nom propre (en première mention),
- d'un référent dénoté comme *animé* et *humain*.

Cette entité fait alors l'objet de reprises dans la suite de l'énoncé (par le biais du pronom « il ») et elle sera ainsi considérée comme saillante.

La saillance peut être vue comme un des moyens utilisés pour hiérarchiser les éléments du discours (Ariel, 1990 ; Chafe, 1994 ; Givón, 1983 ; Prince, 1981, voir Chapitre 2). Elle permet de rendre plus attractif un référent parmi d'autres, par l'utilisation d'expressions référentielles particulières : par exemple, le locuteur préfèrera employer le nom propre « Paul » plutôt que le groupe nominal simple défini « l'homme » s'il souhaite que ce référent joue un rôle dans la suite de l'énoncé. Ainsi, est saillant l'élément le plus présent à l'esprit de l'interlocuteur dans le segment textuel (saillance à effet immédiat, (Landragin, 2005)) ou dans l'ensemble du document (saillance à effet continu). Dans l'exemple ci-dessus, les pronoms réfèrent au même référent saillant « M. Giscard d'Estaing ». Mais ce référent saillant représente aussi un candidat thème potentiel pour le locuteur : « la saillance est à considérer non seulement comme présence du réfèrent dans la mémoire du récepteur, mais aussi comme « disponibilité » d'une unité, du point de vue de l'émetteur, à servir de thème » (Combettes, 1996 : 87).

Nous pensons qu'en identifiant les référents saillants du discours, nous allons détecter les thèmes des documents. L'identification automatique de la saillance référentielle d'une entité passe par l'étude préalable du fonctionnement des mécanismes de reprises référentielles dont elle fait l'objet (*e.g.* de la construction des chaînes de référence[23]) qui seront abordés dans les chapitres 2 et 3 de ce mémoire.

2.5 Bilan

Comme nous venons de le voir, les théories phrastiques dissocient deux pôles : le pôle thématique et le pôle rhématique. Néanmoins, ces deux pôles ne sont pas

[23] Cette notion sera définie dans la section 3.3 de ce chapitre.

considérés par tous comme étant exclusifs (topique et focus peuvent se confondre pour (Dik *et al.*, 1981)) et cette division binaire est à nuancer en terme de degré de force communicative chez (Firbas, 1964). Dans cette dernière vision, les thèmes et les rhèmes sont indépendants de leur structure syntaxique. De ce fait, le thème est lié au contexte précédent (Porhiel, 2005a ; Prévost, 2001) et non pas à une position déterminée dans la phrase.

Cette remarque nous amène à ne pas définir la notion de *thème* à un niveau essentiellement phrastique. En effet, cette perspective ne permet pas de rendre compte de phénomènes plus généraux qui émergent d'une structure plus complexe composée d'un ensemble d'énoncés reliés entre eux[24]. C'est ce que rend possible la notion de thème textuel. Ainsi, les différentes conceptions du thème textuel ont en commun l'idée que le thème textuel est un tout (« discourse topic seem to reduce, organize and categorize semantic information of sequences as wholes. » (van Dijk, 1977b : 132) cité par (Brown et Yule, 1983 : 109)) et forme une unité : une idée générale, un tout organisé dans un cadre thématique, dans une macrostructure ou dans un système de structuration séquentielle.

L'approche purement cognitive et non figée de Chafe semble peu se prêter à un traitement informatique. De la même manière, les approches prônant que le thème textuel (ou ses éléments constitutifs) peut ne pas figurer explicitement dans le texte[25] ne pourront pas être suivies dans une perspective d'identification automatique[26]. De ce fait, nous retiendrons l'approche du thème textuel comme agrégat de thèmes phrastiques et plus spécifiquement le modèle de (Goutsos, 1997) qui privilégie l'utilisation d'indices présents explicitement dans le texte pour signaler les continuités et les ruptures thématiques[27]. Nous souhaitons aussi prendre en compte des indices linguistiques à « longue portée » qui peuvent être présents dans des segments non contigus (tels que les cadres de discours, *cf.* section 3) pour signaler les continuités et les ruptures de thèmes.

Pour identifier de manière automatique les thèmes, nous avons à disposition plusieurs types d'indices de la structure du discours : des marques typo-

[24] Nous rejoignons ainsi (Mondada, 1994) pour qui « Travailler dans le cadre de la phrase correspond ainsi à se cantonner dans une identification statique d'un état informationnel. Par contre, dès que l'on pose la question des processus dynamiques par lesquels des syntagmes deviennent des topics, se développent comme tels ou se transforment voire sont abandonnés, le passage à un cadre discursif s'impose. » (Mondada, 1994 : 44-45), citée par (Grobet, 2002 : 37-38)).

[25] L'approche de (Marandin, 2007) prône aussi que la notion de *thème textuel* ne correspond pas forcément à un constituant dans l'énoncé.

[26] Nous admettons que la perspective automatique choisie peut paraître réductrice puisque, si le changement de thème n'est pas explicitement marqué *via* un marqueur linguistique par exemple, nous ne serons pas en mesure de le détecter.

[27] Néanmoins, nous ne retiendrons pas du modèle de Goutsos la hiérarchie établie entre les marqueurs, notamment le temps des verbes.

dispositionnelles et une série de marqueurs linguistiques de continuité et de discontinuité thématique. Nous présentons quelques indices de cohésion dans la section suivante.

3 Faisceau d'indices[28] de cohésion

Pour signaler les continuations ou les ruptures thématiques, plusieurs dispositifs de cohésion sont utilisés, parmi eux les marques de paragraphe, les cadres de discours, les connecteurs, les chaînes lexicales, les anaphores, les chaînes de référence[29], etc. Ces dispositifs constituent des outils permettant de relier des propositions les unes avec les autres et réduisent, ce faisant, les possibilités d'interprétation :

> « les langues possèdent des marques (ou systèmes de marques) destinées à indiquer les relations qu'entretiennent les unités composant un discours, donc des marques permettant de fournir à un destinataire potentiel des instructions interprétatives propres à favoriser la compréhension » (Charolles, 1988 : 4)

(Charolles, 1988) distingue quatre systèmes de marques (ou plan d'organisation textuelle) qui attribuent une continuité au discours :

- la *période*, constituée d'une série de phrases formant « un tout isolable du reste du texte », est marquée par les connecteurs (par exemple, « de fait ») ;
- les *chaînes*, constituées par des suites d'expressions référentielles référant à la même entité du discours (*i.e* les chaînes de référence) ;
- les *portées*, qui sont des portions de texte initiées par des expressions (les introducteurs de cadres de discours) qui délimitent des domaines temporels, spatiaux, modaux, etc.
- les *séquences*[30], c'est-à-dire le découpage en paragraphes et les organisateurs métadiscursifs (par exemple, « en guise de conclusion », « enfin »).

Ces différents plans d'organisation textuelle interagissent les uns avec les autres de manière permanente et c'est cette interaction qui permet d'assurer la continuité du discours.

[28] Nous emploierons indifféremment les termes « indice » et « marqueur », même si les avis divergent quant à l'emploi de ces termes (Piérard et Bestgen, 2007 ; Ho-Dac, 2007 ; Ho-Dac et Péry-Woodley, 2009).

[29] Nous définissons ces notions dans les sections suivantes.

[30] (Adam, 1990 : 51) utilise plutôt le terme de *segment* pour marquer le découpage textuel.

Dans les sections suivantes, nous proposons une présentation détaillée de différents indices de cohésion : les marques de paragraphe, les chaînes lexicales, les chaînes de référence et les cadres de discours.

3.1 Marques de paragraphe

Nombreux sont les moyens linguistiques présents dans les textes pour signaler les continuités ou ruptures thématiques. Un premier marqueur simple, notamment utilisé par (Passerault et Chenet, 1991) et (Masson, 1995), connaît un regain d'intérêt depuis quelques années (Hoey, 2005 ; Piérard et Bestgen, 2005 ; Fillipova et Strube, 2006) : le changement de paragraphe. C'est en effet par l'alinéa, « le prototype des marqueurs de segmentation à l'écrit » (Piérard et Bestgen, 2006b), que l'auteur d'un document peut signaler une discontinuité thématique (Hofmann, 1989 ; Longacre, 1979). C'est aussi *via* le paragraphe que l'on peut signaler l'unité thématique d'un texte (Goutsos, 1997). Le paragraphe est alors considéré comme une unité[31] « of speech or writing that maintains a uniform orientation » ((Hinds, 1977 : 136), cité par (Dooley, 2007)). (Bessonnat, 1988 : 94) va jusqu'à dire que l'« alinéa est à considérer comme un outil métatextuel, un « signe sur signe » au même titre que les connecteurs, la ponctuation, les reprises anaphoriques » vu qu'il structure le texte.

Les marques de paragraphe sont souvent associées à des marqueurs linguistiques pour signaler une rupture thématique : « paragraph breaks are signals of topic framing. […] they are found to co-occur with other topic signals. » (Goutsos, 1997 : 48)[32]. Afin de vérifier la relation établie entre le type de marqueur utilisé (article défini, indéfini, adjectif, possessif) et sa position dans le paragraphe orthographique (à l'intérieur du paragraphe ou au début du paragraphe), (Piérard et Bestgen, 2006b) ont utilisé une technique classique, le test du Chi² et le rapport de chance qui lui est associé. Le test du Chi² est une comparaison entre les résultats que l'on observe et les résultats théoriques auxquels on aurait pu s'attendre. Dans le cas des marqueurs, le test permet de vérifier s'il existe bien un lien entre le type de marqueur et sa position dans le paragraphe (*e.g.* le test permet de calculer la probabilité que ces deux événements apparaissent ensemble). La formule est la suivante :

[31] Dans cette approche, le paragraphe est une unité structurale et non une unité orthographique.

[32] (Bessonnat, 1988 : 89-90) a établi une liste des marques linguistiques privilégiées pour l'ouverture d'un paragraphe (indicateurs spatio-temporels, changements de noms propres, anaphoriques, connecteurs, reprises, rupture de temps verbaux, marqueurs de sériation) et la clôture d'un paragraphe (termes récapitulatifs, conclusifs, phrase d'appel, phrase-surprise, phrase de clôture, etc.).

$$Chi^2 = \sum \frac{(résultats\ observés - résultats\ attendus)^2}{résultats\ attendus}$$

Si les différences sont très faibles, il en résulte qu'il n'existe aucune relation entre les deux variables. Plus les différences sont importantes, plus la relation est forte entre les deux variables (type du marqueur et position du marqueur dans le paragraphe). L'indice du rapport des chances associé au Chi² permet les comparaisons entre expressions. On peut comparer, par exemple, le rapport entre la chance qu'une phrase contenant une expression temporelle arrive en tête de paragraphe par rapport à celle qu'une phrase ne contenant pas d'expression temporelle arrive en tête de paragraphe.

Les résultats obtenus par (Piérard et Bestgen, 2006b) ont confirmé leur hypothèse initiale : les marqueurs de continuité thématique (adjectifs possessifs et pronoms personnels) se situent plus souvent au sein d'un paragraphe alors que les marqueurs de discontinuité (adverbiaux cadratifs temporels tels que « En 2012, ») apparaissent plus fréquemment en début de paragraphe (voir Figure 6)[33].

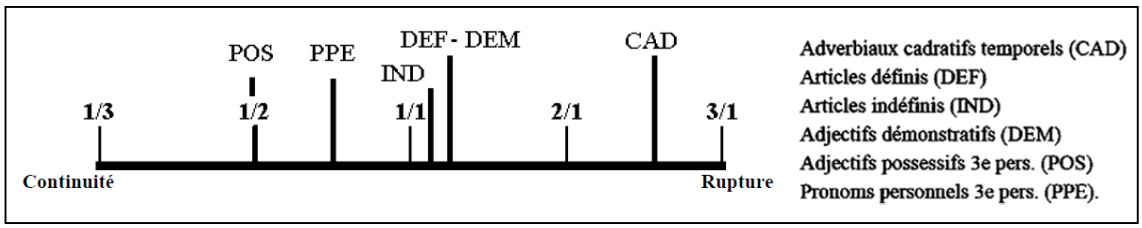


Figure 6 - Distribution des marqueurs selon le rapport des chances (d'après Piérard et Bestgen, 2006b)[34]

Plus spécifiquement, (Schnedecker, 1997, 2005) a montré que le nom propre coïncidait notamment avec le découpage en paragraphes. En effet, « l'alinéa signale au lecteur qu'il vient de traiter une unité de sens et qu'il va passer à une unité ultérieure » (Bessonnat, 1988 : 85), ce qui revient à dire que l'ouverture d'un nouveau paragraphe désactive le référent en cours. De ce fait, si l'on souhaite réinstancier le référent, l'utilisation d'un marqueur de faible accessibilité[35] cognitive tel que le nom propre se révèle nécessaire.

Néanmoins, comme le soulignent ((Brown et Yule, 1983), (Demol, 2010), (Stark, 1988) citée par (Piérard et Bestgen, 2006b)), les paragraphes remplissent d'autres

[33] Voir aussi (Ariel, 1990).

[34] Les valeurs de l'axe gradué s'interprètent de la manière suivante : par exemple, la valeur 3/1 indique qu'un marqueur a trois fois plus de chance d'apparaître en début de paragraphe qu'au milieu d'un paragraphe, la valeur 1/2 indique qu'un élément a deux fois plus de chance d'apparaître au milieu d'un paragraphe qu'au début d'un paragraphe, la valeur 1/1 indique qu'un marqueur apparaît aussi souvent au début d'un paragraphe qu'au milieu d'un paragraphe.

[35] Cette notion est présentée dans le chapitre 2.2.1.

fonctions discursives, comme la mise en relief d'un élément du texte. Il est donc nécessaire d'employer conjointement d'autres indices de continuité ou de rupture de thème, tels que les chaînes lexicales, identifiés par des études linguistiques.

3.2 Les chaînes lexicales

Formalisées par (Morris et Hirst, 1991), les chaînes lexicales sont des marques de cohésion textuelle (Halliday et Hasan, 1976). Elles permettent d'identifier des zones de continuité thématique

> « Often, lexical cohesion occurs not simply between pairs of words but over a succession of a number of nearby related words spanning a topical unit of the text. These sequences of related words will be called *lexical chains.* » (Morris et Hirst, 1991 : 22-23)

ou indiquent des zones de rupture thématique[36] : « When a lexical chain ends, there is a tendency for a linguistic segment to end, as the lexical chains tend to indicate the topicality of segment. » (Morris et Hirst, *ibid.* : 24).

Les chaînes lexicales sont des séquences de mots cooccurrents liés par une relation d'identité référentielle (répétition lexicale, *e.g.* « la décision… la décision »), d'hyponymie (*e.g.* « percheron » est un hyponyme de « cheval »), d'hyperonymie, de synonymie, de méronymie (partie-de, *e.g.* « jambe » est un méronyme de « corps »), d'holonymie ou de collocation (*e.g.* « maladie » … « patient » … « traitement ») ou simplement par une association d'idées (Hirst et St-Onge, 1998 ; Hoey, 1991 ; Hollingsworth et Teufel, 2005). Dans l'exemple ci-dessous, la chaîne lexicale compte 4 « maillons » ({bovins, femelles, animal, animal}) :

> les *bovins* malades sont abattus et totalement détruits, de même que, s'il s'agit de *femelles*, le dernier *animal* auquel elles ont donné naissance durant la période de deux ans ayant précédé ou durant la période ayant suivi l'apparition des premiers signes cliniques de la maladie, si *cet animal* est encore en vie dans le pays ou la région. (corpus *Acquis Communautaire*)

Les maillons des chaînes lexicales peuvent être contenus dans une même phrase, dépasser les frontières de la phrase et même s'étendre au-delà d'un paragraphe (Morris et Hirst, 1991 ; Stairmand, 1996). Les chaînes lexicales constituent ainsi des indices de la structuration textuelle :

[36] Les chaînes lexicales permettraient de délimiter des zones thématiques : « The chains contained in a text will tend to delineate the parts of the text that are "about" the same thing. » (Green, 1996 : 58).

> "Any structural theory of text must be concerned with identifying units of text that are about the same thing. When a unit of text is about the same thing there is a strong tendency for semantically related words to be used within that unit. By definition, lexical chains are chains of semantically related words. Therefore it makes sense to use them as clues to the structure of the text." (Morris et Hirst, 1991 : 35)

Spécifiquement pour les textes expositifs et argumentatifs, (Legallois, 2006) propose de fixer un seuil à partir duquel on peut considérer que les répétitions lexicales interviennent dans la structuration textuelle. Ce seuil de saillance est fixé à 3 répétitions :

> « Aussi est-il nécessaire d'établir un seuil à partir duquel la participation de la répétition dans l'organisation globale devient indiscutablement effective : nous considérerons que l'identification des phrases ayant aux moins trois lexèmes en commun (et non plus un ou deux) permet de constater très simplement qu'un mode d'organisation réticulaire est à l'œuvre dans certains types de textes. » (Legallois, 2006 : 56)

Ce seuil de trois éléments à partir duquel le phénomène de répétition serait bien « installé » dans le texte est aussi utilisé, nous le verrons plus loin, par (Schnedecker, 1997) pour définir la notion de chaîne de référence (*cf.* chapitre 2).

L'utilisation du procédé de répétition lexicale permettrait de thématiser un élément qui ne l'était pas jusqu'alors. Dans l'exemple suivant, c'est la répétition du syntagme nominal défini l'« évaluation » qui « fait monter le référent (ou un des référents) en position thématique » (Richard, 2000 : 79) :

> Les caractéristiques de l'opération susceptibles d'influer sur *l'évaluation des risques* doivent être prises en compte (opérations envisagées, méthodes de travail, échelle, mesures de confinement). *L'évaluation* doit tout spécialement examiner la question de l'élimination des déchets et des effluents. (corpus *Acquis Communautaire*)

La répétition lexicale est aussi un des moyens utilisés pour lever une ambiguïté référentielle (Morris et Hirst, 1991 ; Richard, 2000). En effet, dans l'exemple suivant, la substitution du syntagme nominal démonstratif « cette mesure » par le pronom[37] « elle » mènerait à une ambiguïté référentielle entre plusieurs candidats potentiels (« une réponse uniforme », « la place relative », « la mesure de la satisfaction » et « la décision publique ») :

[37] En suivant la loi de « l'intermède pronominal » (Schnedecker, 1997 : 119)

> Il n'existe pas une réponse uniforme sur la place relative que doit prendre *la mesure de la satisfaction* dans la décision publique. A tout le moins, peut-on considérer que *cette mesure* doit systématiquement être intégrée, à côté d'autres indicateurs fournis notamment par le contrôle de gestion, comme un élément d'aide à la décision. (corpus *La Documentation Française*)

La répétition lexicale permet donc de ré-instancier le référent sous une forme nominale (quasi) identique à celle qui a été utilisée pour l'introduire (Schnedecker, 1997)[38] et évite, ce faisant, l'ambiguïté.

En traitement automatique des langues, les chaînes lexicales et plus spécifiquement les relations de synonymie, d'hyponymie et d'hyperonymie sont couramment utilisées dans certaines méthodes linguistiques pour la résolution de diverses tâches telles que la recherche d'information, le résumé automatique, la détection de termes erronés, etc. (Morris et Hirst, 1991 ; Hirst et St-Onge, 1998 ; Kolla, 2002 ; Oliveira Santos, 2006 ; Jayarajan *et al.*, 2008 ; Eisenstein, 2009). D'autres méthodes statistiques utilisent les répétitions lexicales et les collocations (Hearst, 1997 ; Kan *et al.*, 1998 ; Choi, 2000 ; Brunn *et al.*, 2001 ; Chali, 2001 ; Utiyama et Isahara, 2001 ; Sitbon, 2004 ; Sitbon et Bellot, 2004) pour la segmentation thématique notamment (*cf.* chapitre 4). Ces méthodes s'appuient sur des ressources lexicales (lexiques, dictionnaires, thésaurus, p.e. Roget (Roget, 1977), ontologies, p.e. *WordNet* (Fellbaum, 1998)) pour déterminer la présence d'une relation de synonymie, d'hyperonymie, de collocation, etc. entre deux maillons potentiels d'une chaîne lexicale. Pour (Pincemin, 1999), l'utilisation de ces ressources restreint fortement le nombre des chaînes lexicales potentiellement repérées car « seules les chaînes lexicales prévues par le thésaurus peuvent être repérées, et le thésaurus fixe les interrelations entre les mots indépendamment des usages textuels » (Pincemin, 1999 : 84).

Néanmoins, bien que ces méthodes soient dépendantes de leurs ressources, l'utilisation importante des chaînes lexicales dans la littérature n'est plus à prouver (comme l'atteste l'abondance de références) et elle s'explique, d'une part, par la disponibilité de l'information lexicale contenue dans le texte et, d'autre part, par l'accessibilité des méthodes permettant de les exploiter (Hernandez, 2004)[39]. De plus, dans la plupart des méthodes, les chaînes lexicales sont utilisées en combinaison avec d'autres marqueurs de l'organisation du discours. Par exemple, pour la segmentation thématique, en plus des chaînes lexicales,

[38] (Schnedecker, 1997 : 32) préfère utiliser le terme de *redénomination* plutôt que *répétition* car « le terme de *répétition* rend compte de phénomènes très localisés et catégoriellement limités ».

[39] « It is important to realise that determining lexical chains is not a sophisticated natural language analysis process » (Carthy et Smeaton, 2000 : 2).

(Beeferman *et al.,* 1999) tirent aussi profit des connecteurs, (Passonneau et Litman, 1993) exploitent aussi les connecteurs et des indices prosodiques, (Ferret *et al.,* 2001 ; Vigier *et al.,* 2004) utilisent aussi les cadres de discours, (Hernandez, 2004) use aussi des chaînes anaphoriques. Plus précisément, (Hernandez, 2004) se sert d'un module de construction de chaînes lexicales (CCL, Construction de Chaînes Lexicales) pour améliorer l'identification automatique des thèmes effectuée par un premier module de reconnaissance de chaînes anaphoriques (SRA, Système de Reconnaissance d'Anaphores). Nous nous situons dans cette lignée. Dans notre projet, nous identifierons les relations anaphoriques et de coréférence et nous utiliserons les chaînes lexicales pour le versant statistique de notre approche afin de segmenter thématiquement nos documents *via* l'algorithme *C99* de (Choi et *al.*, 2001)[40] enrichi avec la LSA[41] (*cf.* chapitre 4). Nous rejoignons ainsi (Barzilay et Elhalad, 1997 : 11) :

> « Other indicators can be used to identify discourse structure as well (connectives, paragraph markers, tense shifts). Here again, merging discourse structure approaches with lexical cohesion techniques should lead to further improvement of the source text abstraction. »

3.3 Les chaînes de référence[42]

Les chaînes de référence constituent un système de marques de continuité du discours (Charolles, 1988 ; Schnedecker, 1997). (Chastain, 1975) propose une première définition de *chaîne de référence* comme une « séquence d'expressions singulières apparaissant dans un contexte tel que si l'une de ces expressions réfère à quelque chose, toutes les autres y réfèrent également. » ((Chastain, 1975 : 205), traduit par (Corblin, 1995b : 151)).

D'après (Marandin, 1988), la notion de *thème textuel* recoupe celle des chaînes de référence (« chaîne-objet » dans son approche) car les chaînes de référence permettent de modéliser les dépendances interprétatives établies entre les groupes nominaux d'un texte[43]. C'est en partie *via* les procédés de reprise référentielle (pronominale, démonstrative) que la lecture d'un texte est orientée vers une personne, un objet, un événement. Dans l'exemple suivant, le discours porte sur le

[40] *Cf.* chapitres 4 et 6.

[41] *Latent Semantic Analysis* (analyse sémantique latente).

[42] Les chaînes de référence feront l'objet des chapitres 2 et 3 du présent mémoire.

[43] (Marandin, 1988) a d'ailleurs élaboré la notion de *thème textuel* grâce à celle de *chaîne de référence*.

référent « M. James Carter » qui fait l'objet de reprises par le pronom « il » et le possessif « son » :

> Dans la course à la présidence, *M. James Carter* n'apparaît guère comme un candidat solide. *Il* a perdu au profit de M. Edward Kennedy le soutien d'Etats aussi traditionnellement acquis aux démocrates que ceux de New-York, de Pennsylvanie et de Californie, et *son* emprise sur le parti *s'*effrite chaque jour. (corpus *Le Monde Diplomatique*)

Certaines expressions référentielles contenues dans les chaînes de référence signalent plutôt une continuité thématique tandis que d'autres vont signaler un changement de thème (Fox, 1987 ; Kleiber, 1992 ; De Mulder, 1997 ; Charolles, 1997 ; Corblin, 1995b ; Schnedecker, 1997 ; Schnedecker, 2005). (Fox, 1987) montre que l'emploi d'un pronom de troisième personne signalerait que l'unité de discours est toujours en cours, alors que l'utilisation d'un nom propre complet (*e.g.* « François Hollande ») indiquerait que l'unité de discours est close (et signalerait par là-même l'ouverture d'une nouvelle unité) :

> « The first mention of a referent in a sequence is done with a full NP. After that, by using a pronoun the speaker displays an understanding that the preceding sequence has not been closed down. […] A full NP is used to display an understanding of the preceding sequence containing other mentions of the same referent as closed. » ((Fox, 1987 : 18-19), citée par (Cornish, 1989 : 239))

De même, dans les textes narratifs[44], (Schnedecker, 1997) montre que la redénomination *via* le nom propre indiquerait un changement de point de vue ou délimiterait des unités thématiques. La redénomination « signalerait qu'il va être dit, à propos du référent, des contenus de nature distincte. » (Schnedecker, 1997 : 186). La redénomination du nom propre permettrait ainsi de délimiter les différentes facettes d'un référent. Par exemple, dans :

> En réfléchissant et en *se* promenant, *Athos*$_1$ passait et repassait devant le tuyau du poêle rompu par la moitié et dont l'autre extrémité donnait dans la chambre supérieure, et à chaque fois qu'*il* passait et repassait, *il* entendait un murmure de paroles qui finit par fixer *son* attention. *Athos*$_2$ *s'*approcha, et *il* distingua quelques mots qui *lui* parurent sans doute mériter un si grand intérêt qu'*il* fit signe à *ses* compagnons de se taire, restant *lui*-même courbé l'oreille tendue à la hauteur de l'orifice inférieur. (corpus *Les Trois Mousquetaires*)

[44] Dans les textes narratifs (romans, portraits littéraires, faits divers, textes journalistiques, etc.), le matériau linguistique des chaînes de référence est souvent homogène (Schnedecker, 1997), c'est-à-dire constitué essentiellement de noms propres et de pronoms.

les deux occurrences d'« Athos » délimitent deux facettes à propos du référent, à savoir sa réflexion interrompue par une conversation dans la chambre supérieure et l'attention qu'il porte à cette conversation. De ce fait, deux chaînes de référence sont identifiées :

- chaîne 1 : [Athos… il…il…son]
- chaîne 2 : [Athos…s'…il…lui…il…ses…lui]

D'autres formes semblent dédiées à la clôture de chaînes de référence comme le rapporte (Marandin, 1988) qui utilise la notion de « maillon-fermoir » pour désigner les groupes nominaux démonstratifs (*e.g.* « cette altercation ») doué d'une capacité à résumer (Guillot, 2006) et clore un segment textuel et la répétition de groupes nominaux pleins (*e.g.* « l'altercation… l'altercation ») (voir aussi (Schnedecker, 1997)).

Aux vues des premières caractéristiques des chaînes de référence dégagées (nous étudierons en détail les caractéristiques des chaînes de référence dans les chapitres suivants), l'utilisation des chaînes de référence comme indice de continuité et discontinuité thématique apparaît être un moyen efficace pour participer à la détection des thèmes des documents.

3.4 Les cadres de discours

De même que les chaînes de référence, les cadres de discours (Charolles, 1997) sont des marques d'organisation des discours (Charolles, 2009). Ce sont des

> « unités de segmentation originales venant s'ajouter aux unités typo-dispositionnelles (paragraphe, tiret, puces,...) qui sont des sortes de cadres sous-spécifiés sémantiquement (le critère de regroupement des propositions n'étant pas signalé sauf quand il y a titraison) et ils contribuent de ce fait à l'organisation et donc à la cohésion du discours. » (Charolles et Péry-Woodley, 2005 : 5).

Les cadres ont la capacité de recouvrir dans leur portée sémantique plusieurs propositions (Porhiel, 2006). Ils permettent d'assurer un suivi dans le déroulement du texte (Charolles, 2003). Ainsi, ils « contribuent à subdiviser et répartir les informations apportées par le discours au fur et à mesure de son développement » (Charolles, 1997 : 33) pour faciliter la compréhension. Ce sont donc des guides dotés d'une fonction procédurale et cognitive (Charolles, *ibid.*).

Dans l'hypothèse de l'encadrement du discours, (Charolles, 1997) décrit les cadres de discours comme des segments homogènes par rapport à un critère sémantique

(une localisation spatiale par exemple) spécifié par une expression détachée à l'initiale de la phrase dite *introducteur de cadre*[45]. L'exemple suivant de cadre introduit par une expression temporelle (en italique) invite à interpréter l'ensemble du segment relativement à la date qu'elle définit :

> *Il y a un an à peu près*, qu'en faisant à la Bibliothèque royale des recherches pour mon histoire de Louis XIV, je tombai par hasard sur les Mémoires de M. d'Artagnan, imprimés à Amsterdam, chez Pierre Rouge. Le titre me séduisit : je les emportai chez moi, avec la permission de M. le conservateur, bien entendu, je les dévorai. (corpus *Les Trois Mousquetaires*)

Dans cet exemple, l'introducteur de cadre temporel (« il y a un an à peu près »), détaché en tête de la première phrase, indexe les autres phrases de l'exemple (rôle cadratif (Charolles et Vigier, 2005)). La portée sémantique de l'introducteur de cadre peut donc dépasser la phrase dans laquelle il est situé. Ainsi, les cadres de discours peuvent soit être contenus dans un paragraphe, soit le recouvrir et même regrouper plusieurs paragraphes[46] (Charolles, 1997 ; Charolles, 2003 ; Charolles et Prévost, 2003 ; Charolles, 2009).

(Porhiel, 2004 : 9) définit les cadres comme des « mondes dans lesquels il faut envisager un certain état ou une série d'évènements ». C'est en effet relativement à une certaine portion temporelle ou spatiale par exemple, que l'affirmation pourra s'appliquer. Dans

> *Dimanche 25 septembre*, Valéry Giscard d'Estaing a défendu le principe de la présence d'un candidat de l'UDF dans la compétition présidentielle prévue pour 1995. (corpus *Le Monde*),

c'est relativement à la date précise du 25 septembre 1994 que Valéry Giscard d'Estaing a défendu ce principe.

La notion de cadre de discours recouvre quatre constructions : les univers de discours, les champs thématiques, les domaines qualitatifs (adverbiaux de manière) et les espaces de discours.

- *les univers de discours* délimitent un espace de véridiction (Schnedecker, 2001) où se situent les propos énoncés. Ils constituent, d'après ((Martin, 1983 : 37) cité par (Charolles, 1997)), l'« ensemble des circonstances, souvent spécifiées sous forme d'adverbes de phrase,

[45] (Porhiel, 2004) considère les introducteurs de cadres comme des *balises* permettant de mettre en évidence l'intention informationnelle du locuteur.

[46] (Charolles, 1997) remarque néanmoins que le changement de paragraphe crée une frontière qu'un cadre de discours arrive difficilement à franchir.

dans lesquelles la proposition peut être dite vraie. ». (Charolles, *ibid.* : 25) propose une liste (non exhaustive) d'introducteurs d'univers de discours : *sauf erreur, sauf exception, à la limite, par défaut, à vrai dire, en vérité, blague à part, entre nous, à mon avis, selon X, d'habitude, le plus souvent.* Dans l'exemple suivant, la dispense de l'application des dispositions de la directive ne peut être vraie qu'au regard de la procédure prévue à l'article 21 :

> *Selon la procédure prévue à l'article 21*, un État membre peut, à sa demande, être totalement ou partiellement dispensé de l'application des dispositions de la présente directive pour certaines espèces s'il n'existe normalement pas de reproduction et de commercialisation des semences de ces espèces sur son territoire. (corpus *Acquis Communautaire*)

A ces introducteurs relatifs au degré de certitude du locuteur s'ajoutent des introducteurs liés au cadre temporel et/ou spatial qui vont partitionner l'information, tels que :

> *Dans les cultures destinées à la production des matériels de multiplication de base*, les viroses nuisibles, notamment le court-noué et l'enroulement, doivent être éliminés. Les cultures destinées à la production de matériels de multiplication des autres catégories sont maintenues exemptes de plantes présentant des symptômes de viroses nuisibles. (corpus *Acquis Communautaire*)

Dans l'exemple ci-dessus, l'univers spatial spécifié (*e.g.* « dans les cultures destinées à la production des matériels de multiplication de base ») instaure une relation de contraste entre la proposition jugée vraie (*i.e* qui se vérifie) dans ce cadre, soit « les viroses nuisibles, notamment le court-noué et l'enroulement, doivent être éliminés » et toutes les autres circonstances dans lesquelles cette proposition ne serait pas vraie ((Charolles, *ibid.*) nomme ces autres circonstances des *univers parents*, virtuels). Dans notre exemple, la deuxième phrase rapporte bien l'opposition initiée par le cadre, par l'intermédiaire de l'adjectif « autre » (en gras).

– dans les *champs thématiques* (ou cadres thématiques[47]), (Charolles, *ibid.*) signale des expressions introductrices comme *à propos de X, au*

[47] (Bessonnat, 1988) appelle « marqueurs de sériation » les introducteurs de cadre tels que « en ce qui concerne…, pour ce qui est de…, quant à » ; de son côté (Adam, 2011 : 119-120) parle de marqueur de changement de topicalisation : « Le passage d'un objet du discours à un autre est souvent souligné par des marqueurs de changement de topicalisation comme *quant à* ou *en ce qui concerne.* ».

sujet de X, concernant X, pour ce qui est de X, quant à X, etc. Par exemple :

> *Concernant* la génération des vingt-cinq-trente-cinq ans - la génération d' « après 68 », - cette tendance est marquée par un désir très fort de travailler différemment. (corpus *Le Monde Diplomatique*)

Selon (Charolles, *ibid.*), les expressions détachées de ce type se distinguent des introducteurs à portée véridictionnelle (*i.e.* les univers de discours) parce qu'elles ne vont plus définir des critères de vérité mais plutôt introduire ce qui sera le thème du segment textuel :

> « Ce que marque au premier chef des formules comme *à propos de X, au sujet de X, concernant X, pour ce qui est de X* c'est la volonté du locuteur de signaler que, au moins pour un temps, ce qu'il va dire porte sur X (et non sur Y ou Z), a pour objet X, bref, que le thème de son propos va être X. » (Charolles, 1997 : 26)

Ainsi, dans l'exemple suivant, l'introducteur de cadre thématique « pour ce qui est de » met en exergue le thème du segment « les résultats des « visites mystères » :

> *Pour ce qui est des* **résultats des « visites mystères »**, chaque Bureau de Poste dispose de ses propres résultats et d'un système de comparaison aux meilleurs. (corpus *La Documentation Française*)

D'après (Porhiel, 2004 : 8), « les introducteurs thématiques sont des moyens cohésifs doublement thématiques : ils sont en position détachée, le plus souvent en tête de phrase et ils jouent aussi un rôle thématique » car ils réintroduisent des personnes, situations, ou événements déjà introduits dans le discours (*i.e.* le référent est connu).

De ce fait, les champs thématiques sont à la fois des thèmes phrastiques (par leur position initiale occupée dans la phrase) et des thèmes textuels car ils annoncent un thème qui va occuper une portion textuelle plus ou moins large (Porhiel, 2005b). Les champs thématiques signalent qu'un référent déjà présent dans le discours va occuper l'attention, que ce référent va donc devenir saillant dans la suite du discours. En cela, ce sont des éléments de reprise (Porhiel, 2005b) ; reprise marquée notamment par la présence fréquente de déterminants définis devant les compléments nominaux, comme par exemple :

> *En ce qui concerne* les **vignes-mères destinées à la production de matériel de multiplication standard**, la proportion de pieds

> manquants imputable aux organismes nuisibles visés aux points 5 a) et 5 b) ne doit pas dépasser 10 %. (corpus *Acquis Communautaire*)

Aussi, l'utilisation d'un introducteur de cadre thématique permet de sélectionner un référent parmi un ensemble, comme dans :

> Des dispositions distinctes doivent être établies pour les régimes d'aides « surfaces » et les régimes d'aides « animaux », compte tenu de la nature différente de ces deux régimes d'aides. Les dispositions relatives aux réductions et exclusions doivent tenir compte des particularités des différents régimes d'aide relevant du système intégré. *En ce qui concerne* **les demandes d'aide « surfaces »,** les irrégularités affectent en général une partie des surfaces et les surdéclarations concernant une parcelle peuvent être compensées avec les sous-déclarations d'autres parcelles du même groupe de culture. *En ce qui concerne* **les demandes d'aide « animaux »,** les irrégularités entraînent l'inéligibilité de l'animal concerné. (corpus *Acquis Communautaire*)

Dans l'exemple ci-dessus, la phrase introductrice contient l'ensemble des référents « aide surfaces » et « aide animaux ». Ces deux référents sont ensuite réintroduits l'un après l'autre *via* l'introducteur thématique *en ce qui concerne*.

– les *domaines qualitatifs* regroupent des adverbes tels que *par hasard, prudemment,* des expressions comme *en dépit de X, à l'insu de X,* des constructions absolues du type *pieds nus, les yeux fermés* ou des participiales préfixées telles que *troublé par l'arrivée d'une inconnue, rougie par le soleil, etc.* Il s'agit de cadres introduits par des expressions précisant les aspects qualitatifs des états de choses dénotés comme le but ou les motivations d'un participant au procès. Dans l'exemple ci-dessous, l'expression à l'initiale ne partitionne pas l'information comme c'est le cas pour les introducteurs d'univers de discours ou de champ thématique, mais est liée à la situation :

  > *Pour délibérer valablement*, le Comité consultatif doit réunir au moins la moitié de ses membres. (corpus *Acquis Communautaire*)

(Charolles, 1997) précise que, de même que les introducteurs de champs thématiques, les introducteurs de domaines qualitatifs ont en général une portée limitée à leur phrase d'accueil et entretiennent une relation étroite avec le texte qui précède car ils font souvent référence à des individus ou des situations déjà évoqués.

– les *espaces de discours* regroupent des informations axées sur les aspects métalinguistiques de l'énonciation (p.e. ***En somme**, pour déterminer si elle a juridiction sur une entité particulière « à but non lucratif » gérant un système d'agrément, la Commission fédérale du commerce doit examiner concrètement dans quelle mesure une telle entité permet à ses membres de faire des bénéfices.* (corpus *Acquis Communautaire*) ou de leur disposition dans le texte. Ces dernières expressions se divisent en quatre sous-groupes (Ho-Dac, 1999) :

  a. les espaces d'ouverture (p.e. ***Tout d'abord**, le principe d'une conférence « globalisante », avec un document final engageant l'ensemble des P.C., n'existe plus depuis 1969, date à laquelle s'est réunie à Moscou la dernière conférence mondiale.* (corpus *Le Monde Diplomatique*)),

  b. les espaces d'argumentation (p.e. ***De plus**, la partie relative à la police de proximité, insiste sur la nécessaire complémentarité entre les indicateurs de résultats et les indicateurs de qualité, ainsi qu'entre les indicateurs élaborés en interne et ceux issus d'appréciations externes, tels les sondages d'opinion.* (corpus *La Documentation Française*)),

  c. les espaces de conclusion (p.e. ***En un mot**, nous ne parvenons plus bien à percevoir le lien entre ce que nous faisons et la transformation des choses.* (corpus *Acquis Communautaire*))

  d. et les espaces corrélés (p.e. ***D'une part**, la plupart des pays appliquent la méthode de stratification qui consiste à combiner le parc immobilier total ventilé selon diverses strates avec l'information sur le loyer réel payé dans chaque strate. **D'autre part**, quelques pays appliquent la méthode de l'auto-estimation pour les logements occupés par leurs propriétaires, en demandant aux propriétaires qui occupent leur logement d'estimer le prix auquel il pourrait être loué.* (corpus *Acquis Communautaire*)).

  De même que les cadres temporels et spatiaux, les espaces de discours contribuent au partitionnement du discours : ils découpent le discours en espaces homogènes suivant un critère dispositionnel (Charolles, 1997).

Dans un discours peuvent se trouver un ou plusieurs cadres. Deux cadres de discours adjacents peuvent entretenir deux types de relations (Charolles, *ibid.*) :

– la *coordination*, lorsque l'ouverture d'un nouveau cadre entraîne la fermeture du cadre précédent. Dans l'exemple ci-dessous, l'ouverture du cadre temporel « En avril » ferme le cadre temporel « En janvier » :

> La progression du maire de Paris a également été mesurée par la SOFRES dans son « baromètre » des personnalités que les personnes interrogées souhaitent voir jouer « un rôle important au cours des mois et des années à venir ». *En janvier*, il n'obtenait que 37 % et arrivait en septième position à droite. *En avril*, M. Chirac est quatrième, après M.

> Balladur (56 %), Simone Veil (52 %) et Charles Pasqua (52 %), avec 41 %. (corpus *Le Monde*)

– la *subordination*, lorsque le nouveau cadre s'ouvre dans le précédent, comme dans

> *Chez les bovins*, l'ouverture du cœur et l'incision des ganglions lymphatiques de la tête ne doivent être pratiquées qu'en cas de doute. *Chez la vache*, les mamelles sont ouvertes par une longue et profonde incision jusqu'aux sinus galactophores (sinus lactifères). (corpus *Acquis Communautaire*)

Dans ce dernier exemple, la portée du premier cadre « chez les bovins » dépasse sa phrase d'accueil et s'étend jusqu'à la fin de l'extrait. Le second cadre s'« emboîte » dans le premier et a une portée limitée.

Lorsqu'une expression introductrice de cadre succède à une autre, le choix entre la coordination et la subordination est fonction des connaissances de l'interprétant. Ainsi, ce sont nos connaissances sur le monde qui nous permettent de savoir que « vache » est un hyponyme de « bovin » dans l'exemple ci-dessus. Dans une approche automatique à base de peu de connaissances, le repérage de la portée des introducteurs de cadres constitue un problème important (Bilhaut, 2006 ; Couto, 2006 ; Laignelet, 2009). Par exemple, l'utilisation d'heuristiques telles que « si un nouveau cadre est ouvert, le cadre en cours doit être fermé » n'est pas satisfaisante puisque, comme nous venons de le voir, une relation générique peut s'établir entre les cadres. L'identification automatique des cadres doit tenir compte de ces considérations pour être efficace et peut s'effectuer en combinaison avec d'autres types de marqueurs tels que les marques de paragraphes (*i.e.* l'ouverture d'un cadre de discours coïncide souvent avec le début d'un paragraphe vu que le cadre résiste difficilement à l'alinéa), certains connecteurs, les démonstratifs ou les chaînes de référence (*i.e.* l'ouverture d'un cadre de discours peut coïncider avec l'ouverture d'une chaîne de référence *via* un nom propre par exemple). Néanmoins, même si le problème lié à la portée des cadres est complexe, la notion d'introducteur de cadre dans l'hypothèse de l'encadrement du discours se prête totalement à un traitement automatique.

De ce fait, dans notre approche, parmi les différents types de constructions, nous choisissons d'identifier automatiquement les champs thématiques, les espaces de discours ainsi que les domaines qualitatifs. Nous nous focalisons sur ces trois types de cadre de discours car leur

fonction consiste à introduire le thème du segment ou à apporter des précisions « relatives à ce thème ».

3.5 Bilan

Pour signaler les continuités ou les ruptures thématiques, nous avons vu que plusieurs dispositifs de cohésion étaient disponibles. (Charolles, 1997) regroupe ces marques de cohésion en deux[48] sous-ensembles cumulables :

- les expressions mettant en évidence que deux unités proches doivent être reliées puisque ces unités portent sur le même référent (anaphore, chaîne lexicale, chaîne de référence) ou ont une intention et un contenu propositionnel similaire (connecteurs),
- les expressions indiquant qu'une série de propositions doivent être regroupées en unités (paragraphes, cadres) car ces propositions entretiennent le même rapport avec un certain critère.

D'après (Charolles, 2003), les adverbiaux cadratifs sont, dans une certaine mesure, liés aux chaînes de référence car ces deux types de marqueurs concernent un même objet du discours (pouvoir d'indexation). Ainsi, faisons-nous le choix de cumuler les marqueurs référentiels aux cadratifs pour identifier les thèmes des documents[49]. La collocation de ces deux marqueurs permettrait de confirmer la présence d'un thème persistant (Givón, 1983) par « renforcement d'indices convergents » (Charolles, 1997 : 51).

[48] Notons que dans l'introduction de cette section, nous avons présenté la décomposition en quatre plans d'organisation textuelle de (Charolles, 1988).

[49] Dans notre système de détection automatique de thèmes, l'identification de ces deux types de marqueurs de cohésion constitue le versant linguistique de notre méthode. Les chaînes lexicales seront utilisées par le versant statistique (voir chapitres 4 et 6).

4 Conclusion

La notion de *thème* est une notion couramment utilisée mais elle n'est toujours pas clairement définie à l'heure actuelle. Ainsi, la citation d'Augustin sur le temps, peut-elle être appliquée au thème : « Qu'est-ce donc que le temps ? Quand personne ne me le demande, je le sais ; dès qu'il s'agit de l'expliquer, je ne le sais plus. » (Augustin, *Confessions*, 11 : 17). (Schnedecker, 1997 : 64, note 24) synthétise les critiques formulées à l'égard de la notion de thème en quatre points :

- **un bougé terminologique et conceptuel** (Cadiot et Fradin, 1988), lié aux diverses disciplines et écoles qui ont utilisé cette notion sans la définir explicitement,
- **l'amplitude du phénomène** : à la diversité des niveaux d'analyse (local, global, phrastique, discursif) ne correspond pas une définition propre (Kleiber, 1992),
- **l'hétérogénéité des critères définitoires**, issus de la syntaxe, de la sémantique, de l'ordre des mots d'un système linguistique, de la nature des expressions référentielles,
- **la circularité** (Kleiber, 1992) : utilisation d'un anaphorique pour identifier le thème du discours, au lieu de le définir.

Comme nous avons pu le voir, les différentes approches linguistiques permettent de distinguer deux conceptions du thème : la première considère le thème dans une perspective phrastique (locale), la seconde l'aborde dans une perspective textuelle (plus globale). Les travaux de van Dijk, par exemple, établissent une distinction entre les niveaux microstructurel (le contenu thématique d'un énoncé isolé) et macrostructurel (le contenu thématique qui émerge d'une structure plus complexe composée d'un ensemble d'énoncés reliés entre eux) présents dans un texte. En nous appuyant sur cette distinction, nous ne nous attarderons pas à considérer le thème dans une dimension phrastique car notre objectif consiste à identifier les différents thèmes globaux décrivant les documents. Dans notre approche automatique, nous suivrons l'approche de (Goutsos, 1997) qui utilise des marqueurs linguistiques pour identifier les zones de rupture et de continuité thématiques. Nous définissons donc le thème textuel comme un agrégat des différents thèmes phrastiques, le thème textuel étant identifié *via* des marqueurs linguistiques utilisés seuls ou en combinaison. Ainsi, est considéré comme thème

local à une phrase, le thème phrastique ; le thème *global* du paragraphe constituant l'agrégation des thèmes locaux.

Pour signaler la structure thématique du discours, nous avons vu que plusieurs marqueurs de cohésion étaient spécialisés dans cette fonction. En suivant (Goutsos, 1997 ; Piérard et Bestgen, 2006b), dans notre perspective de traitement automatique, il nous semble nécessaire de cumuler des indices de continuité et de rupture thématique pour mener à la détection automatique des thèmes que nous visons[50]. De ce fait, pour le versant linguistique de notre méthode, nous avons sélectionné deux indices de cohésion fiables pour participer à la détection des thèmes : les cadres de discours (plutôt tournés vers l'aval du discours) et les chaînes de référence (tournées, elles, vers l'amont (Charolles, 2003))[51]. Dans le chapitre suivant, nous focaliserons notre attention sur les chaînes de référence et leur lien avec les thèmes.

[50] En effet, en cumulant les indices nous multiplions les chances d'identifier les thèmes des documents (car un type d'indice peut ne pas être présent dans une portion de texte, ou le texte tout entier).

[51] Voir le chapitre 6 pour l'utilisation de ces marqueurs pour la détection automatique des thèmes textuels.

Chapitre 2

Thèmes et chaînes de référence

Un texte n'est pas qu'une suite linéaire de phrases. En effet, le texte est organisé en segments localement cohérents regroupant des ensembles de phrases ; ces segments étant reliés entre eux *via* la cohérence globale (Charolles, 1988, 1995 ; Bestgen, 2006) et, en surface, par des marques linguistiques de cohésion (Charolles, 1988, 1995). Pour signaler et faciliter l'identification de la structure thématique du discours, plusieurs dispositifs linguistiques tels que les marqueurs cohésifs, les adverbiaux cadratifs ou les expressions référentielles sont utilisés (comme nous avons pu le voir dans le chapitre 1).

Les thèmes textuels (Givón, 1983 ; Marandin, 1988 ; Chafe, 1994 ; Lambrecht, 1994) s'expriment au moyen d'indices référentiels constituant les chaînes de référence. Les chaînes de référence sont constituées par un type de marques

indiquant la cohésion discursive : les expressions coréférentielles d'un texte (Corblin, 1985 ; Charolles, 1988), par exemple les pronoms (« il »), les noms propres (« Barack Obama »), les groupes nominaux (« le colloque international »), les groupes nominaux démonstratifs (« ce colloque »), les possessifs (« son »). Ces expressions référentielles, connectées référentiellement (Corblin, 1995b), constituent une certaine unité, un « ensemble de marques linguistiques qui font système » (Schnedecker, 2002). Les pronoms, par exemple, sont utilisés dans des cas de continuité thématique. Par contre, l'utilisation de groupes nominaux alors que l'accessibilité à l'antécédent demeure forte, indique une rupture de thème (Asher *et al.*, 2006).

Cette relation entre thèmes et chaînes de référence est montrée par (Charolles et Prévost, 2003 : 7) :

> « les expressions référentielles et anaphoriques codent le degré de proéminence des entités dans le modèle mental des locuteurs et interlocuteurs. De même, si on envisage la relation thème/topique – commentaire, l'idée que celui-ci puisse être « à propos de » ne se comprend bien qu'au sujet de quelque chose, autrement dit, là encore d'un référent, qui plus est cognitivement accessible ».

Les liens anaphoriques ou de co-référence (Kleiber, 1994) demeurent jusqu'à présent peu exploités pour la détection automatique des thèmes, malgré leur contribution à l'organisation textuelle.

Nous faisons l'hypothèse que les chaînes de référence représentent des éléments linguistiques fiables pour participer à la détection des thèmes textuels. En effet, les liens référentiels permettent au lecteur de se focaliser sur un référent principal, qui peut constituer par ailleurs le thème du paragraphe. Le participant le plus souvent mentionné au niveau du paragraphe thématique et ultérieurement au niveau du discours constitue le « thème continu » (Givón, 1983 : 7). Pour apprécier la continuité référentielle, (Givón, 1983) a proposé trois mesures : la distance référentielle, les interférences potentielles entre référents (ambiguïté, compétition référentielle) et la persistance qui mesure la durée pendant laquelle l'entité est maintenue dans le discours après avoir été introduite la première fois. Ainsi, la distance, la fréquence et la manière dont une entité est reprise reflètent-elles les intentions du locuteur sur l'importance de cette entité dans le discours.

Dans ce chapitre, après avoir défini la notion de *chaînes de référence* que nous suivons, nous dégageons quelques caractéristiques des chaînes de référence. Nous présentons ensuite plusieurs théories (accessibilité, centrage, optimalité) relatives à l'emploi des expressions référentielles contenues dans les chaînes de référence pour établir, maintenir ou rompre la référence en discours.

1 Les chaînes de référence (CR)

1.1 Définition des CR

Les chaînes de référence[1] (Chastain, 1975 ; Corblin, 1985, 1987, 1990, 1995a ; Charolles, 1988 ; Schnedecker, 1997, 2005), terme métaphorique permettant de rendre compte de la proximité sémantique établie entre les diverses formes renvoyant à un même référent, constituent « la suite des expressions d'un texte entre lesquelles l'interprétation construit une relation d'identité référentielle » (Corblin, 1985 : 123). Par exemple (les maillons de la chaîne de référence sont en gras) :

> **François Hollande** sait que **ses** débuts seront déterminants pour apporter la preuve qu'**il** était « prêt » pour la fonction et imprimer **sa** marque. (*Les Echos*, 09/05/12)

Les chaînes de référence sont constituées d'expressions référentielles référant à la même entité du discours (ces expressions référentielles sont dites « coréférentes ») (Charolles, 1988) qui créent un continuum. Le référent instancié peut être un humain, un événement ou une entité abstraite :

> « Les chaînes sont constituées par des suites d'expressions coréférentielles […]. Seules peuvent appartenir (donner lieu à) une chaîne des expressions employées référentiellement, c'est-à-dire toutes et rien que les expressions nominales (ou pronominales) permettant d'identifier un individu (un objet de discours) *quelle que soit sa forme d'existence (personne humaine, événement, entité abstraite)* » (Charolles, 1988 : 8).

Les chaînes de référence sont un des facteurs qui contribuent à la cohésion textuelle du fait qu'elles apparaissent en continu pour le même référent. Certaines expressions référentielles sont plus cohésives que d'autres (Corblin, 1995b ; Schnedecker, 2005). Par exemple, les pronoms et les groupes nominaux définis simples renforcent la cohésion car ils indiquent que leurs propositions d'occurrences se situent dans le prolongement situationnel l'une de l'autre (*e.g.*

[1] On trouve dans la littérature plusieurs expressions équivalentes à « chaîne de référence » : *chaînage, chaîne anaphorique, continuité* (plus vague), *chaîne référentielle* (Halliday et Hasan, 1976 ; Kleiber, 1989), *chaînes de référence naturelles* (Corblin, 1995a ; 2005), *chaîne indexicale*, *chaîne topicale* (Cornish, 1998), *chaîne-objet* (Marandin, 1988).

« Le conseil d'administration, … il… il… ») (Corblin, 1985 ; Kleiber, 1986, 1994). D'autres expressions référentielles comme les noms propres et les démonstratifs sont moins cohésives et vont plutôt indiquer une rupture (Corblin, 1995b).

1.2 Anaphore et coréférence

Certaines expressions référentielles ne sont pas autonomes : elles nécessitent un support pour être interprétées correctement. C'est le cas des pronoms anaphoriques « il », « les » ou « leurs » (en gras) et du possessif « leur » (en italique) par exemple, dans

> **il les** arrêta pour **leur** faire compliment sur *leur* équipage, ce qui en un instant amena autour d'eux quelques centaines de badauds. (corpus *Les Trois Mousquetaires*)

qui nécessitent, pour être interprétés, des informations issues du contexte, à savoir « les quatre amis » et « M. de Tréville » :

> Près du Louvre les quatre amis rencontrèrent M. de Tréville qui revenait de Saint-Germain ;

Ainsi, il y a « relation d'anaphore entre deux unités A et B quand l'interprétation de B dépend crucialement de l'existence de A, au point que l'on peut dire que l'unité B n'est interprétable que dans la mesure où elle reprend – entièrement ou partiellement – A. » (Milner, 1982 : 18).

Lorsque plusieurs expressions référentielles renvoient au même référent, elles sont dites **coréférentielles**. La coréférence peut s'établir par le moyen d'anaphores (*e.g.* [Paul, il]) ou sans (*e.g.* [Le président de la république française, F. Hollande]). Par exemple, dans le texte suivant, « Jacques Chirac » et « le président du RPR » s'interprètent indépendamment l'une de l'autre (les expressions coréférentes sont en gras) :

> "Le moment n'est pas venu de répondre à cette question", a indiqué, lundi 3 janvier, sur Europe 1, **Jacques Chirac**, interrogé sur la date de sa déclaration de candidature pour l'élection présidentielle. Préférant parler de "la" campagne présidentielle, en général, plutôt que de sa propre campagne, **le président du RPR** a apporté son soutien au "gouvernement", qui "fait le maximum pour redresser la situation", tout en critiquant sur certains points, implicitement, l'action d'Edouard Balladur. (corpus *Le Monde*)

Les chaînes de référence peuvent contenir à la fois des relations de coréférence anaphoriques (*e.g.* [Paul, il]) et des relations de coréférence non anaphoriques (*e.g.* [Jacques Chirac, le président du RPR] de l'exemple ci-dessus).

1.3 Caractéristiques des CR

En suivant (Schnedecker, 1997), les chaînes de référence possèdent plusieurs caractéristiques :

- elles sont bornées : le nom propre est, par hypothèse, l'une de ces bornes car il sert à désigner le référent comme en première mention et de ce fait introduit une forme de rupture dans la chaîne (Schnedecker, 1997 : 2) :

  > « […] les chaînes de référence sont bornées et leurs bornes sont internes à l'espace référentiel lui-même, au sens où ces bornes sont constituées par des expressions référentielles, d'une part, et d'autre part, par la réitération, à un point donné, du nom ayant servi à introduire le référent (*i.e.* la borne initiale). […] Cette hypothèse d'un fonctionnement original des chaînes, grâce à un système d'auto-délimitation, permet également de définir plus précisément les modalités suivant lesquelles elles interagissent avec d'autres plans de l'organisation textuelle. »

- la notion s'applique à un minimum de trois expressions référentielles[2], ce qui la distingue de celle d'***anaphore*** (relation de dépendance entre deux expressions référentielles) ou de ***coréférence*** (suite de toutes les expressions référentielles d'un texte se rapportant au même référent). Les chaînes de référence peuvent contenir dans leurs maillons des expressions référentielles coréférentes et/ou anaphoriques :

  > « les chaînes de référence combinent des liens linguistiquement fondés (liens anaphoriques, au sens large) et des liens fondés sur les inférences autorisées par les connaissances empiriques partagées par le locuteur et le récepteur (liens communicatifs[3]). » (Corblin, 1995a : 2)

- concernant la distance inter-maillonnaire (distance entre les maillons d'une chaîne de référence) : plus les maillons d'une chaîne sont proches, plus la

[2] D'après (Corblin, 1995a : 177), la notion de chaîne « permet de dépasser les contextes de simple succession de deux termes auxquels se limite le plus souvent le linguiste qui sort du domaine phrastique, et ne préjuge pas de la nature des relations dont on verra qu'elles sont en fait hétérogènes. »

[3] *i.e.* la coréférence.

chaîne de référence possède de la substance et de la solidité. C'est le cas notamment dans les textes narratifs, comme l'exemple ci-dessous, où les maillons de la chaîne « Milady » (en gras) sont nombreux et proches les uns des autres :

> **Milady se** releva de toute **sa** hauteur et **Ø** voulut parler, mais les forces **lui** manquèrent ; **elle** sentit qu'une main puissante et implacable **la** saisissait par les cheveux et **l'**entraînait aussi irrévocablement que la fatalité entraîne l'homme : **elle** ne tenta donc pas même de faire résistance et **Ø** sortit de la chaumière. (corpus *Les Trois Mousquetaires*)

– Les chaînes de référence interagissent avec d'autres plans d'organisation textuelle tels que la portée[4] (portion de texte dont l'interprétation dépend d'un cadre ou espace de véridiction, par exemple un cadre temporel « en 2012, » (Charolles, 1988)), la période, les séquences (*cf.* chapitre 1 section 3).

Parallèlement, si l'on envisage d'identifier de manière automatique les chaînes de référence, d'autres caractéristiques des chaînes de référence sont à prendre en compte.

– Concernant leur **substance** :
  a. elles peuvent être *homogènes*, c'est-à-dire se composer de formes relevant de catégories d'expressions référentielles en nombre réduit (essentiellement des noms propres et des pronoms). (Ariel, 1990) relève des proportions de l'ordre de 70% pour les pronoms, qu'elle définit comme « la forme non marquée de la reprise » et que (Schnedecker, 2005 : 95) qualifie de « tout-venant référentiel ». Par exemple :

> **M. Giscard d'Estaing** a proposé une série d'économies. **Il s'**est demandé s'il était « urgent » de créer une chaîne éducative, **il s'**est interrogé sur le financement des interventions militaires ou humanitaires de la France dans le monde, ainsi que sur la pertinence de grands investissements, comme la grande bibliothèque. (corpus *Le Monde*)

  b. elles peuvent, au contraire, être *hétérogènes* et comporter :
    - des catégories diversifiées (différentes sortes de pronoms, de groupes nominaux définis, démonstratifs),

[4] (Schnedecker, 1997 : 13) qualifie les chaînes de référence de « phénomène à longue portée ».

- différentes catégories de groupes nominaux à tête lexicale (noms propres, groupes nominaux définis, groupes nominaux démonstratifs)[5] et un matériau lexical varié, comme dans :

> **Gilles-Eric Séralini**, chercheur français, spécialisé en biologie moléculaire est rattaché à l'université de Caen. Depuis 1997, **il** creuse **son** sillon en analysant les risques des pesticides et des OGM sur les organismes. Fin septembre, **il** a jeté un pavé dans la mare, en publiant une étude dans une revue de toxicologie (*Food & chemical toxicology*) sur les effets à long terme de la nourriture OGM. A peine l'étude parue, les « *contre-feux* », comme dit Séralini, se sont déclenchés. **Le chercheur** serait un anti-OGM notoire, **son** procédé - s'appuyer sur un hebdomadaire, un livre et deux films - nuirait au débat scientifique, **ses** travaux seraient contestables. (*Libération*, 19 octobre 2012).

– Concernant leur **déroulement dans le texte**, les chaînes de référence peuvent :

a. se succéder ; dans l'exemple suivant[6], la deuxième chaîne (en gras et italique) débute à la fin de la première (en gras), la troisième chaîne (en italique) débute à la fin de la deuxième :

> Johannesbourg, années 70. **Un professeur d'histoire** vit une vie de famille sans histoire, entouré de l'affection des **siens** et aveugle aux problèmes politiques et sociaux engendrés par l'"apartheid", régime ségrégationniste qui sévit en Union sud-africaine. **Cet homme paisible** a ***un jardinier noir***, paisible ***lui*** aussi, et soumis à la fatalité. ***Ce jardinier*** a *un fils* et *ce fils* participe à une manifestation d'écoliers violemment réprimée par la police. On apprendra que *l'enfant* est arrêté, puis qu'*il* est mort. (Résumé d'Une saison blanche et sèche, *Télérama*, 15.01.92)

b. s'entrecroiser (alternance entre les maillons de deux ou plusieurs chaînes de référence), par exemple[7] :

> Cependant **Milady**, ivre de colère, rugissant sur le pont du bâtiment, comme une lionne qu'on embarque, avait été tentée de **se** jeter à la mer pour regagner la côte, car **elle** ne pouvait

[5] Ce phénomène est présent dans la moitié des chaînes des portraits journalistiques (Schnedecker, 2005 : 97).

[6] L'exemple est emprunté à (Schnedecker, 1997 : 13).

[7] Dans l'exemple suivant, la chaîne de référence de Milady est en gras, celle du capitaine du navire est en gras et en italique.

> **se** faire à l'idée qu'**elle** avait été insultée par d'Artagnan, menacée par Athos, et qu'**elle** quittait la France sans **se** venger d'eux. Bientôt, cette idée était devenue pour **elle** tellement insupportable, qu'au risque de ce qui pouvait arriver de terrible pour **elle-même**, **elle** avait supplié ***le capitaine*** de **la** jeter sur la côte ; mais ***le capitaine***, pressé d'échapper à ***sa*** fausse position, placé entre les croiseurs français et anglais, comme la chauve-souris entre les rats et les oiseaux, avait grande hâte de regagner l'Angleterre, et ***Ø*** refusa obstinément d'obéir à ce qu'***il*** prenait pour un caprice de femme, promettant à ***sa*** passagère, **qui** au reste ***lui*** était particulièrement recommandée par le cardinal, de **la** jeter, si la mer et les Français le permettaient, dans un des ports de la Bretagne, soit à Lorient, soit à Brest […] (corpus *Les Trois Mousquetaires*)

c. se dédoubler[8] (un groupe d'individus se sépare), comme dans :

> TOUT se brise, s'évanouit sous les coups terribles que j'entends contre la porte. Je me lève en sursaut, et je cours pour ouvrir. **Ils** font irruption, mitraillette braquée, mines féroces; **ils se** dirigent tout de suite vers mon bureau. Je m'aplatis contre le mur, les mains en l'air, sous le dessin de saint Jérôme. Tandis que **l'un d'eux** me surveille, **son** arme contre ma poitrine, ***l'autre se*** met à jeter à terre les livres des rayonnages, à grandes brassées. (corpus *Le Monde Diplomatique*)

ou encore procéder par extraction (*i.e.* une forme de scission). Dans l'exemple suivant, du groupe initial « les trois mousquetaires » est extrait l'élément « Athos » (mais la partition dédoublée n'est pas comptabilisée dans la chaîne) :

> D'ailleurs *les trois mousquetaires* y venaient seuls ; *ils s'*étaient mis en quête et *Ø* n'avaient rien trouvé, rien découvert. ***Athos*** avait été même jusqu'à questionner M. de Tréville, chose qui, vu le mutisme habituel du digne mousquetaire, avait fort étonné son capitaine. Mais M. de Tréville ne savait rien, sinon que, la dernière fois qu'il avait vu le cardinal, le roi et la reine […] (corpus *Les Trois Mousquetaires*)

[8] Se diviser, se partitionner.

d. au contraire, fusionner (cas où des chaînes de référence qui désignent des individus sont rassemblées pour créer un ensemble) :

> **Patrice Landry** n'a rien d'un fou. C'est lui qui, avec sa compagne **Marie-Hélène Marcaud**, a découvert ces empreintes au printemps dernier. Résultat d'une longue et minutieuse enquête. Six mois après l'extraordinaire découverte, **le couple** participait début octobre à la révélation publique du site aux côtés des paléontologues du CNRS […] (corpus *Le Monde*)

Néanmoins, malgré la nature complexe des chaînes de référence et la diversité dont elles peuvent faire l'objet, il est possible de se fonder sur des indices linguistiques (redénomination du nom propre marquant un changement de point de vue, donc un changement de thème) pour prédire celle des chaînes de référence qui fera l'objet de reprises dans le document tout entier[9].

Ce qui nous paraît intéressant dans l'approche de (Schnedecker, 1997), c'est la conception de la référence qu'elle adopte, en termes de continuum textuel, mais aussi la possibilité offerte à une identification automatique des chaînes de référence.

D'une part, dans cette approche, les chaînes de référence sont **bornées** (à la différence des approches considérant toutes les expressions référentielles du texte comme une chaîne de référence). Cette notion de *borne* permet de résoudre en partie[10] l'entrecroisement des chaînes de référence (*i.e.* le bornage des chaînes limite les chaînes « à rallonge »). Elle offre surtout la possibilité de signaler le continuum référentiel du texte constitué par l'ensemble des chaînes de référence possédant le même premier maillon (signalant, de ce fait, que ces chaînes de référence sont coréférentes). En cela, les chaînes de référence, considérées comme de nouveaux types d'unités textuelles (Charolles, 1988 ; Schnedecker, 1997 : 193), interagissent avec les autres plans de l'organisation textuelle.

D'autre part, dans cette approche, les expressions référentielles contenues dans les chaînes de référence sont **marquées** (*i.e.* présentes) **dans le texte** ; les indices référentiels (Landragin, 2011) tels que les accords verbaux ne sont pas pris en

[9] Nous verrons dans la section 2 les contraintes pesant sur l'emploi de certaines expressions référentielles.

[10] Du moins, dans une perspective de TAL.

compte[11] ; ce qui facilite grandement l'automatisation de l'identification des maillons des chaînes de référence[12].

1.4 Cas exclus (provisoirement) des CR

Dans la définition des chaînes de référence adoptée, (Schnedecker, 1997) signale des cas exclus dans son approche :

- la cataphore (par exemple : « j'adore ***ça***, *le chocolat* »), qu'elle qualifie de formes momentanément « orphelines », car la cataphore ne peut pas constituer le premier maillon d'un chaîne de référence (une chaîne de référence ne peut être initiée que par un nom propre ou un groupe nominal), mais elle va se rattacher au premier « bon candidat » initiateur d'une chaîne. Par exemple, dans « **il** viendra finalement, **Paul**. », le pronom cataphorique « il » va se greffer au nom propre « Paul » qui constitue un bon candidat ;
- le nom tête des anaphores possessives, qui sont en partie non coréférentielles. Dans « Camille a fait les soldes. Son mari a regardé le match à la télévision. », on relève deux référents : « Camille » et « mari ». Bien que le déterminant possessif « son » soit dépendant de « Camille », il n'y a pas de coréférence. « Son » coréfère à « Camille » (ce qui est surprenant pour un déterminant) mais la tête lexicale désigne un autre référent (« mari ») : il y a en quelque sorte une double référence ;
- les anaphores associatives, où l'anaphore est dépendante du cotexte (anaphore non coréférentielle). Par exemple, dans « un couple m'a rendu visite hier ; le mari était insupportable » (exemple repris de (Milner, 1982 : 27-28)), ce sont nos connaissances du monde qui nous permettent d'associer « le mari » à « un couple ». C'est une anaphore associative dite de type membre-collection (Kleiber, 2003)[13].

De même que pour la cataphore, les anaphores possessives ou associatives peuvent se greffer à une chaîne de référence existante sans l'altérer ou l'arrêter ; elles participent au réseau référentiel qui peut tramer un texte. Ainsi, elles ne font que

[11] L'absence d'identification de ces indices référentiels peut aussi constituer une lacune.

[12] Les anaphores zéro sont aussi prises en compte dans cette approche. Elles sont, pour notre part, difficiles à identifier (elles nécessiteraient la mise en place de grammaires complexes). En effet, comment pourrions-nous prévoir ces cas de manière exhaustive et en minimisant le plus possible le bruit (*e.g.* les erreurs induites par l'identification de ces cas).

[13] Voir (Kleiber, 1997a, 1997b, 2000) pour une typologie des anaphores associatives (locatives, actantielles, fonctionnelles, méronymiques, etc.).

mettre davantage en évidence le premier maillon de la chaîne. De ce fait, ces trois cas sont finalement inclus dans les chaînes de référence.

Néanmoins, même si nous adoptons la définition des chaînes de référence de C. Schnedecker, les anaphores associatives ne pourront pas être identifiées de manière automatique car elles font appel à des connaissances sur le monde dont notre système automatique est dépourvu[14]. Mais, vu que ces cas participent indirectement aux chaînes de référence, cela ne constitue pas véritablement de problème en soi.

A l'inverse, dans notre approche automatique, nous pourrons tenir compte des cas d'« anaphore passoire » – anaphores nominales inaptes à conditionner le pronom personnel subséquent (Schnedecker, 1997 : 29) – telle que (en gras) :

> « […] *Fangio*. **Cette légende de la course automobile** a laissé *son* nom à l'histoire, mieux encore *il* l'a fait entrer dans le langage courant.[15]»,

qui correspond à une information supplémentaire apportée à propos du référent (que l'interprète éliminera rapidement de sa mémoire). Cette information supplémentaire sera prise en compte par notre système informatique dans son calcul de la référence, même si le genre et/ou le nombre de cette information ne correspond pas nécessairement avec le genre et le nombre du référent (voir chapitre 7, section 4).

1.5 Bilan

Dans cette première partie, nous avons rappelé la définition des chaînes de référence[16] et les caractéristiques qu'elles peuvent avoir et qui les distinguent des notions *d'anaphore* et de *coréférence*.

La notion de *borne* pour une chaîne permet de délimiter les diverses mentions d'un référent *via* la redénomination et d'interagir avec d'autres plans de la structure textuelle. De plus, le nombre minimal de maillons d'une chaîne (trois) permet d'identifier des phénomènes qui se situent au-delà de paires d'expressions référentielles. De ce fait, il est possible de suivre le référent grâce aux

[14] Le traitement des cas d'anaphores associatives feront l'objet de futures extensions de notre système.

[15] Cet exemple d'anaphore « passoire » est emprunté à (Schnedecker, 1997 : 26).

[16] « Convenons d'appeler chaîne de référence la suite des expressions d'un texte entre lesquelles l'interprétation construit une relation d'identité référentielle ». (Corblin, 1985 : 123)

informations distribuées au fil du discours et non plus seulement de manière locale[17].

Dans la partie suivante, nous présentons les diverses approches théoriques conditionnant l'accès et le maintien référentiel dans le discours.

[17] « L'évaluation de l'apport informationnel des anaphores ne saurait exclusivement opérer à partir du segment référentiel ou textuel immédiatement antérieur » (Schnedecker, 2005).

2 Les chaînes de référence et la continuité thématique

Comme nous l'avons vu plus haut (section 1), les chaînes de référence sont constituées d'expressions coréférentielles (noms propres, pronoms, groupes nominaux, etc.). Certaines de ces expressions seraient préférentiellement utilisées pour « ouvrir » des chaînes de référence, d'autres pour les « fermer ». Par exemple, les noms propres ou les groupes nominaux indéfinis signaleraient que leur référent n'est pas ou peu activé dans la mémoire de l'interprète (Ariel, 1990)[18], ils seraient donc des termes initiateurs de chaîne. En revanche, les groupes nominaux démonstratifs (Marandin, 1988) et la répétition de groupes nominaux pleins (Schnedecker, 1997), grâce à leur capacité de rupture discursive, seraient des expressions référentielles de clôture de chaîne.

Pour identifier les termes initiaux et de clôture des chaînes de référence et donc pour faciliter l'identification des chaînes de référence, plusieurs modèles linguistiques (Givón, 1983 ; Ariel, 1990 ; Gundel *et al.*, 1993) ont classé les expressions référentielles en fonction de leur forme et du degré d'accessibilité cognitive du référent (plus une expression référentielle est peu informative et courte, plus le référent est accessible). D'autres théories telles que la théorie dite du « centrage » permettent de prédire le choix d'une expression suivant la présence de son référent dans la proposition précédente.

2.1 La théorie de l'accessibilité

Les expressions référentielles permettent de référer à des éléments parfois déjà cités dans le discours. L'utilisation d'une expression référentielle plutôt qu'une autre résulte de son efficacité à remplir sa fonction d'identification (Clark et Marshall, 1981). Cette dimension a été conceptualisée par (Givón, 1983) puis (Ariel, 1990)[19] sous la forme d'une échelle d'accessibilité où chaque type d'expression référentielle est conditionnée par un degré d'accessibilité du référent. Cette échelle est définie comme une hiérarchie d'expressions référentielles dont les différentes formes linguistiques dépendent du niveau d'accessibilité de l'entité à

[18] En corpus, les maillons initiateurs de chaîne de référence peuvent être de types divers (Schnedecker, 1997) et nous verrons dans le chapitre 3 que cela est en partie lié au genre textuel.

[19] Nous nous intéresserons à l'échelle d'accessibilité d'Ariel.

laquelle elles réfèrent (voir Figure 7)[20], c'est-à-dire selon le degré d'importance du référent :

> "Accessibility theory offers a procedural analysis of referring expressions, as marking varying degrees of mental accessibility. The basic idea is that referring expressions instruct the addressee to retrieve a certain piece of given information from his memory by indicating him how accessible this piece of information is to him at the current stage of the discourse". (Ariel, 2001 : 29)

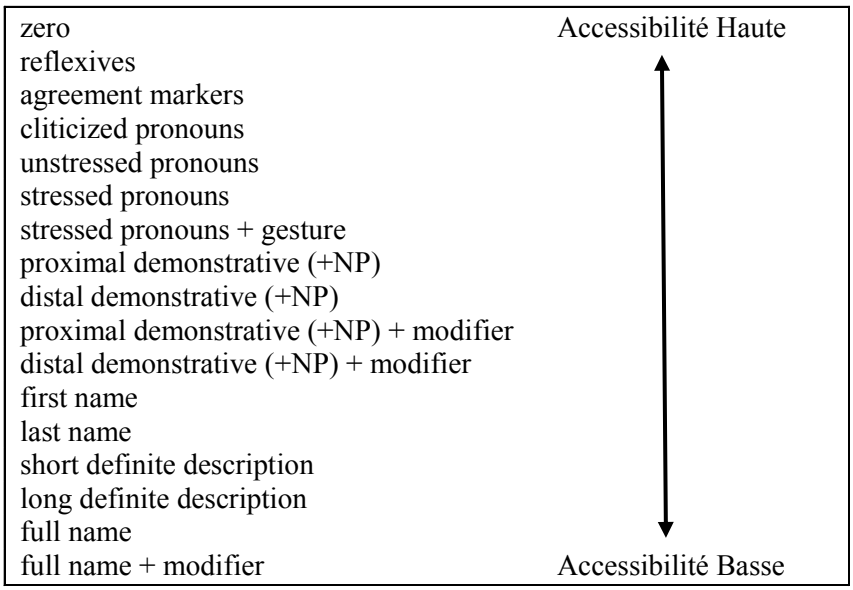


Figure 7 - Echelle d'accessibilité pour l'anglais selon (Ariel, 1990)

Ainsi, trois cas de figure peuvent être dégagés. Tout d'abord, les référents inactifs de faible accessibilité. Ils ont été présentés dans le texte, mais d'autres entités ont été instanciées entre temps. On fera référence à ces référents au moyen de noms propres ou de longues descriptions définies comme par exemple : « François Hollande », « le président de la République française ». Ensuite, dans le cas de l'accessibilité moyenne, on trouve des référents qui ne sont pas dans le focus mais qui ont déjà été activés. Des groupes nominaux démonstratifs réfèreront à ces derniers : « la Mégane 2 … **cette voiture** ». Enfin, est considérée comme un référent de forte accessibilité une entité située au premier plan dans la représentation du discours du lecteur, c'est-à-dire un référent saillant. Un pronom suffit à y faire référence, comme, par exemple, « Barack Obama…**il** ».

La théorie de l'accessibilité considère que le choix d'une expression référentielle est un phénomène bi-directionnel (Ariel, 1996) : le locuteur choisit une expression

[20] Il est à noter que, même si Ariel a construit son échelle pour l'anglais, elle prédit une hiérarchie identique pour les autres langues possédant les formes grammaticales nécessaires.

référentielle en prenant en considération le degré supposé d'accessibilité de l'entité dans la représentation mentale de son interlocuteur, tandis que l'auditeur cherche dans sa représentation mentale du texte une entité correspondant à l'expression référentielle utilisée par le locuteur.

Dans cette théorie, l'accessibilité du référent est conditionnée par quatre facteurs (Ariel, 1990) :

- la distance entre l'expression référentielle et son antécédent,
- le nombre de candidats au poste d'antécédent (*i.e.* s'il y a compétition référentielle ou non),
- la saillance[21],
- l'environnement d'identification des référents (notion d'*unité textuelle*).

Plus la distance entre l'antécédent et l'expression référentielle est grande, plus le degré d'accessibilité diminue[22]. (Asher *et al.*, 2006) ont analysé la distance entre une expression référentielle et son antécédent en termes « d'unités de discours élémentaire ». Leurs résultats confirment la distribution des expressions référentielles proposée par Ariel : les expressions anaphoriques pronominales et les démonstratives sont à une distance plus courte de leur antécédent que des expressions nominales. De même que (Ariel, 1990), (Asher *et al.*, 2006) observent également que les expressions nominales courtes telles que les prénoms ou noms propres seuls sont situés à une distance plus courte que les descriptions définies longues. Ariel précise que cette distance ne se mesure pas uniquement en nombre de mots, mais que les paragraphes, parties (ce qu'elle appelle « unités »), etc., sont autant de sources de distance qui augmentent, de fait, la difficulté pour le lecteur à accéder à l'information antérieure.

Lorsqu'une expression référentielle apparaît dans un contexte ambigu (*i.e.* s'il existe plusieurs candidats potentiels au « poste » d'antécédent), l'accessibilité est diminuée. Dans ce cas de concurrence référentielle (ou de compétition référentielle), une expression spécifique (telle que le nom propre) sera préférée à un simple pronom. C'est ainsi que, pour faire référence à « Paul » dans « Pierre et Paul sont venus », la répétition du prénom « Paul » peut être préférée à l'emploi du pronom « il », pour éviter une ambiguïté référentielle.

[21] Ariel met l'accessibilité sur le même pied d'égalité que la topicalité (« Saliency: The antecedent being a salient referent, mainly whether it is a topic or a non-topic. » (Ariel, 1990 : 29)).

[22] Pour (Gernsbacher, 1989), il s'agit plutôt d'inactivation du référent et non pas de distance. En effet, si le référent est éloigné, d'autres concepts ont été ajoutés au texte entre temps, supprimant alors l'activation du premier référent.

Concernant la saillance des référents, plus un référent est saillant, plus il est accessible. D'après Ariel, les thèmes du discours sont les entités les plus saillantes :

> « in natural discourse [...] topics (mainly discourse topics) constitute the most salient entities more often than not. It seems that topics occupy a privileged position in memory. » (Ariel, 1990 : 28-29).

En effet, les personnages principaux d'un récit sont plus saillants que de simples figurants n'apparaissant que dans certaines scènes. Aussi, l'introduction d'un personnage par un nom propre plutôt qu'une description lui attribue un poids plus fort.

Enfin, concernant l'environnement d'identification des référents, le référent est plus accessible si la dernière mention est située dans la même unité (*i.e.* phrase, paragraphe, etc.) que l'anaphore (le même cadre spatial, temporel ou le même point de vue). La notion d'unité textuelle telle que définie par Ariel est assez large :

> « Unity: The antecedent being within vs. without the same frame/world/point of view/segment or paragraph as the anaphor. » (Ariel, 1990 : 28-29).

Ce dernier facteur souligne l'effet de la structure du discours sur la forme de la référence : les pronoms seront préférentiellement utilisés pour faire référence à des éléments situés dans le même segment discursif.

Ainsi, l'antécédent « Chirac » dans la phrase :

> **Chirac** a fait tout ce qu'**il** a pu pour nous dissuader de voter pour lui. (corpus *Le Monde*)

est un exemple de haute accessibilité référentielle : la distance entre l'antécédent et l'anaphore est minimale ; on trouve un seul candidat au poste d'antécédent (« Chirac ») et l'antécédent est le thème de la phrase, ce qui explique pourquoi le pronom constitue la seule forme de reprise possible. Dans le cas de propositions subordonnées, des pronoms seront préférablement utilisés (ce sont des formes liées) ; en revanche, l'emploi de noms propres sera privilégié pour des propositions coordonnées.

Ariel qualifie cette échelle d'arbitraire dans le sens où, seul l'ordre global prédit est invariant. Par contre, le système des pronoms, par exemple, (leur

organisation, la présence ou l'absence de forme, etc.) va varier selon la langue[23]. Aussi, l'échelle d'accessibilité est universelle car le codage des degrés d'accessibilité fait intervenir trois critères :

- l'*informativité*, c'est-à-dire le niveau d'information contenu dans l'expression référentielle par rapport à l'antécédent ;
- l'*atténuation*, soit la brièveté de l'expression référentielle ;
- la *rigidité*, permettant de rendre compte de l'univocité de l'expression référentielle dans un contexte potentiellement ambigu (cela concerne les noms propres et les pronoms de première et deuxième personne).

En combinant ces trois paramètres, il est possible de prédire qu'une forme référentielle marque une accessibilité haute si elle est peu informative, peu rigide et très atténuée ; ou au contraire qu'elle marque une accessibilité basse si elle est informative, rigide et moins atténuée[24]. Par exemple, une description définie assez longue comme « Le président de la République Française » est plus informative qu'un groupe nominal simple tel que « le président ». « Le président » fait référence à une entité d'accessibilité moyenne, alors que « Le président de la République française » fait référence à une entité de faible accessibilité. Aussi, « Edmund » est un marqueur rigide, mais « Edmund Muskie » est plus rigide car les noms propres complets jouent un rôle d'étiquette (de désignateur rigide (*cf.* (Kripke, 1972) *in* (Kleiber, 1981))) par rapport à une entité supposée unique.

Ainsi, la théorie de l'accessibilité est reflétée par les expressions référentielles utilisées par le locuteur. Le caractère universel de cette théorie ainsi que l'appui théorique issu de la psychologie cognitive en font des atouts importants.

Néanmoins, (Kleiber, 1990) dénonce, dans cette théorie, la trop faible prise en compte du contenu référentiel des expressions référentielles, réduites aux trois paramètres énoncés plus haut. Pour éviter ce raisonnement circulaire, il propose de prendre en compte le sens des expressions référentielles, d'une part, mais aussi la manière dont le locuteur veut présenter le référent. Par exemple, dans la phrase :

> je m'en plaindrai à M. de Tréville, et M. de Tréville s'en plaindra au roi.
> (corpus *Les Trois Mousquetaires*)

Dumas a volontairement répété le nom propre « M. de Tréville » alors qu'il aurait pu utiliser les pronoms « il » ou « celui-ci »/ « ce dernier ».

[23] Dans son étude, (Ariel, 1990) fait référence à plusieurs langues (anglais, hébreu, etc.).
[24] Nous utiliserons ces trois critères dans notre calcul de la référence (*cf.* chapitre 7).

(Reboul *et al.*, 1997), de leur côté, demeurent plus radicaux que Kleiber dans la critique de la théorie de l'accessibilité qui est, pour eux, inutile et constitue une « réintroduction de la notion de cohésion ». Cette théorie serait trop simplificatrice et réductrice car, à travers l'échelle donnée, la référence serait réduite à un phénomène uniquement linguistique.

(Schnedecker, 2005) pointe enfin le manque de prise en compte du genre textuel d'occurrence de la coréférence. Cette position n'est pas partagée par des études telle que (Downing, 2000) qui montre que l'échelle d'accessibilité n'est pas influencée par les différences de genre pour les articles de presse.

Bien qu'elle n'exprime qu'une propriété graduelle, nous retiendrons cette théorie dans notre calcul de la référence. En effet, le classement hiérarchique des expressions référentielles sera utilisé en complément d'autres critères tels que le genre textuel (chapitre 3) pour participer à l'identification automatique des chaînes de référence.

2.2 La théorie du centrage

La théorie du centrage de (Grosz *et al.*, 1995)[25] décrit la cohésion locale produite par l'ensemble des relations entre les énoncés d'un même segment textuel. Elle s'inscrit dans la théorie de la structure du discours de (Grosz et Sidner, 1986). Un des principes fondamentaux de la théorie du centrage prévoit que, parmi les référents d'une phrase, un seul peut être considéré comme le centre préféré (le focus), donc que certaines entités sont plus « centrales » que d'autres.

On distingue plusieurs types de centres : les centres anticipateurs (Ca), le centre rétroactif (Cr) et le centre préféré (Cp). D'après (Cornish, 2000 : 11) :

> « Le Ca est un ensemble de centres (anticipateurs) réalisés ou évoqués dans un énoncé, et dont l'allocutaire pourra anticiper, avec divers degrés de probabilité, qu'ils vont devenir le centre rétroactif (Cr, ou topique de discours local) de l'énoncé suivant. Le Cp évoqué dans un énoncé est « le centre préféré », autrement dit, le centre qui est classé premier dans le Ca de cet énoncé. Le Cr est l'objet de discours psychologiquement le plus saillant à la fois pour l'énonciateur et l'allocutaire au moment où l'expression qui le réalise est employée : c'est l'entité topique de discours local. »

[25] La théorie est plus largement développée dans (Walker *et al.*, 1998), où le modèle des états attentionnels en pile de Grosz et Sidner est remplacé par un modèle en cache.

De là découlent des contraintes sur l'utilisation par le locuteur de différents types d'expressions pour référer à ces entités. Afin de mettre en place ces contraintes, plusieurs règles de transitions sont définies entre les différents centres de deux énoncés (E) consécutifs (voir Tableau 2). Suivant le type des marqueurs référentiels (pronom, groupe nominal défini ou démonstratif, nom propre...) utilisés pour assurer cet enchaînement, on assistera à une continuation, un maintien temporaire, ou un déplacement de la focalisation.

	$Cr(E_i) = Cr(E_{i-1})$ OU $Cr(E_{i-1}) = [?]$	$Cr(E_i) \neq Cr(E_{i-1})$
$Cr(E_i) = Cp(E_i)$	continuation	déplacement en douceur
$Cr(E_i) \neq Cp(E_i)$	rétention	déplacement brutal

Tableau 2 - Synthèse des quatre transitions du centrage (issu de (Walker *et al.*, 1998,6), cité par (Cornish, 2000 : 16))

Pour déterminer le centre préféré (Cp) parmi les centres anticipateurs (Ca), une hiérarchie des référents du Ca est établie suivant leur position (la première position est privilégiée) et leur fonction grammaticale (sujet > objet indirect animé > objet direct > objet indirect inanimé > objet oblique). Selon la théorie du centrage, on ne peut pas, par exemple, mettre en position sujet une entité qui n'a pas été introduite dans l'énoncé précédent, comme le montre, dans les exemples suivants empruntés à (Cornish, 2000[26]), l'impossibilité d'un déplacement brutal (noté «# »dans la dernière phrase) :

a. **Jean** est un type sympa.
b. **Il** a rencontré **Marie** hier.
c.i #Lucie était avec elle.
c.ii **Elle** était avec **Lucie**.

En effet, en a), seul le nom propre « Jean » est le Ca (car le groupe nominal « un type sympa » est une prédication sur « Jean »). En b), le pronom « il » en position sujet exprime le référent introduit dans l'énoncé a), soit « Jean » (qui est le Cr). Dans le centrage, l'énoncé b) constitue une Continuation de l'énoncé a). Mais en c.i), le pronom « elle » reprend un Ca non introduit dans b) (donc ce Ca n'est pas le Cr de b)). Néanmoins, le fait que « Marie » soit désignée en c.i) par un pronom indique que « Marie » est le Cr. Lucie est le Cp car elle est située dans une position « supérieure » dans le Ca de c.i) (*i.e* « Lucie » est en première position et occupe la fonction sujet). Pour éviter ce déplacement brutal, un déplacement en douceur est proposé en c.ii) où le Cr « Marie » est repris par le pronom qui constitue le Cp (car il est en position sujet) de c.ii).

[26] Ces exemples sont eux même empruntés à (Di Eugenio, 1996).

Pour (Péry-Woodley, 2000), la théorie du centrage fournit une « méthode précise d'analyse de la cohérence à l'intérieur d'un segment de discours, elle constitue le volet local d'un modèle qui envisage aussi la cohérence intersegmentale ». Néanmoins, pour (Cornish, 2000), « la préoccupation principale du centrage porte sur les relations de focus d'attention *local* dans un segment de discours donné ». Or, pour être opératoire, cette théorie devrait prendre en compte la structure du discours à l'intérieur comme à l'extérieur des segments. (Grosz *et al.*, 1995) soulignent, d'une part, qu'un centre est associé à un énoncé et non à une phrase et d'autre part, qu'il s'agit d'un objet sémantique et non d'un objet textuel. Dans le centrage, les connexions entre énoncés d'un même segment s'effectuent de manière linéaire, ce qui n'est pas forcément le cas en corpus : il n'est pas rare qu'un énoncé se connecte à un autre énoncé situé plus en amont (dans le même segment ou dans un segment différent).

La théorie du centrage adopte, similairement à l'accessibilité, une perspective axée sur le traitement cognitif des discours mais, elle vise, en plus, des applications informatiques. Cette théorie semble exploitable du point de vue du traitement automatique afin de déterminer si le même centre se répète, donc s'il y a rupture ou continuité thématique. Cependant, (Kleiber, 2002) souligne que le centrage d'attention souffre de plusieurs problèmes, notamment du fait qu'il n'arrive pas à maîtriser totalement les mécanismes de la référence discursive et plus spécialement ceux des marqueurs référentiels. Cette théorie ne proposerait pas, selon l'auteur, d'échelle de distribution des différents types d'expressions référentielles, mais elle considèrerait que le pronom est l'expression référentielle signalant la continuité[27] par excellence. Or, la continuation du même centre préféré est possible à partir de plusieurs expressions référentielles, comme dans :

- **Un hélicoptère** s'est écrasé hier. **Il** volait au-dessus des collines.
- **Un hélicoptère** s'est écrasé hier. **Cet hélicoptère** volait au-dessus des collines.
- **Un hélicoptère** s'est écrasé hier. **L'hélicoptère** volait au-dessus des collines.

contrairement aux prédictions de cette théorie.

Enfin, pour (Kleiber, 2002), le centrage ne prédit pas correctement l'emploi des marqueurs référentiels. Il s'appuie sur l'exemple suivant :

> Il était une fois **un prince**. **Ce prince** / *Il vivait dans un pays de neige et de brumes.

[27] La « saillance discursive » par excellence pour (Walker *et al.*, 1998 : 5).

Dans cet exemple, le centre anticipateur (Ca) « un prince » est aussi le centre préféré (Cp). Selon le centrage, le maintien du Cp doit s'effectuer par l'intermédiaire du pronom personnel dans l'énoncé suivant. Or, Kleiber estime que seul le démonstratif peut être jugé acceptable dans ce contexte (bien que les démonstratifs marquent un changement sur le plan référentiel suivant le centrage).

Ainsi, la portée et la puissance prédictive de la théorie du centrage apparaissent limitées lors de leur application en corpus (Walker et *al*, 1998), (Cornish, 2000). De plus, les nombreux problèmes relevés par (Cornish, 2000) et (Kleiber, 2002) nous invitent à émettre des réserves quant à l'utilisation du centrage, en l'état, pour notre projet. Dans la section suivante, nous présentons une reformulation du centrage par (Beaver, 2004) *via* la théorie de l'optimalité.

2.3 Reformulation du centrage par la théorie de l'optimalité

(Beaver, 2004) a reformulé les règles proposées par la théorie du centrage (Grosz et Sidner, 1986) en termes de contraintes pour pouvoir les utiliser en sémantique[28]. Pour cela, il s'est appuyé sur la théorie (générative) de l'optimalité de (Prince et Smolensky, 1993).

2.3.1 La théorie de l'optimalité

La théorie de l'optimalité (Prince et Smolensky, 1993) est un cadre formel[29] qui repose sur une série de concepts simples que (Dal et Namer, 2005 : 2) synthétisent de la manière suivante :

- les langues doivent satisfaire un certain nombre de contraintes universelles de bonne formation[30],
- à l'intérieur d'une langue donnée, des contraintes peuvent entrer en conflit,
- certaines contraintes sont violables (*violability*),

[28] Beaver souhaite ainsi rapprocher les communautés en psycholinguistique et TAL qui ont travaillé sur la résolution des anaphores avec des communautés de sémantique et pragmatique.

[29] La théorie de l'optimalité est largement utilisée en phonologie et en syntaxe.

[30] Un exemple de contrainte serait que la préposition possède toujours un argument (et cet argument pourra, suivant la langue, être situé à droite (comme en français) ou bien à gauche (comme en japonais)).

- on peut ordonner[31] les contraintes (*ranking*), selon que leur infraction est ou n'est pas fatale (par convention, « *! » et « * » notent respectivement chacun de ces cas),
- les différences entre les langues du monde sont ramenables[32] à des différences dans l'ordonnancement des contraintes.

Le processus d'optimisation permettant de déterminer le candidat optimal en sortie (*output*), pour une entrée donnée (*input*), se schématise en quatre étapes (Lyngfelt, 2000) :

Input → GEN (générateur) → EVAL (évaluation) → Output

Ainsi, la fonction GEN génère tous les candidats potentiels. Puis, la fonction EVAL évalue les candidats suivant l'ordre des contraintes. Après quoi le candidat optimal émerge en sortie et tous les autres candidats sont jugés agrammaticaux.

D'après la théorie de l'optimalité, tout output linguistique est optimal car il subit la violation la moins coûteuse (Ašić, 2008). En effet, la violation d'une contrainte ne rend pas forcément une phrase agrammaticale si cette contrainte a été violée au profit d'une contrainte plus forte (*i.e.* classée plus haut dans la hiérarchie).

Le résultat de l'application du processus est présenté sous forme de tableau (voir Tableau 3), où une transgression est notée « * » (la contrainte est violée), une violation fatale est notée « ! » et le candidat optimal est pointé par « ☞ » (le candidat optimal peut soit ne violer aucune contrainte, soit en violer une ou plusieurs si elles peuvent être violables dans le modèle) :

Contrainte A >> Contrainte B

Candidats	*Contrainte A*	*Contrainte B*
Candidat 1	*	
☞ Candidat 2		*
Candidat 3	* !	

Tableau 3 – Exemple de représentation en tableau

Au-dessus du tableau, « >> » précise que la contrainte A est située plus haut dans la hiérarchie que la contrainte B. Dans notre exemple, le candidat 3 a violé une contrainte fatale, il ne peut donc pas être le candidat optimal. Les candidats 1 et 2 ont violé chacun une contrainte. Or, suivant l'indication fournie au-dessus du tableau, la contrainte A est située plus haut dans la hiérarchie, donc le

[31] L'ordonnancement est spécifique à la langue (il est lié à la variation linguistique).
[32] Ce terme est employé par les autrices.

candidat 2 représente le candidat optimal (car il n'a pas violé la plus haute contrainte).

En guise d'illustration « concrète », reprenons l'exemple célèbre de l'analyse sémique de « chaise » *vs* « fauteuil » de (Pottier, 1964). Dans cette configuration :

- les contraintes à valider sont les divers sèmes :
  a. « avec dossier »,
  b. « sur pieds »,
  c. « pour une seule personne »,
  d. « pour s'asseoir »,
  e. « avec bras » ;

- les candidats potentiels sont :
  1. « tabouret »,
  2. « pouf »,
  3. « canapé »,
  4. « chaise »
  5. « fauteuil ».

Dans cet exemple, les contraintes ne sont pas ordonnées et aucune contrainte n'est fatale. Le tableau est le suivant (voir Tableau 4) :

Candidats	*Contrainte A*	*Contrainte B*	*Contrainte C*	*Contrainte D*	*Contrainte E*
Candidat 1	*				*
Candidat 2	*	*			*
Candidat 3			*		
Candidat 4					*
☞ Candidat 5					

Tableau 4 – Représentation en tableau de l'analyse en sèmes pour « fauteuil »

Le candidat 2 a violé trois contraintes, le candidat 1 en a violé deux, les candidats 3 et 4 ont violé chacun une contrainte. Seul le candidat 5 (« fauteuil ») n'a violé aucune contrainte. De ce fait, le candidat 5 est le candidat optimal pour cette série de contraintes.

Comme précisé en introduction de cette section, la théorie de l'optimalité fournit des contraintes universelles plutôt applicables au domaine de la phonologie (Ašić, 2008). Par exemple, deux grandes familles de contraintes vont être distinguées dans le domaine phonologique :

- les contraintes de *marque*, qui vont porter sur les représentations structurales des candidats. Par exemple, la contrainte *ATTAQUE : une syllabe doit avoir une attaque.*
- les contraintes de *fidélité*, qui vont pénaliser la présence d'une divergence entre l'input et le candidat. Par exemple, les containtes *MAX : maximisation des éléments de l'input dans l'output*, vont pénaliser les suppressions ; les contraintes *DEP : dépendance des éléments de l'output dans l'input*, vont, elles, pénaliser les épenthèses[33].

Dans notre projet, nous allons utiliser l'approche de Beaver qui a adapté les contraintes de la théorie de l'optimalité à la résolution des anaphores.

2.3.2 L'approche de Beaver

(Beaver, 2000, 2004) a redéfini la théorie du centrage dans le cadre de la théorie de l'optimalité. Il en a reformulé les divers types de centres en un système de contraintes basées sur la notion de *thème phrastique* qu'il définit de la manière suivante :

> "the entity referred to in both the current and the previous sentence, such as the relevant referring expression in the previous sentence was minimally oblique. If there is no such entity, the topic is undefined". (Beaver, 2000 : 24 ; 2004 : 12)

Pour définir le thème phrastique comme une marque de cohérence discursive, (Beaver, 2004 : 12-13) propose une série de 6 contraintes à satisfaire, classées par ordre d'importance (de la plus importante à la moins importante)[34] :

- AGREE: les expressions référentielles ont le même genre et le même nombre que leurs antécédents,
- DISJOINT: les arguments du même prédicat doivent être disjoints (*e.g,* les arguments d'un même prédicat ne peuvent pas être en relation de coréférence),
- PRO-TOP: le thème est pronominalisé,
- FAM-DEF: les groupes nominaux définis sont « familiers », c'est-à-dire qu'ils réfèrent à une entité déjà mentionnée,

[33] Modification phonétique permettant de faciliter l'articulation d'un mot ou d'un groupe de mots via l'insertion d'un phonème (par exemple, la prononciation du mot « pneu » en « peneu » ([pønø]) présente un « e » épenthétique).
[34] L'ordre des contraintes est arbitraire.

- COHERE: le thème de la phrase en cours est le même que celui de la phrase antérieure,
- ALIGN: le thème est en position sujet.

Ces contraintes, reformulées en termes de relations entre les antécédents et les anaphores, sont appliquées par Beaver de manière hiérarchique pour la résolution des anaphores pronominales pour l'anglais. Conformément à la théorie de l'optimalité, les contraintes sont violables et l'antécédent optimal (*i.e.* préféré) est sélectionné parmi une série de candidats potentiels s'il satisfait le plus de contraintes possibles.

(Beaver, 2004 : 17) fournit l'exemple suivant pour illustrer l'application des contraintes :

a) Jane$_i$ likes Mary$_j$.
b) She$_k$ often visits her$_l$ for tea$_m$.
c) The woman$_n$ is a compulsive tea drinker.

La phrase a) ne contenant pas d'anaphore, le Tableau 5 reporte l'application des contraintes pour la phrase b). En colonne, on retrouve les 6 contraintes classées par ordre d'importance (la contrainte la plus à gauche est la plus importante) et en ligne figurent les différentes associations entre antécédents et anaphores. On peut observer qu'aucun candidat ne viole la contrainte AGREE (la plus importante) vu que toutes les expressions référentielles sont au féminin singulier. Par contre, la contrainte COHERE est violée par l'ensemble des candidats car le thème de la phrase a) n'est pas défini. La contrainte DISJOINT n'est pas satisfaite lorsque les deux pronoms (« she » et « her »), arguments du verbe « visit », réfèrent à la même entité (« Jane » ou « Mary »). La contrainte PRO-TOP est violée dans les deux cas où il n'existe pas de référence anaphorique. De son côté, la contrainte FAM-DEF est violée par les six derniers candidats lorsqu'un pronom n'est pas considéré comme anaphorique. La dernière contrainte ALIGN permet de départager les candidats 1 (k=i, l=j) et 3 (k=j, l=i) car elle privilégie la lecture en parallèle où le pronom en position sujet reprend le thème en position sujet. De ce fait, le candidat optimal est le candidat 1, soit la prédiction que « Jane often visits Mary for tea ».

Example (b)	AGREE	DISJOINT	PRO-TOP	FAM-DEF	COHERE	ALIGN
☞ $k = i, l = j$					*	
$k = l = i$		*			*	
$k = j, l = i$					*	*
$k = l = j$		*			*	
$k = i, l \notin \{i, j\}$				*	*	
$k = j, l \notin \{i, j\}$				*	*	
$k \notin \{i, j\}, l = i$				*	*	*
$k \notin \{i, j\}, l = j$				*	*	*
$k, l \notin \{i, j\}, k \neq l$			*	**	*	*
$k = l \notin \{i, j\}$		*	*	**	*	*

Tableau 5 – Structure en tableau de la phrase b.

De même que (Gegg-Harrison et Byron, 2004), nous pensons que la méthode de (Beaver, 2004) possède plusieurs avantages : tout d'abord, l'utilisation de contraintes à valider pour identifier les paires antécédent-anaphores se prête bien à une implémentation informatique ((Beaver, 2004) propose l'algorithme COT qui utilise cette méthode). Ensuite, cette méthode laisse la possibilité de violer certaines contraintes. Aussi, cette méthode est modulable car il est possible d'ajouter, de supprimer ou de modifier l'ordre hiérarchique des contraintes suivant la langue et l'objet d'étude (résolution pronominale, résolution nominale, etc.). Enfin, la représentation en tableau permet de filtrer de manière claire et visible les reprises anaphoriques.

3 Conclusion

Les chaînes de référence constituent l'un des dispositifs permettant de rendre un texte cohésif. Nous avons émis l'hypothèse que les chaînes de référence représentent un type d'éléments linguistiques à prendre en compte pour détecter les thèmes. En effet, les liens référentiels permettent de pointer un référent « saillant » qui peut représenter localement le thème du paragraphe ou constituer le « thème continu » (Givón, 1983) au niveau du discours (lorsque le maillon initiateur d'une chaîne de référence se répète sur plusieurs chaînes).

Nous avons présenté des modèles linguistiques du discours qui ont classé les expressions référentielles dans des échelles d'accessibilité en fonction de leur forme ou qui ont mis en place une série de règles à utiliser pour évaluer les expressions référentielles candidates suivant un référent donné (en termes de continuité thématique ou de transition thématique).

Parce que les expressions référentielles jouent le rôle de signal de changement de thème, d'une part, et au vu des résultats obtenus par (Asher *et al.*, 2006) confirmant certaines prédictions d'Ariel en terme de distance, d'autre part, nous choisissons de suivre l'approche d'Ariel pour hiérarchiser de manière automatique les expressions référentielles susceptibles de constituer des chaînes de référence. Dans notre système d'identification de chaînes de référence, pour trier les diverses expressions référentielles candidates pour un référent donné (paires antécédent-anaphore), nous utilisons l'approche informatique modulable du centrage dans sa version modifiée par (Beaver, 2004).

Cependant, ni l'accessibilité ni le centrage ne prennent en compte les propriétés des chaînes de référence telles que le genre textuel. Nous proposons dans le chapitre suivant des études sur les propriétés des chaînes de référence dans des corpus issus de plusieurs genres.

Chapitre 3

Genre textuel et chaînes de référence

Outre le matériau lexical permettant d'obtenir une échelle d'éléments saillants, d'autres éléments à prendre en compte sont les propriétés des chaînes de référence. Des études antérieures telles que (Tutin *et al.*, 2000), (Jenkins, 2002), (Schnedecker, 2005), (Condamines, 2005) ou (Goutsos, 1997) insistent sur l'importance d'un aspect qui demeure relativement peu étudié ou exploité à l'heure actuelle par les systèmes de détection automatique de thèmes ou de coréférence : le genre textuel. Ainsi, pour ces auteurs, les genres conditionneraient-ils l'introduction du référent (première mention) et les formes privilégiées de sa reprise. Par exemple, dans les portraits journalistiques, les chaînes de référence introduites par des noms propres sont privilégiées (Jenkins, 2002 ; Schnedecker, 2005). Elles comportent principalement trois types de constituants à fonction référentielle : les noms propres, les pronoms et les groupes nominaux (voir l'exemple ci-dessous[1]).

> **George W. Bush** entend contre-attaquer sur le terrain culturel. **Il** a réussi, en 2000, à **se** donner l'image d'un rancher texan (faisant presque oublier Yale, Harvard, **son** grand-père financier et sénateur, **son** père président…). **Il** a montré qu'**il** avait vécu : problèmes d'alcool qu'**il** a surmontés, Dieu qu'**il** a rencontré, équipe de base-ball qu'**il** a rachetée… À la différence de Kerry, **il** parle comme un Américain moyen. Pas de grandes phrases, du bon sens, des petites vannes. Fort de cette image john-waynesque, **il se** présente comme le seul à avoir les tripes – le « caractère » comme disent les républicains – pour affronter sans flancher le principal problème du moment : le terrorisme. Les électeurs, a-t-**il** déclaré la semaine dernière, doivent choisir entre « *une Amérique qui dirige le monde avec force et confiance, et une Amérique qui est hésitante face au danger …* » (…). (*Libération*, 04/03/04)

Les chaînes de référence sont principalement homogènes dans ce genre (elles sont constituées essentiellement de noms propres et de pronoms). Par là même, la structure des chaînes de référence et le choix des expressions référentielles qui les constituent sont dépendants du genre.

[1] Cet exemple est emprunté à (Schnedecker, 2005 : 15).

Dans le domaine du TAL, plusieurs travaux proposent des modèles cognitifs pour le calcul de la référence ou exploitent des indices linguistiques de surface pour des tâches telles le dialogue homme-machine (Salmon-Alt, 2001), la génération automatique (Manuélian, 2003) ou la détection de thèmes (Hernandez, 2004). Cependant, malgré le nombre important de travaux sur la référence, peu de modèles opérationnels rendent compte des propriétés des chaînes de référence suivant le genre.

Ainsi, posons-nous comme hypothèse que les chaînes de référence possèdent des propriétés déterminées par leur genre d'occurrence. Le genre conditionnerait la manière dont elles sont construites. En ce sens, les chaînes de référence seraient dépendantes du genre textuel, quant au matériau lexical utilisé, à la distance entre les maillons ou à la longueur (nombre de maillons) de la chaîne. Nous avons mené une étude comparative des chaînes de référence sur un corpus composé de divers genres textuels pour vérifier notre hypothèse et pour prendre en compte ces paramètres dans notre calcul de la référence (chapitre 7). Compte-tenu de la variété des genres textuels présents dans les archives internes des moteurs de recherche, nous identifions les chaînes de référence relatives à des personnes mais aussi à des objets abstraits (par exemple, un phénomène tel que le réchauffement climatique).

Dans les sections suivantes, nous présentons l'étude en corpus de genres diversifiés qui a été menée et la typologie des chaînes de référence obtenue. Nous effectuons ensuite une étude de cas sur le genre des faits divers.

1 Impact du genre sur la composition des chaînes de référence : une étude en corpus

Nous présentons l'étude d'un corpus composé de divers genres permettant de construire les ressources linguistiques nécessaires à l'identification automatique des chaînes de référence[2]. En suivant le « principe de modestie » d'(Adam, 2004 : 94), cité par (Jenkins, 2002) – selon lequel il faut se limiter, dans l'approche linguistique des genres, à des objectifs descriptifs –, nous nous contenterons de décrire les spécificités des chaînes de référence suivant le genre textuel.

1.1 Objectifs de l'étude

Selon (Lin et Hovy, 1997 ; Corblin, 1995b), chaque genre de texte comporte des régularités spécifiques dans sa structure discursive. Le genre textuel peut ainsi influencer le type des expressions référentielles présentes dans un texte (Tutin *et al.*, 2000 ; Condamines, 2005) ainsi que le choix des diverses désignations du même référent (Cornish, 1998). Les études antérieures menées sur la corrélation entre genre discursif et chaînes de référence se sont appuyées uniquement sur l'étude d'un genre textuel précis (*i.e.* le portrait journalistique pour (Jenkins, 2002) et (Schnedecker, 2005), les textes expositifs pour (Goutsos, 1997)) ou se sont limitées à n'analyser qu'une catégorie d'anaphores (pronominales (Tutin *et al.*, 2000) ; hyperonymiques (Condamines, 2005)). Cela a mené (Demol, 2010 : 7) à constater le caractère restreint de son étude menée uniquement sur les articles de presse :

> « Nous sommes consciente qu'en limitant notre étude à des articles de presse, nous n'offrons qu'une image partielle des emplois du démonstratif, les anaphores constituant en effet un domaine de recherche sensible aux propriétés du genre et du registre. ».

Dès lors, pour vérifier l'impact du genre textuel sur la composition des chaînes de référence, nous avons constitué un corpus composé de divers genres (analyses

[2] Cette étude a fait l'objet d'une publication aux 6èmes journées de linguistique de corpus (Longo et Todirascu, 2009).

politiques, lois européennes, éditoriaux, roman, rapports publics). Nous rejoignons en ce sens (Baumer, 2011a, 2011b) qui a aussi utilisé un corpus de deux genres (fiction littéraire et portraits journalistiques) pour comparer le comportement des chaînes de référence. Ainsi, étudions-nous celles-ci selon des propriétés dépendantes du genre, comme leur longueur, la nature ou la fréquence de leurs maillons. A l'issue de cette étude, les spécificités dégagées vont permettre de paramétrer notre outil suivant ce type de variation. Par exemple, notre système sera en mesure de « prédire » le nombre de maillons des chaînes de référence attendu pour un genre particulier ou bien d'attribuer la priorité (un score élevé) à un candidat appartenant à la catégorie grammaticale « préférée » d'un genre défini.

1.2 Le corpus multi-genres

Compte-tenu de la variété des genres de documents susceptibles de figurer dans les archives internes des moteurs de recherche, nous avons choisi d'étudier des extraits (50 000 *tokens*)[3] issus de cinq genres différents (textes narratifs et non narratifs), répartis de la manière suivante (voir Tableau 6) :

GENRE	SOUS-CORPUS	PERIODE	NB MOTS
Analyses politiques	*Le Monde*	2004	110 012
Editoriaux	*Le Monde Diplomatique*	1980-1988	114 037
Roman	*Les Trois Mousquetaires (Dumas)*[4]	1844	105 068
Lois européennes	*Acquis Communautaire (Steinberger et al.)*	2006	116 702
Rapports publics	*La Documentation Française*	2001	106 765
		TOTAL	552 584

Tableau 6 - Répartition du corpus issu de genres textuels différents

Le corpus de genres divers ainsi constitué obéit aux critères d'élaboration affirmant que l'objectif de l'étude doit guider le mode d'élaboration du corpus (Habert et *al.*, 1997).

Les analyses politiques[5] du *Monde* traitent de la préparation des divers partis politiques à l'élection présidentielle de 2007. Une compétition présidentielle est

[3] Nous avons travaillé sur des extraits de notre corpus de départ (de 500 000 *tokens* (mots et signes de ponctuation)), tant l'annotation manuelle des CR est une tâche complexe. Nous souhaiterions poursuivre l'annotation de ce corpus de référence, vu qu'il n'en existe pas encore pour le français.

[4] Notre choix s'est porté sur un roman libre de droits.

[5] Ou articles d'information (Agnès, 2009 : 30)

installée et Valéry Giscard D'Estaing réaffirme le principe d'une candidature de l'UDF. De leur côté, les éditoriaux du *Monde Diplomatique* abordent la création de deux centres d'étude français – le CERMOC et le CEDEJ – au Proche-Orient. L'extrait issu du roman *Les Trois Mousquetaires* relate l'arrivée de d'Artagnan au Bourg de Meung. Le portrait du jeune homme et de sa monture y sont décrits de manière détaillée. Dans les lois européennes issues de l'*Acquis Communautaire*, sont abordées les relations à établir entre la Commission Européenne et les autorités des Etats Membres pour que chacune des parties puisse prendre les mesures adéquates en temps voulu. Enfin, les rapports publics de *La Documentation Française* portent sur une comparaison de la satisfaction des usagers des services publics et privés à l'égard des produits commercialisés.

Les sujets des cinq genres textuels étudiés sont divers et on peut remarquer également (voir exemples ci-dessous) que les chaînes de référence possèdent des différences de forme (types d'expressions référentielles utilisés) et de type de référent (référence à des humains et des non humains) suivant le genre :

- Corpus *Le Monde :*
  (1) **M. Giscard d'Estaing** a proposé une série d'économies. **Il s'**est demandé s'il était « urgent » de créer une chaîne éducative, **il s'**est interrogé sur le financement des interventions militaires ou humanitaires de la France dans le monde, ainsi que sur la pertinence de grands investissements, comme la grande bibliothèque.

- Corpus *Le Monde Diplomatique :*
  (2) **Le Centre de documentation d'études juridiques, économiques et sociales (CEDEJ),** installé au Caire, ne recoupe que partiellement les objectifs de **son** homologue libanais. Né dans le cadre de la coopération culturelle française de type classique, **il** a opéré, voilà trois ans environ, une mutation qui a suscité l'intérêt des partenaires égyptiens en multipliant **ses** activités avec l'aide de collègues dont quelques volontaires du service national actif.

- Corpus *Les Trois Mousquetaires :*
  (3) Car notre jeune homme avait **une monture**, et **cette** monture était même si remarquable, qu'**elle** fut remarquée.

- Corpus *Acquis Communautaire :*
  (4) Au cas où la parité de la monnaie d'**un État membre** par rapport à l'unité de compte définie ci-dessus serait réduite, le montant de la quote-part de capital versé par **cet État** sera ajusté, proportionnellement à la modification intervenue dans la parité, moyennant un versement complémentaire effectué par **cet État** en faveur de l'Agence et limité au

montant des avoirs effectivement détenus dans la monnaie de **cet État membre**.

- Corpus *La Documentation Française :*

(5) De ce fait, **la recherche de la satisfaction des usagers se** présente différemment selon le type de service public. On constate ainsi sans surprise que les services publics dont le mode d'action privilégié est la conviction, accordent une importance plus grande à **la recherche de la satisfaction de leurs publics**.

Il n'empêche que, globalement, c'est bien dans le cadre d'un rapprochement des méthodes des gestions publique et privée (développement des démarches « qualité », du contrôle de gestion et de la mesure des performances) qu'est réapparu l'intérêt porté à **la satisfaction de l'usager.**

Pour pouvoir mettre en place notre module d'identification automatique des chaînes de référence, nous étudions les spécificités des chaînes de référence suivant le genre.

1.3 Critères de l'étude

Pour cette étude, nous avons annoté manuellement les chaînes de référence pour en déterminer les propriétés pertinentes relativement à un genre donné[6]. L'étude des chaînes de référence que nous avons effectuée est inspirée des travaux de (Schnedecker, 2005). Ainsi, pour chaque genre, nous examinons les chaînes de référence suivant cinq critères :

- la longueur moyenne (en nombre de maillons) des chaînes de référence,
- la distance moyenne (en nombre de phrases) entre les maillons,
- la catégorie grammaticale privilégiée des maillons des chaînes de référence suivant le genre,
- la classe grammaticale des premiers maillons des chaînes de référence,
- la correspondance entre le thème phrastique et le premier maillon des chaînes de référence (sur ce point, cf. infra chapitres 1 et 2).

[6] Nous rappelons que nous avons mené notre étude sur des extraits (50 000 *tokens*) de notre corpus multi-genres.

1.3.1 Longueur moyenne des chaînes de référence

Pour ce premier critère, est comptabilisé le nombre moyen de maillons contenus dans une chaîne de référence suivant le genre (seules les chaînes de référence de trois maillons au moins sont comptabilisées, selon la définition de (Schnedecker, 1997) que nous suivons).

CORPUS	LONGUEUR MOYENNE
Le Monde	4
Le Monde Diplomatique	3,67
Acquis Communautaire	3
Les Trois Mousquetaires	9
La Documentation Française	3,4

Tableau 7 - Longueur moyenne (en nombre de maillons) des chaînes de référence suivant le genre textuel

L'étude a révélé quelques différences (voir Tableau 7). Par exemple, nous avons constaté que la longueur des chaînes de référence était de trois maillons en moyenne pour les lois européennes de l'*Acquis Communautaire* (6) alors qu'elle était trois fois plus élevée en moyenne pour le roman (7) :

(6) « L'Agence peut emprunter sur les marchés financiers d'un État membre dans le cadre des dispositions légales s'appliquant aux emprunts intérieurs, ou, à défaut de telles dispositions dans **un État membre**, quand **cet État membre** et l'Agence se sont concertés et se sont mis d'accord sur l'emprunt envisagé par celle-ci. L'assentiment des instances compétentes de **l'État membre** ne peut être refusé que si des troubles graves dans les marchés financiers sont à craindre.

(7) C'était **un bidet du Béarn**, âgé de douze ou quatorze ans, jaune de robe, sans crins à la queue, mais non pas sans javarts aux jambes, et qui, tout en marchant la tête plus bas que les genoux, ce qui rendait inutile l'application de la martingale, faisait encore également **ses** huit lieues par jour. Malheureusement les qualités de **ce cheval** étaient si bien cachées sous **son** poil étrange et **son** allure incongrue, que dans un temps où tout le monde se connaissait en chevaux, l'apparition **du susdit bidet** à Meung, où **il** était entré il y avait un quart d'heure à peu près par la porte de Beaugency, produisit une sensation dont la défaveur rejaillit jusqu'à **son** cavalier.

Cette différence significative entre la longueur des chaînes de référence de ces deux genres s'explique en ce que les lois européennes font intervenir de nombreux référents, donc que la compétition référentielle y est forte. En cas de compétition

référentielle, la redénomination [d'un nom propre] est considérée comme une fermeture du référent en cours (donc une ouverture d'une nouvelle chaîne de référence) (Schnedecker, 2005), d'où le faible nombre de maillons pour ce type de chaîne de référence. En revanche, dans l'extrait du roman, on assiste à de nombreux passages descriptifs, propices aux longues chaînes de référence.

1.3.2 Distance moyenne entre les antécédents

Nous avons déterminé ici le nombre de phrases séparant chacun des maillons d'une même chaîne de référence (0 pour la même phrase, 1 pour la phrase suivante, etc.). Pour ce second critère, nous avons observé que la distance entre les maillons des chaînes de référence des analyses politiques du *Monde* n'excède pas une phrase en moyenne, tandis que, pour les rapports publics de *La Documentation Française*, la distance est supérieure à deux phrases entre le second maillon et le troisième (voir Tableau 8). Pour ce dernier genre, on retrouve une technique particulière : introduction du référent, maintien (sans compétition référentielle, donc la distance est plus grande entre les maillons 2 et 3 de la chaîne) et rappel avant la fermeture de la chaîne de référence (Goutsos, 1997).

CORPUS	RANG DU MAILLON			
	1à 2	2 à 3	3 à 4	4 à 5
Le Monde	0,4	1	1	-
Le Monde Diplomatique	0,3	1,3	0	2
Acquis Communautaire	0	1,25	-	-
Les Trois Mousquetaires	0	0,25	0	1,3
La Documentation Française	0,4	2,4	0,5	-

Tableau 8 - Distance moyenne entre les maillons (en nombre de phrases) suivant le genre textuel[7]

1.3.3 Catégorie grammaticale la plus fréquente des maillons des chaînes de référence

Ce troisième critère permet d'identifier la ou les catégorie(s) grammaticale(s) privilégiée(s) des maillons des chaînes de référence issues de chaque genre. Ainsi, on constate (Tableau 9) une part importante de noms propres (30,8%) dans les

[7] Par soucis de lisibilité, ne sont reportés dans ce tableau que les maillons des rangs 1 à 5 (même si les chaînes de référence des *Trois Mousquetaires* ont une longueur moyenne de neuf maillons).

maillons des chaînes de référence du *Monde*, alors que cette catégorie n'est pas représentée dans les autres genres textuels[8].

CORPUS	CATEGORIE GRAMMATICALE DES MAILLONS					
	Np	Pr	GNdef	GNindef	GNposs	GNdem
Le Monde	30,8	15,4	23,1	0	23,1	7,7
Le Monde Diplomatique	0	25	50	0	25	0
Acquis Communautaire	0	10	20	40	10	20
Les Trois Mousquetaires	0	35,9	20,5	10,3	28,2	5,1
La Documentation Française	0	33,3	33,3	16,7	16,7	0

Tableau 9 - Répartition des catégories grammaticales des maillons des chaînes de référence suivant le genre (en %)

Aussi, la moitié des maillons des chaînes de référence des éditoriaux du *Monde Diplomatique* sont-ils des groupes nominaux définis (GNdef) alors que 40% des maillons sont des groupes nominaux indéfinis (GNindef) dans les lois européennes de l'*Acquis Communautaire*. Cette dernière observation pour les lois européennes est corrélée au faible nombre de maillons relevés en moyenne dans les chaînes de référence (critère 1). En effet, les mesures décidées par la Commission Européenne ont un caractère générique qui doit s'appliquer à tout Etat Membre de la Communauté ; d'où la présence massive d'indéfinis (on aura par exemple : « un Etat Membre », « une décision », « une mesure »). Les deux derniers genres textuels comptent des pronoms (environ un tiers) et respectivement 28,2% de possessifs (roman *Les Trois Mousquetaires*) et 33,3% de GN définis (rapports pulbics de *La Documentation Française*) dans leurs maillons.

Les disparités observables entre les fréquences des catégories des maillons présentes dans chaque genre textuel sont, dans une certaine mesure, révélatrices des spécificités des chaînes de référence suivant le genre.

1.3.4 Catégorie grammaticale privilégiée du premier maillon des chaînes de référence

Nous nous sommes intéressée ici à définir la catégorie privilégiée des maillons utilisés en première mention (Schnedecker, 1997) dans chacun des genres. Pour les analyses politiques, on relève essentiellement des noms propres en première mention, avec des redénominations de « Valéry Giscard d'Estaing » dans

[8] Plusieurs noms propres étaient bien présents dans les autres genres textuels étudiés, mais seuls les noms propres figurant dans des chaînes de référence de trois maillons au moins (suivant la définition des chaînes de référence que nous avons adoptée) ont été comptabilisés et reportés dans le tableau.

l'ensemble du document. Ce sont des descriptions définies qui se retrouvent souvent en première mention des chaînes de référence des éditoriaux. En effet, comme les sujets abordés dans les éditoriaux concernent un point de vue à propos des actualités du moment, les auteurs considèrent que les références aux entités présentes dans leurs écrits sont acquises par leurs lecteurs. Il en est de même pour les rapports publics qui comptent en majeure partie des groupes nominaux définis dans les premiers maillons des chaînes (par exemple, « la satisfaction des clients », « la mesure de la satisfaction des usagers »). Ici, la mesure de la satisfaction des clients est le sujet (thème) du document tout entier et est abordée sous divers angles (la place, la mesure, la recherche de la satisfaction des clients).

Enfin, les groupes nominaux indéfinis dominent dans les premiers maillons des chaînes de référence des lois européennes, cela étant lié à la portée générique des lois européennes. Il en est de même pour les premiers maillons des chaînes de référence du roman, puisque l'extrait étudié présente les acteurs (« un jeune homme », « une monture ») qui feront l'objet de reprises dans la suite du discours (« le Gascon », « le jeune d'Artagnan », « cette monture », etc.).

1.3.5 Correspondance entre le thème phrastique et le premier maillon des chaînes de référence

Pour ce dernier critère, nous souhaitions savoir dans quelle mesure il était possible de regrouper les chaînes de référence contenant le même thème phrastique (entendu ici comme le sujet de la phrase). Nous avons donc comptabilisé les cas où le premier maillon des chaînes de référence coïncidait avec le thème de la phrase en cours. On observe ainsi que le premier maillon est le thème phrastique dans 80% des cas pour les analyses politiques *du Monde* par exemple, mais qu'il n'est que de l'ordre de 40% pour les rapports publics issus de *La Documentation Française* (Tableau 10).

CORPUS	CORRESPONDANCE THEME PHRASTIQUE
Le Monde	80
Le Monde Diplomatique	100
Acquis Communautaire	60
Les Trois Mousquetaires	60
La Documentation Française	40
Moyenne	68

Tableau 10 - Correspondance entre le premier maillon des chaînes et le thème phrastique (en %)

Les cinq critères utilisés dans notre étude (longueur des chaînes de référence, distance entre les maillons, catégorie grammaticale du 1er maillon, catégorie grammaticale la plus fréquente et correspondance avec le thème phrastique) permettent de mettre en place une typologie des chaînes de référence suivant leur genre d'accueil.

2 Typologie des CR suivant le genre textuel

L'étude des chaînes de référence dans un corpus multi-genres a-t-elle permis de mettre au jour leurs propriétés spécifiques suivant le genre textuel (voir Tableau 11). Il ressort de cette étude que la composition des chaînes de référence (nature du premier maillon, distance moyenne entre les antécédents, catégorie grammaticale des maillons, etc.) est effectivement fortement tributaire du genre d'accueil des chaînes de référence.

Ces « tendances » vont être utilisées pour configurer notre module d'identification des chaînes de référence selon le genre textuel. Par exemple, si le document à traiter est une analyse, notre système sera paramétré pour identifier des chaînes de référence courtes (d'une longueur moyenne de quatre maillons), qui débuteront de préférence par un nom propre et dont les maillons seront compris dans la phrase en cours ou dans la phrase suivante. Le nom propre coïncidera souvent avec le thème phrastique, ce qui signifie que plusieurs chaînes de référence indiqueront le même thème.

corpus (50 000 *tokens*) / critères	Le Monde	Le Monde Diplomatique	Acquis Communautaire	Les Trois Mousquetaires	La Documentation Française
Longueur moyenne	4	3,7	3	9	3,4
Distance moyenne entre antécédents (nb de phrases)	0,8	0,9	0,6	0,4	1,1
Catégorie 1er maillon	Noms propres	GN définis	GN indéfinis	GN indéfinis	GN définis
F des maillons	30 % Np	50 % GN définis	40 % GN indéfinis	36 % pronoms	- 33 % pronoms - 33 % GN définis
Correspondance thème -1er maillon	80 %	100 %	60 %	60 %	40 %

Tableau 11 - Typologie des chaînes de référence suivant le genre textuel

Dans la section suivante, nous proposons une étude quantitative et qualitative (quelle est la composition des chaînes de référence ?, quel type de référents

humains y trouve-t-on ?) fine des chaînes de référence dans un corpus issu d'un sous-genre journalistique : le fait divers.

3 Etude de cas : les faits divers

Dans cette section, nous présentons une étude de l'impact des genres sur la composition des chaînes de référence dans les faits divers menée en collaboration avec C. Schnedecker[9].

3.1 Objectifs de l'étude

Cette étude porte sur la manière dont les chaînes de référence sont conditionnées par le genre de leur texte d'accueil. Nous avons choisi de prendre comme genre d'appui les faits divers car ils offrent plus d'un avantage. D'une part, il s'agit de textes courts, ce qui permet de rendre compte de l'intégralité des chaînes de référence. D'autre part, ils permettent d'appréhender un genre de presse dont les caractéristiques thématiques et structurelles sont désormais bien établies (*cf.* Barthes, 1964 ; Auclair, 1970 ; Dubied, 2000), ce qui facilite théoriquement l'établissement de corrélations. Enfin, la concentration des phénomènes sur des formes brèves est de nature, peut-être, à faire ressortir des problèmes passés jusque-là inaperçus.

Nous avons établi un corpus de 46 textes (9 838 mots/59 097 caractères), indexés sous la rubrique des Faits divers du *Républicain Lorrain*, quotidien régional de la région lorraine (édition de Metz) et collectés durant l'été 2011 sur le site du journal. Notre objectif est de déterminer quelles sont les expressions référentielles mobilisées dans ce genre, quel est leur degré de solidarité, comment elles sont initiées et reliées les unes aux autres ; bref, en quoi elles répondent à un mode de composition précis.

Dans cette optique, nous commencerons par rappeler quelques caractéristiques notoires des faits divers, susceptibles de peser sur la composition des chaînes de référence. Nous procéderons à une étude quantitative du corpus construit pour cette étude, qui fera ressortir le matériau linguistique des chaînes de référence des faits divers et, le cas échéant, permettra de dégager quelques modèles de chaînes. Enfin, nous procéderons à l'étude qualitative de notre corpus, en insistant particulièrement sur quelques spécificités référentielles des faits divers de nature à questionner les approches au long cours de la coréférence.

[9] Cette étude a fait l'objet d'une publication dans les actes du Congrès Mondial de Linguistique Française (Schnedecker et Longo, 2012).

3.2 Les faits divers : quelques caractéristiques

3.2.1 Caractéristiques thématiques

Suivant la définition de (Fragnon, 2007 : 254), les faits divers « narrent des événements sociaux, eux-mêmes à la marge de l'espace social », dont ont été soulignés les écarts par rapport au déroulement quotidien des choses : ainsi Auclair a-t-il parlé de « rupture de l'univers réglé où l'homme trouve sa sécurité », (Barthes, 1981) de « déviations causales » : « en vertu de certains stéréotypes, on attend une cause, et c'est une autre qui apparaît ». Il est question aussi, dans la littérature, de dérogation à une norme ou d'extraordinaire[10]. L'exemple (8) illustre la déviation causale dont parle Barthes : un moniteur d'auto-école est censé connaître la règlementation en matière de circulation routière et, théoriquement, devrait éviter les excès de vitesse. Or, non seulement il commet une infraction mais celle-ci est, qui plus est, d'importance :

> (8) **Un automobiliste de 28 ans**, exerçant la profession de moniteur d'auto-école, a été contrôlé samedi à plus de 200 km/h au lieu des 130 km/h autorisés sur l'autoroute A16 à la hauteur de Pont-de-Metz (Somme), près d'Amiens. **Le conducteur**, contrôlé à 214 km/h, a eu son permis retiré et la voiture est partie à la fourrière.

3.2.2 Caractéristiques structurelles

3.2.2.1 Brièveté

Le fait divers se caractérise par sa brièveté[11]. Les textes de notre corpus comptent en moyenne 208 mots[12]. Structurellement,

> « <Ces> brèves sont composées d'un titre, d'une indication géographique et du texte proprement dit. […] Par souci de concision, les brèves répondent aux questions fondamentales (qui ? quoi ? quand ? où ? comment ?) : lieux et horaires de l'incident, sujet de l'acte délictueux, victimes, et, actions de l'anti-sujet policier. » (Fragnon, 2007 : 256),

[10] Cf. la synthèse de (Dubied, 2004 : 84-85).

[11] Il existe aussi des faits divers longs composés d'articles judiciaires ou dont l'importance est telle qu'elle mérite une large couverture (*cf.* Fragnon, 2007 : 256).

[12] Le plus court compte 52 mots et le plus long 579.

C'est ce que montre (9) où la mention du lieu figure généralement à gauche du titre ; elle est reprise dans le texte sous la forme d'une expansion du syntagme nominal (« une buraliste de Nancy ») ou d'un circonstant localisateur.

(9) **Nancy**. Buraliste braquée

(p1) **UNE BURALISTE DE NANCY** a été agressée, hier à 6h, alors qu'ELLE ouvrait SON commerce. (p2) << Deux hommes encagoulés et gantés ont surgi derrière ELLE alors qu'ELLE SE dirigeait vers la réserve de SON bar-tabac. (p3) Chacun portait une arme de poing. (p4) Ils ont bousculé LA GERANTE QUI est alors tombée à terre, avant de demander le coffre. (p5) Terrorisée, LA VICTIME a expliqué qu'il n'y en avait pas. (p6) Ils se sont fait remettre 4 000 € en chèques et espèces. (p7) Les deux agresseurs ont ensuite pris la fuite dans une direction inconnue, laissant LA BURALISTE sous le choc. (p8) ELLE a été soignée à l'hôpital central de Nancy pour une estafilade à l'épaule gauche, due à SA chute. >> (p9) L'affaire est confiée à **la sûreté départementale de Meurthe-et-Moselle.**

Les gendarmes et les agents de l'UTR de Bitche ont coupé la circulation sur la RD 35, un axe fréquenté qui relie Bitche à Deux-Ponts (Allemagne).

L'information qui constitue le fait divers apparaît dans les titres et/ou sous-titres, eux-mêmes repris et expansés dans la première phrase du texte ; elle est développée par le biais d'une narration constituant la partie centrale du fait divers (entre chevrons dans l'exemple (9))[13]. La partie narrative est, le cas échéant, délimitée par des syntagmes nominaux à caractère métadiscursifs comme les faits, par exemple dans (10) :

(10) L'auteur d'une fusillade survenue dans le cadre d'une opération de représailles, à la mi-juin, à Frouard, s'est constitué prisonnier cette semaine auprès de la gendarmerie. Il a été écroué en dépit de ses dénégations. Il reconnaissait avoir fait feu avec son fusil, mais pas avoir voulu tuer. Il est poursuivi néanmoins pour tentative d'homicide volontaire. Un mandat de recherche était lancé contre lui. **Les faits** s'étaient déroulés en deux temps, route de Liverdun. *Un passant avait été pris à partie et bastonné à coup de chaîne par deux frères turbulents, vivant dans une maison de MMH. Un peu plus tard, la victime avait envoyé des amis donner une leçon à ses agresseurs.* […]

[13] On peut à l'instar de certains auteurs y voir la réalisation du fameux schéma quinaire de Propp, avec, dans le cas de (9), la force transformatrice (p2-3), l'action (p4-6), la force équilibrante (p7-8) et la situation finale (p9).

3.2.2.2 Les individus dans les faits divers

L'une des particularités du fait divers tient à ce qu'ils racontent des événements touchant « des gens de tous les jours qui, à un moment de leur vie, ont vécu une situation exceptionnelle, souvent dramatique, et ont accompli un geste qui a sauvé des vies. » (Dubied, 2004 : 53). Dubied (2004 : 227) évoque également « des quidams, […] saisis dans leur vie de tous les jours ; des « Monsieur-tout-le-monde ». De là viendrait que « les termes employés pour caractériser le sujet sont les plus neutres possibles » (Fragnon, 2007 : 257). Nous aurons l'occasion d'en évaluer les répercussions sur les chaînes de référence.

On peut d'ores et déjà souligner que cela transparaît dans la titraille. En effet, 67% des titres font référence à un individu humain dont les modes de désignation sont assez systématiques. Nous avons ainsi repéré deux « patrons », schématisés sous **i.** et **ii.**, classés en fonction de la structure dominante : les SN expansés par un participe passé employé comme adjectif dominent dans 39% des cas (nous y incluons les cas où le titre se limite au participe passé) ; ils sont suivis (20%) par des phrases « complètes » instanciant le protagoniste principal, le plus souvent en fonction grammaticale de sujet. On observe, à part (cf. **iii.**), des SN, parfois déverbaux, dont les arguments sont laissés vides :

i. **[Déterminant + Nom] + participe passé : 39%**
Accident mortel : conducteur poursuivi
Un homme tué par balles sur une aire de l'A31
Marseille. Garçonnet tué par un chauffard en fuite
Uckange. Poignardé à son domicile

ii. **Phrase complète : 20%**
Corse : un détenu violent s'évade d'un hôpital
Une octogénaire périt dans un incendie

iii. **Syntagme nominal [déverbal] : 20%**
Nancy. Mort suspecte dans un incendie
Imling. Le sauvetage de papy courage

3.2.3 Caractéristiques fonctionnelles

Certaines des visées du fait divers[14] ont un lien direct avec son « personnel » comme le nomment (Dubied et Lits, 1999) à la suite de Hamon. En effet, qu'il s'agisse de dénoncer les dysfonctionnements sociaux et d'en stigmatiser des groupes (Fragnon : 262-265) ou, au contraire, de décrire la condition humaine à des fins cathartiques ou d'« agrégation tribale » (expression de Maffesoli, citée *in*

[14] Voir (Fragnon, 2007), (Dubied, 2004 : 84-85) et (Deleu, 2005 : 14-15).

Dubied : 73-74), l'individu y est montré dans ses aspects les plus « stéréotypés », voire confiné dans des « rôles thématiques » : le petit voyou, le criminel par accident, la victime âgée, etc. Bien entendu, cela aura des conséquences sur la dénomination des personnages et leur expression référentielle.

3.3 Les expressions référentielles dans les faits divers : les grandes tendances

3.3.1 Quelques chiffres

Sur un corpus de 9838 mots, nous avons recensé un total de 905 expressions référentielles référant à des personnes *i.e.* des entités individuelles (le motard sarrois en (11)) ou collectives (en italiques ou soulignées dans le texte), ce qui équivaut à une expression référentielle tous les 11 mots. Pour des textes d'une longueur moyenne de 208 mots, la densité est donc relativement importante. Cela signifie, en effet, que 19% environ des unités du texte sont des expressions référentielles renvoyant à des référents humains. Le texte (11) montre, en effet, qu'un référent au moins est instancié toutes les lignes :

(11) Schweyen **Un motard sarrois** tué dans une collision
Alors qu'**il** venait tout juste de passer la frontière avec un groupe d'amis et **ø se** dirigeait vers Bitche, **un motard sarrois** a trouvé la mort hier, à 10h, sur la RD 35 à hauteur de Schweyen.
Le pilote menait le groupe de deux-roues. **Il** aurait tenté de dépasser un camion dans une courbe à droite, pourtant marquée par une ligne blanche continue. Un autre poids lourd arrivait en face. La collision était inévitable. **Le pilote** et **sa** moto ont été projetés sur le bas-côté. **Joachim Platt**, 52 ans, est mort sur le coup.
Des pompiers français et allemands sont intervenus sur les lieux de l'accident. **Ces derniers** ont pris en charge DEUX CHAUFFEURS ROUTIERS D'OUTRE-RHIN, légèrement blessés et surtout choqués.
LES GENDARMES ET LES AGENTS DE L'UTR DE BITCHE ont coupé la circulation sur la RD 35, un axe fréquenté qui relie Bitche à Deux-Ponts (Allemagne).

Les 46 textes du corpus totalisent 89 chaînes de référence[15]. Plus de 60% de chaînes de référence comportent un nombre situé entre 3 et 6 maillons, ce qui fait que, du point de vue de la longueur moyenne, elles sont conformes à ce qui a été constaté pour le genre journalistique par (Jenkins, 2002 ; Schnedecker, 2005).

Les chaînes de référence se répartissent comme suit (cf. Tableau 12) dans les différents textes : la majorité des textes – presque 40% – sont centrés sur un personnage et un tiers sur 2. Par contraste, les textes restants, centrés sur plus de 2 personnages, ne constituent qu'un quart du corpus.

Nombre de textes	Nombre de chaînes de référence / texte	Pourcentage de textes concernés
18	1	39%
16	2	35%
9	3	19.5%
3	4	6.5%
46	--	100%

Tableau 12 - Répartition des chaînes de référence selon les textes

3.3.2 Catégories grammaticales des expressions référentielles

Les expressions référentielles se répartissent en différentes catégories grammaticales, que synthétise le Tableau 13 :

Catégories d'expressions référentielles	Npr	SNØ	SN indéfinis		SN poss.	SN définis		SN dém.	Pronoms					Dét. poss.
			nus	expansés		nus	expansés		dém.[16]	réfléchis	Ø	relatifs	Pers.	
Nombre d'occurrences	29	8	71	75	25	175	114	11	5	63	14	29	173	104
Pourcentages	3.2%	1%	8%	9%	2.5%	20%	12.5%	1.3%	0.5%	7%	1.5%	3%	19%	11.5%
		18%			35%			1.8%	11.5%				30.5%	

Tableau 13 - Répartition des catégories grammaticales des expressions référentielles

[15] Comme le montre l'annotation du texte (11), sont considérés comme « maillons » les syntagmes nominaux ou pronominaux (pronoms personnels, réfléchis, etc.) explicites, les anaphores zéros et déterminants possessifs du fait qu'ils équivalent à « de SN ». Nous n'avons pas pris en compte ce que certains (cf. Cornish, 1986 ; Landragin, 2011) nomment « indices », comme par exemple les accords des participes passés de « *Ces derniers ont pris en charge deux chauffeurs routiers d'outre-Rhin, légèrement blessé**s** et surtout choqué**s** »* qui sans référer à proprement parler réinstancient d'une certaine manière les référents dans le texte.

[16] Même si le regroupement des expressions démonstratives nominales et pronominales est peu orthodoxe, ce choix nous est paru pour cette étude pertinent compte tenu de la capacité commune des démonstratifs à opérer une forme de rupture avec le cotexte.

3.3.2.1 Un équilibre inattendu entre les expressions dites de haute et moyenne accessibilité référentielle

Suivant la terminologie d'(Ariel, 1990), les expressions dites de haute accessibilité référentielle comme le pronom personnel, auquel nous ajoutons le déterminant possessif du fait de son équivalence avec le pronom personnel, constituent 30% des expressions référentielles du corpus, ce qui, ajouté aux expressions pronominales liées syntaxiquement (anaphore zéro, pronoms réfléchis et relatifs), aboutit à un pourcentage de 42%. De ce fait, les pronoms dominent nettement le corpus. Mais ils sont suivis de très près par les expressions dites de moyenne accessibilité référentielle dont le pourcentage s'élève à 36,3% et qui se répartissent entre les SN définis (35% dont 20% de SN « nus », 12,5% de SN expansés et 2,5% de possessifs) auxquels s'ajoutent 1,3% de démonstratifs.

A priori, ces chiffres démentent les pourcentages d'(Ariel, 1990) dont le corpus fait apparaître 70% de pronoms personnels, ce qui les met en tête des emplois en tant que marques du topique, de la continuité référentielle. Cela étant, les pourcentages que nous obtenons s'expliquent. En effet, les textes de notre corpus sont pluri-référentiels, *i.e.* instancient au moins deux référents humains. Or, le nombre de référents – et, partant, les modalités de la reprise anaphorique – augmente les risques d'ambiguïtés pronominales. C'est ce que montre (13) – version remaniée de (12) – où la première apparition du pronom serait susceptible de brouiller l'interprétation coréférentielle[17] :

(12) Football : **le président d'un club** agressé par un ex-joueur
Les dirigeants du Sporting-Club de Moulins-lès-Metz pensaient en avoir terminé avec les problèmes extra-sportifs. Ils les ont violemment rattrapés le week-end dernier. A l'issue de l'entraînement de l'équipe seniors, **le président du club de football** a été agressé par *un ancien joueur*, exclu après avoir été suspendu par la fédération pour des menaces sur un arbitre. *L'auteur* a profité de l'obscurité du parking pour **le** frapper au visage, avec une batte de base-ball. […]

(13) A l'issue de l'entraînement de l'équipe seniors, **le président du club de football** a été agressé par *un ancien joueur*, exclu après avoir été suspendu par la fédération pour des menaces sur un arbitre. ***Il*** a profité de l'obscurité du parking pour **le** frapper au visage, avec une batte de base-ball.

[17] En fait, comme le montre (Schnedecker, 1997 : 52-63) cette ambiguïté n'est, en l'occurrence, que potentielle.

De ce fait, le fort pourcentage d'expression de moyenne accessibilité référentielle pourrait, en première instance, prévenir les risques d'ambiguïtés[18].

3.3.2.2 Les expressions de faible accessibilité référentielle : domination des syntagmes nominaux indéfinis

Les expressions de faible accessibilité référentielle s'élèvent à 21%. 3,2% d'entre elles sont constituées de noms propres[19]. Ceux-ci ne figurent que dans 17 (soit 37%) des 46 textes de notre corpus et rarement en première mention (dans 6 des 87 chaînes de référence, soit 6.9% des cas). C'est le cas de (11) vu plus haut où le nom propre apparaît dans le 7ème maillon[20]. En revanche, (14) illustre un cas de nom propre en position de 1er maillon :

> (14) Waltembourg Percuté et tué en marchant sur la RN4
> **Frédéric Jambois**, 38 ans, d'Imling, est décédé, hier matin, sur la RN4. Percuté par une voiture, il marchait avec des amis sur la voie rapide. […]

La majorité des expressions référentielles de faible accessibilité référentielle est constituée par les syntagmes nominaux indéfinis dont le pourcentage s'élève à 18% et se répartit de manière équilibrée entre les syntagmes nominaux « nus » (8%) et expansés (9%)[21]. A quoi s'ajoutent 1% de syntagmes nominaux sans déterminants qui figurent exclusivement dans les titres (cf. supra 3.2.2.2).

3.3.3 Des patrons de chaînes de référence ?

Il est difficile de dégager des patrons de chaînes dans les faits divers, tant est grande la disparité des chaînes de référence, aussi bien du point de vue des expressions référentielles que de leur longueur. Néanmoins, il ressort de nos analyses que les chaînes de référence des faits divers sont plutôt brèves, comptant en moyenne 3,42 maillons. Plus d'un tiers des chaînes de référence en comptent 3 à 4 (*cf.* Figure 8) ; un deuxième tiers (29%) 5 et 6. Cette longueur tient évidemment à la brièveté inhérente au genre du fait divers lui-même et se rapproche de celle des textes expositifs rappelée plus haut.

[18] Nous verrons plus bas que d'autres raisons, propres au genre, motivent ce type d'expressions référentielles.

[19] Qui constituent 17,5% de cette catégorie.

[20] Sur cette question, *cf.* (Charolles, 1987).

[21] Les SN indéfinis nus constituent 40% des expressions de faible accessibilité référentielle et les indéfinis expansés 42,5%, soit 82,5% du total.

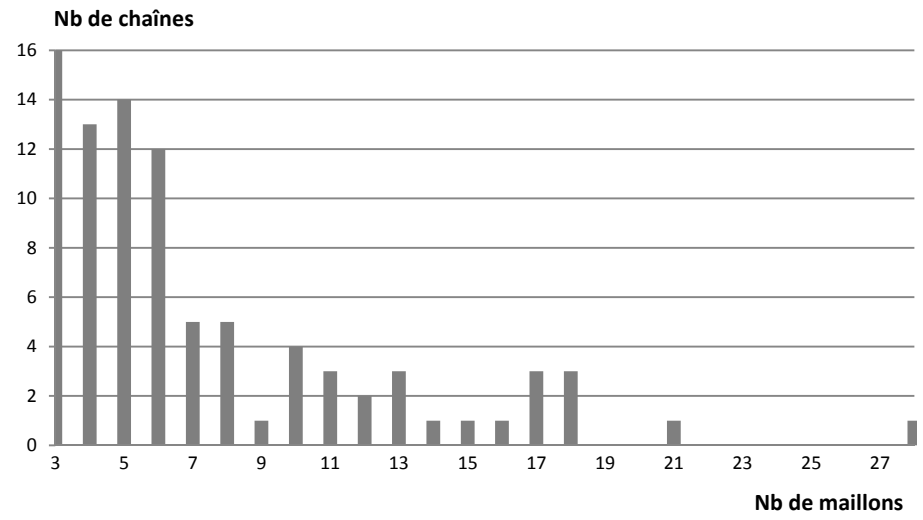


Figure 8 – Longueur des chaînes de référence dans les faits divers

Deuxièmement, le premier maillon est constitué à 45% de SN indéfinis, à 27% de SN définis et, pour les plus faibles pourcentages, à 8% de pronoms personnels ou de déterminants possessifs. Dans 84% des chaînes de référence, le 1er maillon du titre coïncide avec la 1ère chaîne du texte, comme dans (15) et (16) :

(15) **Il** tente de s'immoler dans un terrain vague
Un jeune homme âgé de 29 ans, demeurant à Blainville-sur-l'Eau, en Meurthe-et-Moselle, a tenté de mettre fin à **ses** jours, samedi dernier, en **s'**immolant par le feu.

(16) Mirecourt. **Gendarme poignardé** : **sa** santé s'améliore
Les jours **du commandant de la communauté de brigade de Mirecourt** (Vosges) ne semblent plus en danger. Poignardé à la gorge par un individu, **le militaire** a été opéré, vendredi soir, au CHU de Nancy.

Troisièmement, en prenant en compte uniquement les 3 premiers maillons – donc le seuil-plancher des chaînes dans notre conception –, on a pu dégager 3 patrons correspondant à des débuts de chaînes de référence configurées comme suit :

i. SNdéf pro pro : 9%

(17) **Le conducteur de la voiture**, un homme de 83 ans, originaire d'Erstroff, est indemne. **Il** est rentré chez **lui** après l'accident.

ii. SNindéf. pro pro : 9%

(18) **un blessé** et des dégâts
Vendredi après-midi, alors qu'**il** rentrait chez **lui** à pied, un habitant de Folschviller a été surpris par l'orage de grêle qui s'est soudainement abattu sur la région de Saint-Avold … .

iii. SNdéf. SNdéf. SNdéf. : 7%

(19) Alertés par des riverains, à 16h30, **les sapeurs-pompiers** ont déployé la grande échelle, ainsi que le fourgon pompe-tonne de Forbach. Mais à l'arrivée **des secours**, la locataire, Marie-Louise Pezzotta, avait déjà succombé, suite à l'inhalation de monoxyde de carbone. Âgée de 85 ans, la victime vivait seule, selon les premiers éléments de l'enquête. Selon la police de Forbach, si la cause exacte du sinistre reste encore à déterminer, la thèse de l'accident est privilégiée. « Il est difficile de dire dans quelle pièce l'incendie s'est déclaré », indique l'adjudant-chef **des sapeurs-pompiers** Laurent Wetzl, qui a dirigé l'intervention.

Notons enfin que 18% des chaînes de référence sont dépourvues de pronoms personnels. Nous y reviendrons dans la section 3.4.1.

3.3.4 Premier bilan

Cette analyse quantitative aura fait ressortir la spécificité des chaînes de référence des faits divers à quatre points de vue :

- la pluri-référentialité des textes : 61% d'entre eux comprennent au moins 2 référents, ce qui les distingue d'autres genres de la presse comme, par exemple, les portraits ;
- le type d'expression référentielle dominante : ce sont les syntagmes nominaux, tous déterminants confondus, qui dominent (55% des expressions référentielles) alors que les pronoms n'excèdent pas 42% ;
- le grand « perdant » des faits divers est le nom propre qui n'atteint que 3,2% des cas et dans des emplois qui diffèrent de ce qui a été observé dans les romans ou les portraits ;
- la composition : les chaînes de référence sont plutôt brèves (3,42 maillons en moyenne), plutôt hétérogènes, tant au plan des catégories grammaticales instanciées que des têtes lexicales des syntagmes nominaux et assez difficiles à modéliser, puisque seuls 3 patrons ont émergé.

C'est toutefois l'analyse qualitative qui, comme on va le voir, est la plus riche d'enseignements.

3.4 Les expressions référentielles dans les faits divers : analyse qualitative

3.4.1 Diversité des noms d'humains

Au plan qualitatif, la première caractéristique notoire des chaînes de référence de faits divers tient à la diversité des têtes lexicales des syntagmes nominaux dont le Tableau 14 donne un aperçu[22].

Qui plus est, la diversité nominale est à la fois récurrente dans les faits divers du corpus, distribuée sur l'ensemble des référents instanciés dans un texte comme l'illustre (20)[23], et inhérente aux syntagmes nominaux anaphoriques dont la tête lexicale varie d'une reprise à l'autre, comme le montrent, dans (20) ci-dessous, la CR du personnage principal de l'octogénaire et (9) supra.

(20) Forbach. *UNE OCTOGENAIRE* périt dans un incendie

Hier après-midi, un incendie s'est déclaré au quartier du Creutzberg à Forbach. Le feu a pris au premier étage d'une maison mitoyenne, au 6, avenue de Stiring-Wendel. Alertés par DES RIVERAINS, à 16h30, **LES SAPEURS-POMPIERS** ont déployé la grande échelle, ainsi que le fourgon pompe-tonne de Forbach. Mais à l'arrivée DES SECOURS, *LA LOCATAIRE*, *MARIE-LOUISE PEZZOTTA*, avait déjà succombé, suite à l'inhalation de monoxyde de carbone. Âgée de 85 ans, ***LA VICTIME*** vivait seule, selon les premiers éléments de l'enquête.

Selon LA POLICE DE FORBACH, si la cause exacte du sinistre reste encore à déterminer, la thèse de l'accident est privilégiée.

« Il est difficile de dire dans quelle pièce l'incendie s'est déclaré », indique L'ADJUDANT-CHEF **DES SAPEURS-POMPIERS** LAURENT WETZL, qui a dirigé l'intervention.

En effet, les flammes ont entièrement ravagé l'appartement.

LES SAPEURS-POMPIERS ont longuement inspecté la toiture et les conduites de façon à empêcher tout risque de reprise du feu.

[22] La classification est encore hasardeuse, mais le fait est que l'étude des noms d'humains reste à faire. Nous n'avons pas inclus les noms à occurrence unique.

[23] Par contraste, par exemple, avec les portraits journalistiques où la diversité des maillons se limite généralement à la chaîne de référence du référent principal.

3.4.2 Des sous-catégories de noms en nombre limité

Les types de noms sont limités à 4 sous-catégories (voir Tableau 14) :

- les noms « généraux » (*homme, femme*), auxquels s'ajoutent les noms dits de phase, qui désignent l'individu par ses caractéristiques identitaires fondamentales (sexe et âge) ; ces noms constituent un quart des emplois ;
- les noms de fonction/profession dont le pourcentage s'élève à 21% mais dont une grande partie (19%), « incontournable » dans les faits divers, renvoie notamment aux forces de l'ordre, aux sauveteurs ou à la justice (désignée sous le nom de *parquet* (16 occurrences), cf. *supra*, ex. (11), (19)). Ce sont d'ailleurs ces noms qui composent les chaînes les plus régulières, fondées sur la réitération de la tête lexicale, et donnent lieu au patron **vi.** ;
- la troisième catégorie, qui comprend 33% des syntagmes nominaux, réunit 1) les noms indiquant les relations entre le protagoniste principal du fait divers et ses proches (famille, voisinage mais aussi ses proches géographiquement, saisis dans leur ancrage spatial pourrait-on dire) ; auxquels s'ajoutent 2) les noms de gentilés et, enfin, 3) les noms saisissant les actants plus ou moins directs du « drame » décrit : acteurs (*agresseur, vengeur*) ou patients (*blessé, victime*) ainsi que les personnages plus extérieurs, simplement spectateurs (*témoin*) ;
- la dernière (1%) rassemble les rares noms indiquant l'empathie du locuteur (cf. *supra* 3.2.2).

Noms Généraux (16%)	**Noms de Phase (9%)**	**Noms de Fonction/ Profession (21%)**	**Noms Relationnels (32%)**			**Gentilés (1%)**	**Axiologique Empathique (1%)**
			relations sociales	**ancrage spatial**	**ancrage situationnnel**		
Personne (7) Individu (5) Homme (49) Femme (17)	Jeune / vieil homme (7) Personne âgée (4) Enfant (12) Petit garçon (3) Fillette (3) N en *-aire* (9) (*quinquagénaire*) Jeunes (4) Retraité (4)	Gendarmes (19) Gendarmerie (8) Enquêteurs (6) Police (24) Policier (13) Pompiers (24) Buraliste Automobiliste (8)	Mère (5) Fille (10) Fils (3) Compagnon (6) Voisin (18)	Habitant (12) Riverain (4) Occupant (3) Résident (2)	Agresseur (2) Vengeur Blessé (16) Victime (35) Suspect (6) Témoin (6) Détenu (2) ***Secours (23)*** ***Touriste*** Rescapé Conducteur (20) Gérante ***Auteur (4)***[24]	Meusien Mosellan Maizièrois Messin (3)	Bon samaritain Chauffard (3) Malheureux (2)

Tableau 14 – Classement (provisoire) des têtes lexicales des syntagmes nominaux dans les faits divers

[24] Les noms en italiques signalent un classement « faute de mieux ».

Là encore, les fonctions d'agrégation tribale ou d'identification/projection cathartique attribuées aux faits divers, rappelées plus haut, sont largement servies par le lexique qui sert, pour l'essentiel, à décliner l'identité des protagonistes au moyen des noms (sexe, âge, profession), communs à tous les humains et à les resituer dans un environnement ou dans des situations relativement ordinaires, banals. Autant d'éléments qui ramènent le personnel des faits divers à « monsieur tout le monde » et créent une certaine proximité avec le lecteur de faits divers.

3.4.3 La relative uniformité et objectivité des informations délivrées par les modifieurs

Près de 21% des expressions référentielles du corpus sont expansées, principalement par des compléments du nom (CDN) ou des adjectifs (éventuellement des participes passés employés comme adjectifs)[25], ainsi que le synthétise le Tableau 15. Ces informations concernent généralement le personnage principal du fait divers.

Type de modifieur	Complément du nom	Adjectif
Nombre	106	58
Pourcentage	56%	31%
Sous-catégories	Localisateur (32%)[26]	Localisation (20%)
	Age (12%)	Participe passé (31%)

Tableau 15 – Répartition et information des modifieurs

Les CDN dispensent, dans presque la moitié des cas, des informations relatives à la localisation géographique des référents (21)[27] – et ce, de manière systématique pour les forces de l'ordre, sauveteurs, etc. – ou à leur âge (22) ; quant aux adjectifs ou assimilés (cf. *supra* 3.3), ils informent principalement sur les faits (cf. *supra*, 3.2.3) ou la localisation (23) :

(21) d'un habitant **de Jezainville** / un habitant **de Folschviller** / au parquet **de Metz** / les enquêteurs de la compagnie **de Toul** et de la section de recherches **de Nancy**

(22) Un automobiliste **de 28 ans** / d'une mère **de 19 ans** / son petit-fils **d'un an** / un Mosellan **de 74 ans**

(23) Cette automobiliste **hollandaise** / Un quadragénaire **uckangeois** / un motard **sarrois** / un retraité **mosellan**

[25] Les relatives sont exploitées dans 13% des cas.

[26] Les pourcentages sont ici calculés sur le total de chacune des catégories.

[27] Avec un « jeu » entre têtes lexicales et modifieurs puisque les noms délivrent, le cas échéant, ces informations : *Une octogénaire périt dans un incendie/Un jeune Maiziérois.*

Les modifieurs sont assez peu utilisés pour « qualifier » physiquement ou moralement les référents. A cet égard, (24) et (25) constituent les deux exceptions de notre corpus. Cette neutralité peut paraître paradoxale dans la mesure où les faits narrés prêtent facilement le flanc à l'émotion/indignation, etc. Mais, d'après la teneur des syntagmes nominaux, il semble bien que les auteurs de faits divers sont, sans doute par devoir et par déontologie, astreints à une forme d'objectivité.

(24) Un détenu réputé **très violent**

(25) Un passant avait été pris à partie et bastonné à coup de chaîne par deux frères **turbulents**, vivant dans une maison de MMH.

3.4.4 Zoom sur ...

Nous proposons une analyse de deux catégories de noms en nombre conséquent dans notre corpus : les noms « relationnels » et les noms généraux d'humains.

3.4.4.1 … les syntagmes nominaux relationnels

Les noms relationnels sont la sous-classe de noms la plus nombreuse de notre corpus (32% des syntagmes nominaux). Ils sont remarquables, lexicalement, par leur absence d'autonomie référentielle, que traduit, entre autres, la difficulté ressentie à leur emploi isolé (cf. Herslund, 1996, entre autres) :

(26) *J'ai rencontré **un frère** (ex. de Herslund, 1996 : 36)

Aux noms dénotant les relations sociales, bien repérés dans la littérature, telles que la parentèle (27) (26% des noms relationnels) ou le voisinage (28), nous ajouterons d'autres noms[28] renvoyant à d'autres formes de relations : spatiales (29) et (30) (14% des noms relationnels) ou celles que nous appellerons, faute de mieux et provisoirement, situationnelles (60%) comme *témoin* ou *victime* dans (31) :

(27) **L'ex-compagnon d'une mère de 19 ans** retrouvée morte devant son domicile dans les Deux-Sèvres est suspecté de l'avoir étranglée.

(28) Il était environ 20h45 quand les secours ont été alertés par **des voisins**.

(29) Rapidement, la fumée envahit la cage d'escalier, empêchant **plusieurs habitants** de s'échapper

(30) Le choc a été tel qu'Ali Simsek, l'un des **deux occupants du véhicule**, a été éjecté.

[28] Encore peu pris en considération dans les classifications nominales et donc encore peu étudiés.

(31) Selon **plusieurs témoins**, **la victime** s'était proposée pour le ramener à son domicile.

Ces dernières sous-catégories de noms sont bien dépendantes comme le montre le test appliqué ci-dessous :

(32) *J'ai rencontré un habitant/un occupant/un témoin/une victime

Toutefois, à la différence des noms de parenté, la dépendance des deux derniers types de noms résulte de leur statut d'« agent » (agresseur) ou d'« acteur » (p.e. *conducteur/habitant*[29]), issu de la « transposition » selon Benveniste d'une construction verbale transitive (*agresser/conduire/habiter*) dont les arguments peuvent figurer dans les CDN (cf. *supra*, ex. (21) et (30)) ou se laisser récupérer *via* le contexte.

Ces types de noms sont suscités par le genre du fait divers, dont les auteurs collectent les informations, pour une grande part, auprès des forces de l'ordre et au moyen d'enquêtes de voisinage, menées auprès des proches (familiaux ou locaux) des protagonistes. Ils montrent également que l'expression référentielle des faits divers est totalement focalisée sur le protagoniste et sa situation dramatique, qui servent de point d'ancrage aux anaphores (cf. *infra*).

Au plan théorique, la présence de ce genre de nom dans les chaînes de référence pose un double problème car elle montre que leur matériau a une origine diverse et que, par voie de conséquence, celui-ci a un empan, ou un potentiel d'utilisation sur la chaîne, variable.

Pour ce qui concerne l'origine, ou la source, on peut distinguer entre :

- celles qui sont alimentées par le cotexte propositionnel, comme, par exemple le syntagme nominal défini *la victime* (*cf.* (20) *supra*) dont l'interprétation dépend du contenu d'une proposition antérieure (*avait succombé suite à l'inhalation de monoxyde de carbone*) et qui, selon la conception que l'on a de la source, peut être considéré comme anaphore associative (ancrée dans la situation et récupérée par les processus inférentiels décrits par Reichler-Béguelin, 1989)[30] ou comme une anaphore prédicative (puisque le SN fournit une information nouvelle). Dans cet

[29] Pour ces noms d'agents, voir (Benveniste, éd. 1974), (Anscombre, 2001), (Luquet, 1994) sur la différence de construction entre les noms d'agent du type d'*enquêteur* (qui vient d'*enquêter*) et *conducteur* (qui provient du supin de *conducere* > *conductus*). Pour les noms en *-ant*, voir (Anscombre, 2003), pour qui il s'agit de noms d'acteurs, dotés d'une moindre agentivité que les noms en *-eur*, et (Ulland, 1993).

[30] (M.-J. Reichler-Béguelin, 1989) a proposé une typologie de ce type d'anaphores.

ordre d'idées, il conviendrait d'ailleurs d'analyser également des syntagmes nominaux (bi-)nominaux appelés « discursifs » par (Bartning, 1996 : 30) :

(33) Petit détail, **l'homme à l'échelle** est un retraité de 76 ans.

(34) Après contrôles, **l'individu au scooter** est entendu à l'hôtel de police.

(35) Cuisinier au restaurant Les Loges, installé à quelques centaines de mètres de l'accident, **le mis en cause** n'aurait jamais dû prendre le volant.

- celles qui proviennent du faisceau de propriétés intrinsèques au référent (p.e. les noms généraux cf. (36) ci-dessous, de phase, etc.) ;

(36) Malheureusement, il est trop tard pour **Pietro Lancellotti**, chauffeur-livreur de la société Autobar qui exploite des distributeurs automatiques de boissons et de sandwiches. **L'homme**, un Colmarien de 44 ans, père de deux enfants, est recroquevillé sur le siège conducteur de son camion frigorifique. Il a la tête à moitié arrachée. La vitre de sa portière a volé en éclats.

La Figure 9 synthétise la double nature des maillons dans les chaînes de référence :

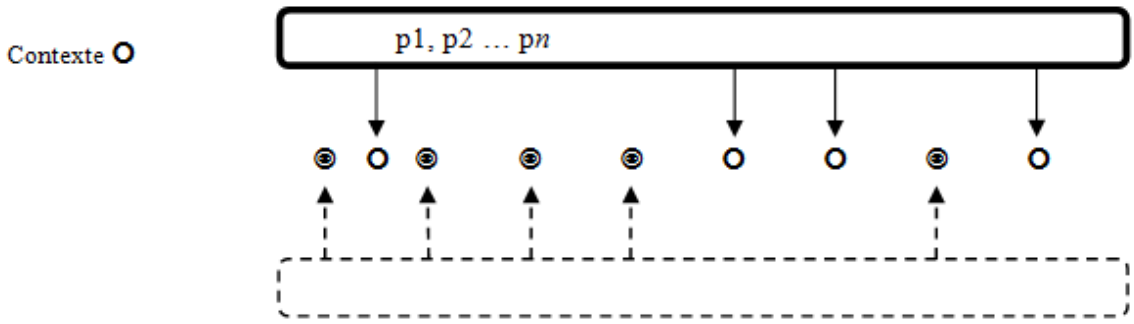


Figure 9 – Sources des maillons dans les chaînes de référence

De là il résulte que les syntagmes nominaux dont l'information dépend des propriétés intrinsèques du référent sont susceptibles d'être employés tout au long d'une chaîne de référence : on dira alors qu'ils ont un empan large, voire maximal. Par contraste, les syntagmes nominaux situationnels ou relationnels, fortement dépendants du cotexte, opèrent une saisie référentielle qui reste ponctuelle, ce qui fait qu'ils ont un empan « étroit ». C'est ce que traduisent notamment les modalités de leur reprise : dans (37), *la victime*, dont le référent est masculin, est repris par un pronom congruent en genre mais sur une brève

distance[31], l'usage consistant apparemment à réinstancier assez rapidement le référent par un syntagme nominal congruent en genre :

(37) **La victime** est tombée sur un chemin bétonné menant au hall d'entrée. **Elle** gisait sur le ventre quand les secours sont arrivés. Entourés de nombreux riverains, les pompiers ont tenté de réanimer **l'octogénaire** sur place avant de **le** transporter dans l'ambulance où **il** est décédé.

(38) Samedi, vers 19h, **la victime âgée de 29 ans et connue pour sa fragilité**, avait pris **son** vélo pour **se** rendre sur un terrain vague de Blainville. **L'homme s'**était aspergé d'essence avant d'y mettre le feu.

De là vient aussi que, certains emplois, comme en (39) où le référent désigné par *la victime* était manifestement connu par ses voisins antérieurement aux faits, doivent être considérés comme ce que (Fauconnier, 1984 : 47 et 49)[32] nomme des connexions entre « espaces-temps » :

(39) La police de Capellen a été rejointe par la police criminelle chargée par le parquet de rechercher des empreintes. La Protex a également envoyé son service d'aide psychologique afin d'assister *l'habitante de la rue qui connaissait* ***la victime*** *ainsi que ses voisins*, tous très choqués par ce geste désespéré.

Bref, il ressort de ces observations que la composition des chaînes de référence est plus complexe qu'il n'y paraît compte tenu de l'origine variable de ses maillons. Cela ne semble guère avoir d'incidence dans une perspective d'annotation de corpus, où l'on « enregistre » les maillons les uns après les autres – encore faudrait-il le vérifier – mais, au plan cognitif, où l'on considère que le processus de traitement se fait de manière incrémentielle, en enrichissant le « fichier » d'un même référent au fur et à mesure du traitement des occurrences, il se pourrait que la disparité des maillons soit de nature à faciliter *vs* entraver le processus interprétatif. Ce qui serait, bien entendu, là encore, à vérifier.

3.4.4.2 … les noms généraux d'humains

La seconde catégorie de noms qui, en nombre, suit les relationnels est constituée par les noms généraux d'humains, du type *homme, personne, individu,* pour lesquels le pourcentage s'élève à 16% des syntagmes nominaux dont la tête est un nom commun. Leur originalité, dans les faits divers, est de faire office de 1er

[31] Lorsque la victime renvoie à un référent féminin, la reprise pronominale peut opérer sur une longue distance. Quand le référent est masculin, il est assez rapidement réinstancié sous la forme d'un syntagme nominal masculin.

[32] Qui donne d'ailleurs en guise d'illustration des extraits de faits divers.

maillon dans 21% des cas[33], titraille exclue (cf. (41) et (42)), et, dans 40% des cas, d'anaphore. Ce point est important dans la mesure où, dans d'autres genres textuels, les syntagmes nominaux à hyperonymes ont plutôt tendance à faire office de reprises anaphoriques comme l'illustrent ((43)-(44)) :

(40) Le 10 juillet dernier, **un homme** est mort après avoir été éjecté d'une voiture incontrôlable

(41) **Un homme** a été tué de plusieurs balles, dans la nuit de lundi à mardi.

(42) Quatre **personnes** sont décédées dans la nuit de samedi à dimanche dans un accident de la circulation dans le Cantal

(43) Le charme de **M. Giscard d'Estaing** opère. **Il** garde de la distance mais **ø** sait séduire, avec humour**. L'homme** est aussi craint et **il** impose un secret total (*Le monde,* 07/06/2003)

(44) **Mon père** voulait devenir ingénieur comme lui. Qu'ai-je reçu en héritage de **cet homme** au destin tragiquement suspendu ? (Frantext, Flem, *Lettres d'amour*, 2006)

Ce mode de présentation initial, qui s'appuie sur les expressions référentielles spécialisées dans la référence indéfinie épistémique[34] selon (Martin, 2006 : 17), a pour effet d'anonymiser les référents et de contribuer à ce que les spécialistes du fait divers appellent la « platitude » des personnages (Dubied, *op. cit.* : 146) *i.e.* à l'absence d'attributs contradictoires et fixes.

3.5 Bilan

De cette étude ressortent au moins trois résultats. Premièrement, nous avons fait valoir, sur la base d'une étude quantitative et qualitative des expressions référentielles, que les chaînes de référence des faits divers sont originales au triple plan de :

- leur nombre : les faits divers sont, dans leur grande majorité (60% des cas), pluri-référentiels et instancient, qui plus est, des référents « obligés » comme les forces de l'ordre, de la sécurité et la magistrature ; corollairement, la densité des expressions référentielles dans les textes est assez forte ;

[33] Pourcentage calculé sur l'ensemble des noms d'humains généraux du corpus (78).

[34] « Dans l'indéfinition épistémique (dans l'indétermination), l'objet n'est pas déjà déterminé au moment de l'énonciation : soit que le locuteur estime qu'il ne l'est pas pour l'interlocuteur ; soit il est incapable lui-même de spécifier l'objet dont il s'agit parmi tous les objets de même nature […]. » (Martin, 2006 : 19).

- la dimension : les chaînes de référence sont, dans leur ensemble, brèves et comptent en moyenne 3,42 maillons ; cela tient évidemment à la brièveté même du genre ainsi que, dans une moindre mesure, à la pluri-référentialité ;
- la composition : les maillons des chaînes de référence sont hétérogènes du point de vue de leur catégorie grammaticale qui se manifeste dans des proportions inattendues par rapport aux prédictions de certaines théories (par exemple l'accessibilité) puisque les formes de moyenne et haute accessibilité prévalent ; pour ce qui concerne les syntagmes nominaux, leur tête lexicale est éminemment variable ; en outre, il est apparu que l'hétérogénéité des chaînes de référence était renforcée, pour ainsi dire, par l'origine des maillons selon qu'ils s'ancrent dans la situation *vs* dans le faisceau de propriétés du référent : cela a des incidences sur leur empan long *vs* étroit selon leur nature ;

En cela, nos observations donnent des assises linguistiques non négligeables à celles des théoriciens du fait divers, qui, même lorsqu'ils examinent précisément, ainsi que le fait (Dubied, *op. cit.* : 227-245), le personnel du roman, en restent à des considérations sémiotiques très générales. Nos observations montrent également que les chaînes de référence des faits divers se différencient aussi bien de genres non journalistiques (comme les textes de lois) et de sous-genres journalistiques (comme le portrait). La corrélation entre genre et chaînes de référence se confirme donc, ce qui fait des chaînes de référence un critère fort pour distinguer les genres et devrait inciter les approches des genres fondées sur l'exploration de gros corpus, à l'instar de celle de (Biber, 1989, 1994, 1995), à systématiser la prise en compte de ce paramètre.

Au plan théorique, la distinction entre anaphores « situationnelles » *vs* anaphores fondées sur les propriétés du référent montre, par ailleurs, que la composition des chaînes de référence, généralement schématisée de manière linéaire et « plate » par une série de maillons, n'est que le résultat d'expressions référentielles dont le point d'ancrage est éminemment variable avec des incidences sur le traitement cognitif qui mériteraient d'être explorées.

Enfin, à travers la typologie, encore sommaire, des noms d'humains qui constituent, dans les faits divers, l'essentiel des expressions référentielles, nous avons montré que la distinction entre anaphores situationnelles vs fondées sur les propriétés du référent, nécessite une étude fine du lexique dénotant l'humain.

4 Conclusion

Le type de chaîne de référence est-il une caractéristique du genre textuel. Les études que nous avons effectuées nous permettent de dégager des tendances (Adam, 2004), des traits caractéristiques des chaînes de référence dans les genres étudiés (rapports publics, faits divers, éditoriaux, etc.). Ces tendances vont permettre à notre système de détection automatique de chaînes de référence *RefGen* (*cf.* Partie III) de « s'attendre à », de rechercher en priorité certains éléments (catégorie du premier maillon d'une chaîne de référence, distance entre les maillons, longueur des chaînes de référence) plutôt que d'autres.

Néanmoins, (Tutin *et al.*, 2000 : 24) reconnaissent que, même si leur étude a porté sur des textes appartenant *a priori* au même genre (*i.e.* des textes informatifs), il demeure des différences de paramètres (nature des reprises, distance entre expressions référentielles) qui posent problème pour mettre en place des stratégies de résolution d'anaphores automatiques globales (*i.e.* pour tous les genres confondus). Remarquons que, dans leur étude, les catégories de genres utilisées de même que les étiquettes sont plutôt à « gros grains », ce qui n'est pas le cas dans nos deux études. En effet, nous avons travaillé sur des textes de genres journalistiques (articles, faits divers, éditoriaux), un roman, des articles de lois, des rapports publics.

Cependant, les résultats que nous avons obtenus sont tout de même à nuancer du point de vue de la taille des corpus d'étude, du choix des genres et du caractère multi-séquentiel des genres (Adam, 2004, 2011). En effet, nous avons travaillé en majeure partie sur des corpus de taille réduite, vu la difficulté pour annoter les chaînes de référence. Aussi, le choix d'utiliser un roman libre de droits (les *Trois Mousquetaires*) a pour conséquence une différence notable d'époque par rapport aux autres genres étudiés. Enfin, il est important de souligner qu'au sein même d'un genre, par exemple, le roman, il est courant de rencontrer différents types de séquences : des descriptions, des narrations, des argumentations, etc. (Adam, 2004), ce qui rend d'autant plus difficile l'essai de caractérisation des chaînes de référence.

Couplées aux autres indices statistiques et linguistiques, les spécificités des chaînes de référence suivant le genre dégagées à l'issue de ces études en corpus vont permettre à notre système de détecter automatiquement les thèmes des documents. Dans la partie suivante, nous présentons un état des lieux des systèmes automatiques de détection des thèmes existants, en nous focalisant sur

les méthodes statistiques et linguistiques de segmentation thématique (chapitre 4) et sur les systèmes de résolution de la coréférence (chapitre 5).

PARTIE II
Aspects automatiques : systèmes de détection de thèmes et de coréférence

Si, jusqu'à présent, la notion de *thème* ne fait toujours pas l'unanimité dans les études linguistiques, le point de vue adopté du côté du traitement automatique des langues (TAL) demeure plutôt consensuel, bien qu'assez réducteur. En effet, pour la plupart des systèmes de détection automatique de thèmes, la notion de *thème* se limite souvent à considérer le thème comme *ce sur quoi porte la phrase ou le texte*. Ainsi, dans un système de TAL, les définitions et les modèles linguistiques décrivant les propriétés des thèmes ne trouvent qu'une applicabilité limitée. Cette « simplification » de la notion de *thème* est liée à la relative complexité de la tâche de détection automatique de thèmes : « Extracting themes from a dataset is the most challenging task in analyzing qualitative data » (Davi *et al.*, 2005 : 89). De ce fait, suivant les systèmes, les thèmes peuvent être associés à des événements[1] (Allan, 2002), être dépendants du domaine (Ferret *et al.*, 2001) ou bien représenter un vecteur[2] de mots avec leurs probabilités (Hearst, 1997). Les thèmes des documents peuvent alors être considérés comme des thèmes composites (*i.e.* le thème du discours et son contexte) ayant une structure hiérarchique (Bilhaut et Enjalbert 2005 ; Bilhaut, 2006, 2007) ou bien comme des agrégats de thèmes phrastiques (Goutsos, 1997). Mais, les thèmes des documents sont liés les uns aux autres par des connecteurs, par des procédés de répétitions,

[1] « a topic is defined to be a set of news stories that are strongly related by some seminal real-word event » (Allan, 2002 : 2). Par exemple, les articles traitant de la première greffe de larynx artificiel menée le 07 octobre 2013 constituent un thème, de même que ceux portant sur la méthodologie adoptée, les matériaux utilisés pour l'opération, etc. En revanche, des articles portant sur une greffe de foie effectuée le même jour n'appartiennent pas au même thème.

[2] La représentation vectorielle permet d'associer à chaque mot un ensemble de valeurs reflétant sa similarité avec les autres mots du document. Cela permet de représenter l'importance des mots dans le document.

par des marqueurs référentiels (marqueurs lexicaux cadratifs, anaphores et chaînes de référence) explicites. Ce sont ces marques de cohésion qui participent à la détection des thèmes des documents et que nous proposons d'identifier automatiquement.

Dans cette seconde partie, nous présenterons les principales méthodes et systèmes (statistiques et linguistiques) de détection automatique de thèmes (chapitre 4). Le chapitre 2 nous a permis de montrer l'importance de la contribution des chaînes de référence à la détection des thèmes des documents (*i.e.* ce sont des marques d'introduction, de maintien et de rupture thématique). Néanmoins, comme nous le verrons dans le chapitre 5, les systèmes actuels d'identification automatique de la (co)référence, qu'ils fonctionnent par apprentissage ou par règles, ne prennent que partiellement en compte ces éléments.

Chapitre 4

Systèmes automatiques pour la détection de thèmes

Sous la notion de *détection automatique de thèmes* sont regroupées plusieurs problématiques distinctes mais étroitement liées : la caractérisation des thèmes abordés dans un document, la segmentation de ce document en segments thématiquement homogènes et la détection des thèmes proprement dite (Rossignol, 2005).

La *caractérisation des thèmes* suppose de réaliser en amont une liste ainsi qu'une description des thèmes potentiellement abordés dans un texte. Cette description sous-entend une connaissance préalable du contenu des documents et oblige souvent à faire un « choix » parmi des thèmes prédéfinis et donc forcément généraux[1]. En cela, la caractérisation des thèmes s'écarte de notre perspective automatique qui a pour objectif de détecter les thèmes contenus dans le document lui-même, sans connaissance préalable, afin de caractériser au plus près les différents thèmes abordés dans le document. Le second point, la *segmentation thématique*, comprend l'identification des lieux de changement de thème dans le document (Ferret, 2006a), d'une part, et le découpage en segments thématiquement homogènes, d'autre part. La segmentation thématique permet ainsi de délimiter les différentes zones thématiques du document. Les systèmes de segmentation thématique insèrent des marques de début et de fin de segment dans le texte ; une marque de fin étant toujours suivie par un début de segment, sauf pour le dernier segment (Hernandez et Grau, 2002). Le dernier aspect, la *détection des thèmes* à proprement parler, est la tâche liant les thèmes aux segments thématiques déterminés à l'issue des deux premières opérations. Dans notre approche, cette tâche consiste à associer à chaque segment thématique le ou les thèmes identifiés automatiquement (*via* les marqueurs linguistiques).

La plupart des travaux en TAL se focalisent essentiellement sur la segmentation thématique. La segmentation thématique peut être considérée comme une tâche à part entière, mais elle peut aussi constituer une première étape permettant d'améliorer les performances d'un système de recherche d'information (*e.g.* pour le résumé automatique, l'indexation, l'extraction d'information (Boufaden *et al.*, 2002, 2010), la détection de thèmes dans des émissions télévisuelles (Claveau et

[1] Suivant le système de détection de thèmes mis en place, un document peut n'être décrit que par un seul thème général (*e.g.* santé, histoire, littérature, économie). Cela est le cas des modèles probabilistes tels que WSIM (Brun *et al.*, 2002) qui réalise, en amont, l'apprentissage des mots caractéristiques d'un thème pour associer ensuite un document à un de ces thèmes prédéfinis.

Lefèvre, 2011a ; Guinaudeau, 2011 ; Bouchekif *et al.*, 2013), la navigation textuelle (Ellouze, 2010), la structuration thématique des textes (Hernandez, 2004), l'amélioration d'un système de question-réponse (Prince et Labadié, 2007 ; Foucault *et al.*, 2013), la détection de l'organisation textuelle *via* une base de voisins distributionnels (Adam, 2012), etc.). Dans notre approche, c'est dans la perspective d'amélioration d'un moteur de recherche global que nous utilisons la segmentation thématique. En effet, la segmentation d'un texte en unités plus petites en améliore l'indexation (Callan *et al.*, 1992 ; Hearst et Plaunt, 1993) et la recherche thématique (Nomoto et Matsumoto, 1996). De plus, la délimitation en segments permet de se positionner directement dans une partie du document, afin de faciliter la navigation. Notons que nous nous intéressons à la segmentation intratextuelle, c'est-à-dire que nous cherchons à segmenter en thèmes un document, pas à retrouver les frontières entre des extraits textuels provenant de documents différents[2], comme il est fréquent de le constater dans les campagnes d'évaluation (Labadié et Prince, 2008a).

Partant, différentes méthodes d'analyse thématique automatique permettent de définir la structure thématique d'un document. On identifie généralement deux « familles » de méthodes (Bilhaut, 2006) :

- d'une part, les méthodes *statistiques* (ou quantitatives) qui se fondent sur la notion de cohésion lexicale (Halliday et Hasan, 1976). Ces méthodes de segmentation linéaire du texte utilisent, pour la majorité, la répétition des mots comme indice ;
- d'autre part, les méthodes *linguistiques* qui exploitent divers marqueurs de cohésion (cadres de discours, anaphores, connecteurs), indices et formes typodispositionnelles (*e.g.* marques de paragraphes, tirets) pour signaler la structure thématique (Hearst et Plaunt, 1993 ; Ferret *et al.*, 2001 ; Ferret, 2009).

Dans ce chapitre, nous présentons des travaux issus de chacune de ces deux approches (sections 1 et 2). Plutôt complémentaires que concurrentes (Ferret, 2006a), elles ont donné lieu à des méthodes hybrides statistiques-linguistiques que nous exposerons dans une troisième section. Loin de viser une présentation exhaustive de ces divers travaux, notre objectif est de positionner notre méthode dans ce vaste domaine.

[2] En effet, dans les campagnes d'évaluation telles que DEFT2006 (DEfi Fouille de Texte, https://www.lri.fr/~aze/fdt/DEFT06/), le corpus d'évaluation proposé est composé de fragments textuels issus de plusieurs documents concaténés. L'objectif du défi revient à retrouver les frontières entre ces divers extraits.

1 Systèmes statistiques de segmentation thématique

Dans le chapitre 1, nous avons vu que les marques de cohésion textuelle telles que les chaînes référentielles constituaient des guides pour l'interprétation et pour la mise en évidence d'éléments centraux dans le texte.

La segmentation thématique vise à découper les textes suivant les thèmes qui y sont abordés. Les techniques employées pour la segmentation automatique[3] sont, pour la plupart, fondées sur la notion de *cohésion lexicale* (Halliday et Hasan, 1976). L'hypothèse sous-jacente est qu'une portion textuelle présentant une forte cohésion lexicale (marquée notamment par la répétition lexicale) possède une grande probabilité de représenter un segment thématique.

Parmi les nombreux algorithmes développés pour la segmentation thématique, on dégage habituellement deux familles de méthodes (Hernandez, 2004 ; Adam et Morlane-Hondère, 2009 ; Adam, 2012) :

- les méthodes *ascendantes* qui parcourent le texte de manière linéaire suivant une fenêtre d'observation glissante (voir section 1),
- les méthodes *descendantes*, basées sur une matrice de similarité entre segments thématiques d'abord calculée au niveau global (pour l'ensemble des unités du texte), avant de délimiter les points de rupture (voir section 2).

Des méthodes statistiques *hybrides*, issues de ces deux familles de méthodes, ont aussi été mises en place (voir section 3). Nous présentons ci-après quelques méthodes issues de chaque famille ainsi que des méthodes statistiques hybrides.

1.1 Méthodes ascendantes

Les méthodes ascendantes, dans la lignée de (Youmans, 1991), (Hearst, 1997), (Reynar, 1994) et (Salton *et al.*, 1996), procèdent à une segmentation linéaire du texte suivant des critères quantitatifs. Le principe est de rechercher les ruptures

[3] Nous considérons dans ce chapitre uniquement les systèmes ne nécessitant aucun apprentissage en amont (les systèmes par apprentissage utilisent un corpus segmenté manuellement comme modèle d'apprentissage et définissent *a priori* les thèmes recherchés). De même que (Rossignol et Sébillot, 2003), nous souhaitons faire émerger les thèmes des documents eux-mêmes.

thématiques et de les identifier lorsqu'un segment du document présente un moins grand nombre de mots traitant du thème. Ainsi, si un terme possède une fréquence élevée dans l'ensemble du document, il ne permet pas de déterminer un thème particulier du texte, alors que sa répétition dans une zone textuelle limitée permet de caractériser le thème du segment.

1.1.1 La méthode par blocs thématiques de (Salton *et al.*, 1996)

(Salton *et al.*, 1996) figurent parmi les premiers à avoir eu l'idée de segmenter le texte en blocs thématiques dans une perspective de recherche d'information. Les frontières entre blocs signalent alors les lieux de changement de thème.

Dans leur approche, le texte est d'abord découpé en paragraphes (suivant le découpage paragraphique existant). Puis, des calculs de proximité lexicale entre paragraphes adjacents permettent de fusionner les paragraphes. Ces calculs s'effectuent suivant des mesures de distances vectorielles, chaque bloc étant représenté par un vecteur lexical (*i.e.* une représentation des occurrences des termes du bloc). La fusion se déroule de manière itérative, en parcourant le texte du début à la fin et de gauche à droite, jusqu'à ce que celui-ci ne contienne plus que des blocs thématiquement homogènes. A l'issue de cette segmentation en blocs thématiques, une représentation en graphe permet de visualiser les relations thématiques établies entre les blocs dans le document. Par exemple, dans la Figure 10, le graphe représente les 28 relations établies entre les 16 paragraphes (p1 à p16) du texte 21385 traitant du tabac : son histoire, ses effets sur la santé, son arrêt.

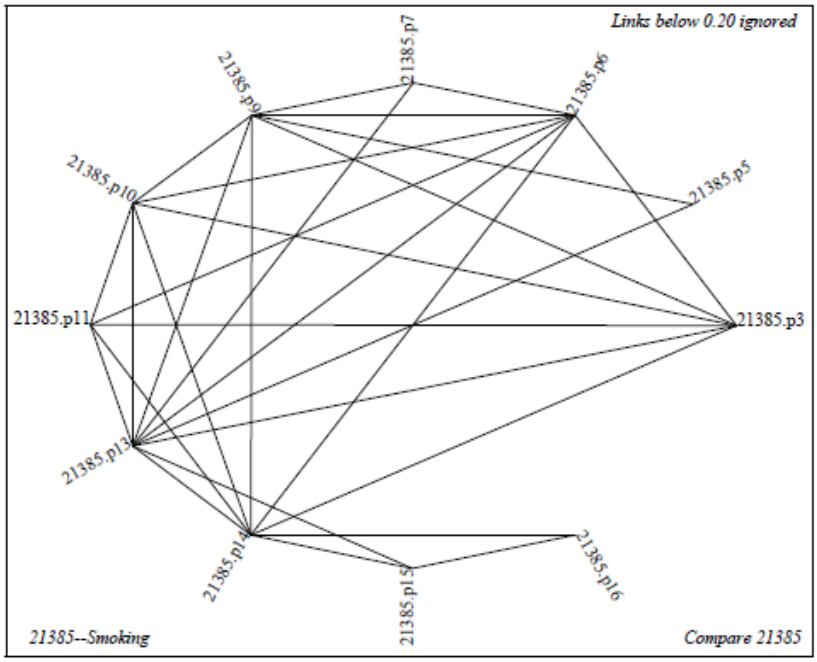


Figure 10 – Représentation en graphe de blocs thématiques, d'après (Salton *et al.*, 1996 : 54)

D'après les auteurs, les relations existant entre des paragraphes contigus d'un texte créent des « segments de texte » représentant des unités fonctionnelles (par exemple, introduction, conclusion). Des relations existant entre des paragraphes éloignés les uns des autres créent des « thèmes de texte » (*i.e.* des liens thématiques). Ainsi, le type de structure de texte est-il caractérisé par la manière dont interagissent ces deux plans de relations :

- si le découpage en segments correspond au découpage en thèmes, le texte est homogène (*i.e.* les thèmes sont traités les uns après les autres de manière indépendante),
- si le découpage en segments ne suit pas le découpage en thèmes, soit le texte comporte des thèmes centraux, soit il comporte de nombreux sous-thèmes entrecroisés.

Ces informations de structures permettent alors de sélectionner les passages les plus pertinents d'un article suite à une requête utilisateur. Néanmoins, la caractérisation des thèmes abordés dans les documents n'est pas traitée, ce qui limite l'utilisation de cette méthode pour nos travaux.

1.1.2 La méthode du *TextTiling* de (Hearst, 1997)

Dans la lignée de (Salton *et al.*, 1996), la méthode du *TextTiling* de (Hearst, 1994, 1997) étudie la distribution des termes d'un document et leur récurrence

pour en identifier les frontières thématiques. *TextTiling* a été développé pour traiter des textes expositifs contenant de nombreux paragraphes et une structure du discours plutôt marquée par le contenu thématique : « in expository text [...] the subject matter tends to structure the discourse more so than characters, setting, and so on » (Hearst, 1997 : 40).

La première étape de la méthode consiste à segmenter le document en blocs de trois à cinq phrases ou pseudo-phrases (voir Figure 11). Les blocs ainsi établis sont comparés deux à deux (étape 2) et une valeur de similarité entre blocs adjacents est calculée par une mesure du cosinus (comprise entre 0 et 1). La mesure du cosinus permet de calculer le cosinus de l'angle formé par les vecteurs des mots contenus dans deux documents dérivés à partir de la fréquence des termes contenus dans chaque document. Elle est donnée par l'équation suivante. Etant donné des blocs *b1* et *b2*, alors :

$$\cos(b1, b2) = \frac{\sum_{t=1}^{n} w_{t,b1} w_{t,b2}}{\sqrt{\sum_{t=1}^{n} w^2_{t,b1} \sum_{t=1}^{n} w^2_{t,b2}}}$$

où t s'étend à l'ensemble des termes du document, $wt,b1$ est le poids tf.idf assigné au terme t dans le bloc *b1* (le poids tf.idf correspond au nombre de termes communs et au nombre de fois qu'ils apparaissent dans l'ensemble du document).

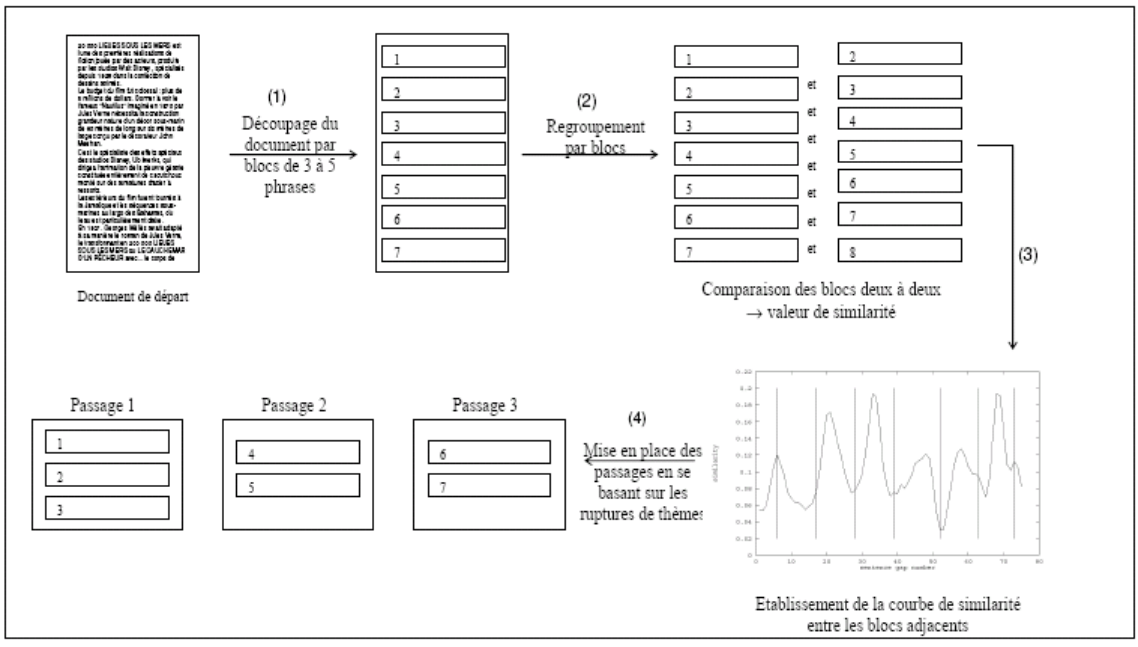


Figure 11 - Méthode du *TextTiling* (issu de (Maisonnasse et Tambellini, 2005 : 6))

Un score de cohésion lexicale est ensuite attribué à chaque bloc du texte en fonction du bloc qui le suit afin d'établir une courbe de similarité (étape 3). L'algorithme procède enfin au marquage des frontières thématiques suivant l'écart

calculé entre les blocs adjacents *via* l'utilisation d'une fenêtre glissante. Les frontières de blocs présentant les écarts les plus importants (selon une valeur seuil) sont sélectionnées comme frontières thématiques (étape 4).

TextTiling a été évalué sur 12 articles de magazine (entre 1800 et 2500 mots chacun) en comparant la segmentation thématique automatique obtenue avec celle de 7 juges humains. Le système obtient des performances de 63,4 %. L'auteur note une relative absence d'accord inter-annotateurs pour certaines frontières thématiques, ce qui peut constituer une difficulté pour l'évaluation de l'outil.

L'algorithme de (Hearst, 1994, 1997) utilise donc les relations de cohésion lexicale locale pour segmenter les textes en segments de plusieurs phrases ou paragraphes, qui reflète la structure du texte en sous-thèmes (ou passages). Dans cette méthode (de même que celle de (Salton *et al.*, 1996)), la rupture thématique ne pourra être déterminée qu'entre deux blocs, bien qu'elle puisse avoir lieu au milieu d'un bloc. Par conséquent, le choix de la taille du bloc constituerait un premier problème pour le bon fonctionnement du système (Nomoto et Matsumoto, 1996). Aussi, cette méthode effectue une analyse à gros grain de la structure du texte ; elle est donc nécessairement moins précise que des méthodes à granularité plus fine. Néanmoins ce choix de granularité est justifié par l'objectif ultime de cette méthode qui n'est pas seulement d'identifier les frontières thématiques mais aussi d'étiqueter leur contenu correctement. (Hearst, 1997 : 33) fournit ainsi l'exemple de découpage et d'attribution de sous-thèmes à partir d'un texte scientifique nommé *Stargazers*. Ce document a pour thème principal l'existence de la vie sur Terre et sur les autres planètes. L'auteur décrit son contenu comme la succession des sous-thèmes suivants (les numéros indiquent le numéro des paragraphes) :

1-3	Intro - the search for life in space
4-5	The moon's chemical composition
6-8	How early proximity of the moon shaped it
9-12	How the moon helped life evolve on earth
13	Improbability of the earth-moon system
14-16	Binary/trinary star systems make life unlikely
17-18	The low probability of non-binary/trinary systems
19-20	Properties of our sun that facilitate life
21	Summary

(Khalis, 2006) signale une limite à cette méthode concernant le contenu même des documents. En effet, la présence d'un nombre important de titres et/ou de petits paragraphes dans le document génère de mauvais résultats car l'algorithme de

TextTiling ne prend pas en compte l'organisation hiérarchique des textes[4] (Hernandez, 2004 ; Rossignol, 2005 ; Pimm, 2008). De même, lorsque des zones de texte présentent peu de répétitions lexicales, l'algorithme doit faire des choix arbitraires pour segmenter. (Hearst, 1997) précise que dans ce dernier cas l'algorithme doit utiliser des informations additionnelles telles que des marqueurs linguistiques ou des informations supplémentaires sur la cohésion lexicale.

1.1.3 *Segmenter*, méthode par chaînes lexicales (Kan *et al.*, 1998)

De son côté, *Segmenter* (Kan *et al.*, 1998) effectue une segmentation thématique linéaire basée sur les chaînes lexicales[5] présentes dans le texte en attribuant un score à chaque paragraphe en fonction des poids et des types de liens qui y sont contenus.

La méthode de segmentation exécute séquentiellement une série de trois étapes. La première étape repère les catégories d'information qui reflètent le contenu thématique du document (*i.e.* les groupes nominaux, les groupes de noms propres, les pronoms personnels et possessifs). Ensuite, les unités similaires sont regroupées par leur tête lexicale (*e.g.* les occurrences de « mammifères marins » sont incluses dans celles de « mammifères ») et un seuil de filtrage des mots est mis en place. La seconde étape relie les mots apparentés suivant un critère de proximité (on parle de chaînage des répétitions). De ce fait, si deux occurrences d'un terme sont situées dans plusieurs phrases, ces occurrences vont être liées pour constituer une seule unité. Ce procédé est alors répété jusqu'à ce qu'aucun regroupement ne soit plus possible[6]. La différence de taille du lien entre deux occurrences dépend des catégories de termes (*e.g.* la distance du lien pour les noms propres est la plus grande possible alors que celle des formes pronominales est la plus petite possible).

Une fois les liens établis, un poids des chaînes est donné suivant des étiquettes marquant les liens des occurrences des termes dans les paragraphes concernés

[4] Cette limite est applicable à toutes les méthodes de segmentation linéaire.

[5] Comme nous avons pu le voir dans le chapitre 1, une chaîne lexicale est la liaison de toutes les occurrences d'un terme trouvées au fil du texte. Une chaîne lexicale peut être considérée comme rompue par l'algorithme suivant la distance choisie (en nombre de phrases, de mots ou de propositions entre deux occurrences) par l'utilisateur. Lorsqu'une phrase correspond au point de départ d'un grand nombre de chaînes lexicales (nombre à fixer par l'utilisateur), elle constitue une frontière thématique.

[6] Cette idée est une interprétation de la notion de chaîne lexicale présentée dans (Morris et Hirst, 1991).

(voir Tableau 16). Ces étiquettes indiquent le début d'un lien (*Front*), une occurrence qui n'est pas le début du lien (*During*), la fin d'un lien dans le paragraphe précédent (*Rear*) ou bien une activité nulle (*No link*). Pour chaque type d'étiquette et de catégorie de terme, un score de segmentation est attribué. Les paragraphes dont les liens persistent (*During*) se voient attribuer une valeur négative, signe de maintien du thème. Les paragraphes étiquetés *Rear* ou *Front* ont un score positif, ce qui suppose un changement de thème.

Term	**Paragraph Type with respect to term**				**Link**
Type	front	rear	during	No link	**Length**
Proper NP	10	8	-3	*	8
Common NP	10	8	-3	*	4
Pronouns & Possessives	1	13	-1	*	0

Tableau 16 - Poids des chaînes selon l'étiquette du paragraphe et la catégorie d'unité selon (Kan *et al.*, 1998 : 199)[7]

Ce processus de pondération est appliqué à chaque terme, puis les poids attribués aux éléments des paragraphes sont totalisés. Les marques de segmentation sont alors placées au début des paragraphes ayant les scores les plus élevés (troisième étape)[8]. La mise en place d'un seuil pour délimiter les segments est effectuée par l'intermédiaire d'une 0-somme des poids pour chaque unité (*i.e.* un contrepoids égal au poids attribué aux segments est attribué aux paragraphes n'ayant aucune activité). Cela garantit que la somme des poids de tous les paragraphes est égale à 0, donc que cela correspond bien au texte entier. Ainsi, pour un paragraphe donné, les résultats de pondération de la 0-somme sont-ils soit positifs et indiquent alors le début d'un segment thématique, soit négatifs (*i.e.* continuité du thème).

(Kan *et al.*, 1998) ont évalué leur système à partir d'un corpus de vingt articles courts (entre 800 et 1500 mots) issus du *Wall Street Journal* et de *The Economist*. Ils notent que les performances de leur système sont en moyenne 10% supérieures par rapport à celles obtenues par *TextTiling* (Hearst, 1997).

Segmenter offre un traitement spécifique aux catégories grammaticales porteuses de thèmes et met en évidence les liens entre phrases. Néanmoins, le chaînage des répétitions (*i.e.* le regroupement des groupes nominaux similaires sous la

[7] Les valeurs attribuées sont obtenues par apprentissage.

[8] Il est important de noter que dans cette méthode, les formes pronominales se voient appliquer une légère modification dans la pondération des paragraphes. En effet, vu que la majorité des référents des pronoms se situe avant le pronom, les paragraphes étiquetés *Front* (score de 1) ne sont pas pondérés fortement mais l'accent est plutôt mis sur les paragraphes *Rear* (score de 13).

même tête) apparaît fastidieux si l'on traite un corpus comptant de nombreux documents car cette technique nécessite le parcours de chacun des textes.

1.1.4 Le système *SeLeCT* (Stokes *et al.*, 2004)

De même que (Kan *et al.*, 1998), (Stokes *et al.*, 2004) proposent une approche basée sur les chaînes lexicales pour identifier les changements thématiques entre des transcriptions d'émissions radiodiffusées. Dans leur approche, une rupture thématique correspond à la fin d'un reportage dans une émission (*i.e.* à chaque changement de reportage dans une émission correspond un changement thématique). Dans *SeLeCT*, la notion de *cohésion lexicale* n'est pas uniquement utilisée pour la répétition de termes, mais elle prend aussi en compte des associations entre termes (*e.g.* « bébé royal » ↔ « Kate Middleton ») identifiées *via* des statistiques sur les cooccurrences établies à partir d'une série de reportages issus du corpus TDT1 (Topic Detection and Tracking, (Allan *et al.*, 1998))[9].

Le système de segmentation thématique *SeLeCT* contient trois modules : un *tokeniser*, un *chaîneur* et un *détecteur*. Ainsi, le texte est tout d'abord étiqueté, les groupes nominaux simples et complexes sont identifiés (*i.e.* les descriptions définies, les noms propres complets et leurs *alias*[10]) et localisés dans le texte. Puis, le *chaîneur* recherche les relations entre *tokens* (*i.e.* noms, noms propres) afin de former des chaînes lexicales. Un nouveau *token* est ajouté à une chaîne existante s'il est relié à au moins un autre *élément* de la chaîne. S'il s'agit d'un groupe nominal, des critères supplémentaires de distance et de force de relation (*e.g.* répétition, synonymie, généralisation/spécialisation, etc.) entre les *tokens* sont appliqués afin d'éliminer des chaînes parasites. Le processus continue jusqu'à ce que toutes les chaînes lexicales du texte aient été identifiées. Le *détecteur* parcourt alors les chaînes du texte et délimite les frontières thématiques suivant l'hypothèse qu'une forte concentration de début de chaînes et de fin de chaînes dans le texte signale les frontières d'un reportage dans une émission.

Afin d'éviter les problèmes d'accord entre annotateurs, les auteurs n'ont pas utilisé un corpus annoté manuellement en segments thématiques pour l'évaluation de leur système. Ils ont fait le choix de modifier la tâche d'évaluation qui consiste alors à rechercher les frontières entre des reportages concaténés. Evalué sur 1000 reportages issus de CNN et Reuters, *SeLeCT* obtient des performances plus élevées que *TextTiling*. Néanmoins, les performances obtenues sont différentes

[9] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC98T25

[10] Un *alias* est une variante lexicale d'une même entité (*e.g.* « B. Obama » et « Barac Obama »).

pour les articles de Reuters (68,3%) par rapport aux transcriptions de reportages de CNN (54,6%). Cette différence serait liée au type de corpus (écrit *vs* oral) qui ne véhiculerait pas le sens *via* les mêmes catégories de termes (*i.e.* le texte écrit véhiculerait plus de sens par les noms et les adjectifs alors qu'il s'agirait des verbes et des adverbes pour le texte oral (Blanche-Benveniste, 1990)). De ce fait, n'identifiant que les éléments lexicaux du texte (alors que d'autres marqueurs de clôture seraient à prendre en compte), *SeLeCT* éprouverait des difficultés pour retrouver les frontières thématiques des textes transcrits de l'oral[11].

1.1.5 Bilan

Nous avons vu que les méthodes ascendantes procèdent en une segmentation quantitative plate et linéaire du texte. Vu que ces méthodes n'utilisent pas de ressources spécifiques, elles sont peu coûteuses et elles sont dotées d'une grande robustesse. De plus, elles sont applicables rapidement à grande échelle (vu qu'elles ne nécessitent pas d'adaptation au domaine).

Toutefois, ces méthodes ne s'appuient que sur une information lexicale de base (Labadié, 2008) et souffrent d'un manque de précision pour la détermination des frontières de segments thématiques (Hernandez, 2004) car elles n'effectuent pas une analyse fine (*i.e.* à l'échelle de la phrase ou de la proposition). (Ferret, 2006b, 2007 ; Bilhaut, 2006) soulignent aussi que ces méthodes sont mises en défaut lorsqu'on les confronte à des synonymes. En effet, ne comportant pas ou peu de connaissances externes, ces systèmes éprouvent des difficultés pour reconnaître des mots sémantiquement proches.

Enfin, (Hearst, 1997), de même que (Kan *et al.*, 1998) évaluent leurs systèmes en comparant leurs segmentations automatiques par rapport à des segmentations effectuées par des juges humains. Or, l'accord inter-annotateurs ne dépasse guère les 65% (certains annotateurs privilégient la sur-segmentation tandis que d'autres découpent de plus grands pans textuels), ce qui ne permet pas d'obtenir un bon modèle de comparaison pour une évaluation.

1.2 Méthodes descendantes

A la différence des méthodes ascendantes qui définissent les frontières thématiques d'un document en comparant des blocs situés dans un voisinage

[11] Dans (Stokes, 2003), l'auteur propose des heuristiques permettant de pallier ces difficultés liées au type de corpus.

proche, les méthodes descendantes comparent tous les blocs du texte afin de maximiser la valeur de la cohésion lexicale dans chacun des segments. La plupart de ces méthodes se basent sur une représentation matricielle du document. (Reynar, 1994) figure parmi les premiers à avoir proposé ce mode de représentation pour la segmentation thématique.

1.2.1 *Dotplotting*, méthode de segmentation par représentation graphique (Reynar, 1994, 1998)

L'algorithme de *Dotplotting* (Reynar, 1994, 1998)[12] se base sur une représentation graphique des positions des occurrences des termes du texte à segmenter. La détermination des ruptures thématiques s'effectue soit manuellement par l'examen du graphe, soit automatiquement *via* un algorithme d'optimisation.

L'algorithme s'effectue en 4 étapes (voir Figure 13 page 144). Après un prétraitement (lemmatisation et suppression de mots de classe ouverte (*e.g.* « be », « have ») et le repérage des fins de phrases ou de paragraphes, le texte est parcouru afin d'identifier les répétitions de lemmes contenus dans le texte (étape 1). Lorsqu'un terme apparaît à deux positions du texte x et y, quatre points (x, x), (x, y), (y, x) et (y, y) sont représentés sur un graphe, ce qui permet de visualiser les zones du texte comportant de nombreuses répétitions. Par exemple, pour les phrases : « Le chat dort. Le chat miaule. Ce félidé dort souvent. », les lemmes « le », « chat » et « dormir » sont répétés :

phrases	**Le**	*chat*	dormir.	**Le**	*chat*	miauler.	Ce	félidé	dormir	souvent.
n° mot	1	2	3	4	5	6	7	8	9	10

Les coordonnées des lemmes répétés (Tableau 17) sont alors reportées (*e.g.* dans les phrases de l'exemple ci-dessus, « le » est répété en position 1 et 4. De ce fait, les coordonnées de « le » sont (1,1), (1,4), (4,1) et (4,4)) afin d'obtenir un graphe des répétitions (Figure 12). On observe qu'une diagonale se dessine, représentant un terme vis-à-vis de lui-même.

[12] L'algorithme proposé par (Reynar, 1994) est une adaptation pour la segmentation de la méthode des nuages de points présentée par (Helfman, 1994) pour la recherche d'information.

lemme / points	**le**	*chat*	dormir
x,x	1,1	2,2	3,3
x,y	1,4	2,5	3,9
y,x	4,1	5,2	9,3
y,y	4,4	5,5	9,9

Tableau 17 - Coordonnées des lemmes répétés

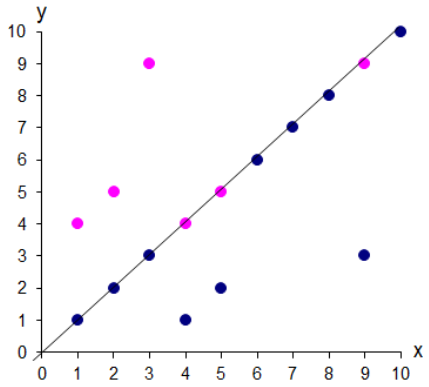


Figure 12 - Tracé des répétitions

En établissant le tracé des répétitions pour chacun des lemmes contenus dans quatre articles du *Wall Street Journal*, (Reynar, 1994) a obtenu un graphe de Dotplot (étape 2)[13]. Sur le graphe de Dotplot, on remarque une forte concentration de points situés de part et d'autre de la diagonale principale, formant des carrés (*i.e.* les 3 zones sombres dans le graphique). Ces zones denses correspondent aux différentes « régions » thématiques des articles. A partir de cette observation visuelle des limites de régions, un algorithme est mis en place, consistant à maximiser les régions de points de forte densité délimitées dans les carrés le long de la diagonale, donc à minimiser de ce fait les régions extérieures de faible densité. La densité est calculée pour chaque unité d'aire en divisant le nombre de points situés hors d'une région par l'aire de cette région (étape 3). Une fois les densités des régions extérieures calculées, l'algorithme sélectionne une limite qui correspond à la densité extérieure la plus faible. Cette limite représente la première rupture thématique. D'autres limites sont ajoutées jusqu'à ce que la densité extérieure croisse ou bien que le nombre de segments soit atteint (étape 4).

Dotplotting a été évalué sur un corpus de 600 articles concaténés aléatoirement (soit 150 concaténations de 2 à 8 articles). L'évaluation a consisté à retrouver les frontières entre articles. Les performances moyennes obtenues oscillent entre 28,25% (reconnaissance stricte des frontières) à 44,6% avec une marge d'erreur de reconnaissance des frontières de 3 phrases.

(Reynar, 1994, 1998) a ainsi proposé une représentation graphique du texte permettant de visualiser rapidement les segments thématiques. L'algorithme permet à l'utilisateur de sélectionner une liste de limites de phrases ou de paragraphes. Néanmoins, chacune des limites est déterminée en fonction de la limite qui précède. De ce fait, la rupture n'est ni déterminée de manière globale,

[13] (Reynar, 1994) a adapté la méthode graphique nommée Dotplotting par (Church, 1993), initialement conçue pour aligner les corpus bilingues.

ni de manière locale (contexte). De plus, (Choi, 2000) montre que les valeurs de similarité des segments provenant de textes courts ne sont pas suffisamment fiables et qu'il faut leur préférer le classement. Il propose alors une amélioration de *Dotplotting* dans son algorithme *C99* (voir 1.2.2).

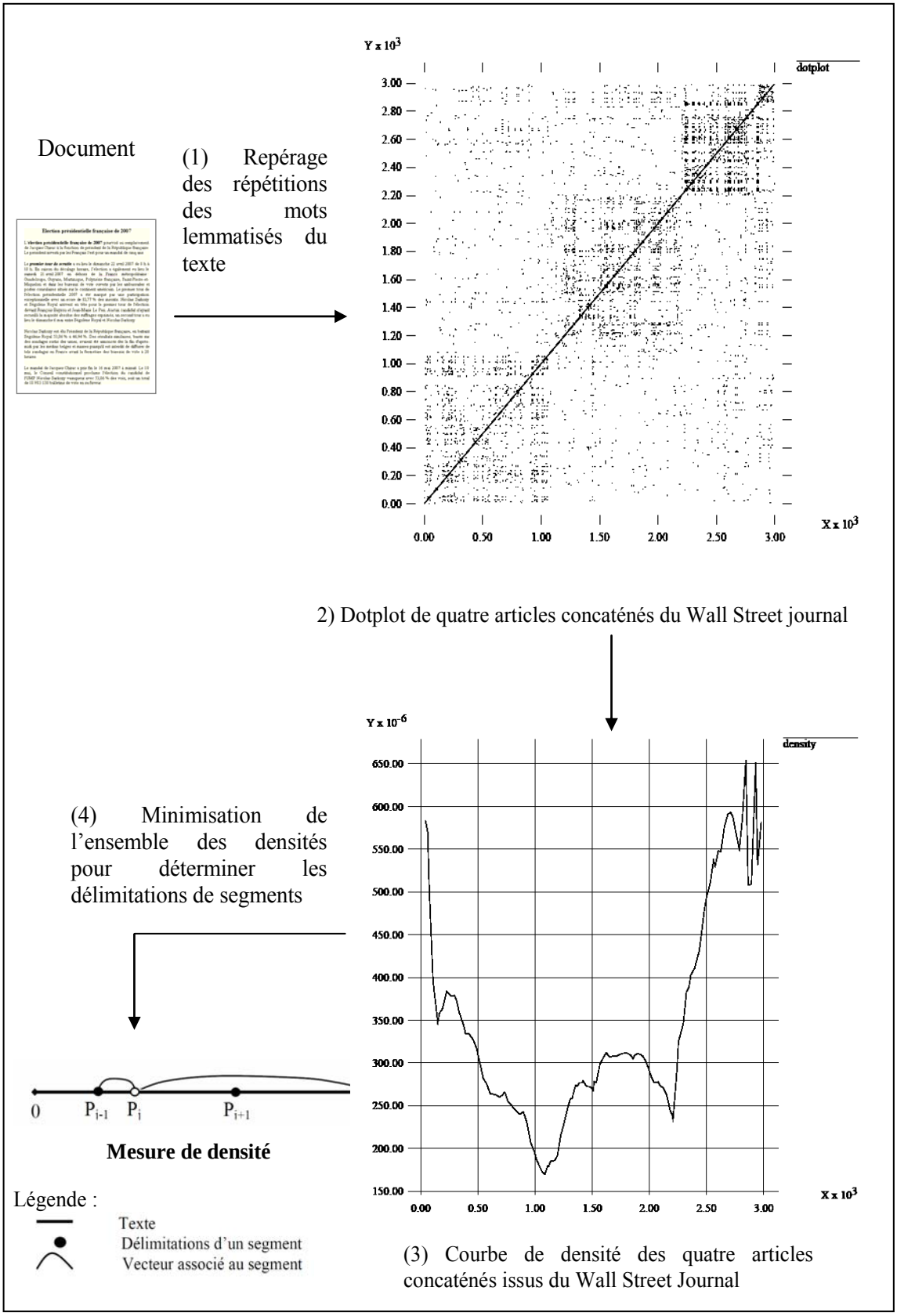


Figure 13 - Principe du *Dotplotting* de (Reynar, 1994, 1998)

1.2.2 *C99*, méthode de segmentation par similarité (Choi, 2000 ; Choi *et al.*, 2001)

1.2.2.1 *C99* (Choi, 2000)

(Choi, 2000) a proposé un algorithme de segmentation thématique à partir de calculs de similarité entre phrases, *C99*[14], basé sur les travaux de (Reynar, 1994, 1998). Partant du principe que les valeurs de similarité de segments issus de textes courts ne sont pas significatives statistiquement, *C99* utilise une représentation vectorielle des phrases et mesure la similarité de l'ensemble des vecteurs qu'il combine à un système de classement.

L'algorithme *C99* divise tout d'abord le document à segmenter en phrases. Les mots peu informatifs (*i.e.* article, pronom, verbes très fréquents, etc.) sont supprimés et les mots restants sont lemmatisés. Puis l'indice de similarité (indice du cosinus[15]) est calculé entre les paires de phrases, contiguës ou non. Cette mesure est appliquée pour toutes les paires de phrases afin de générer la matrice de similarité (voir Figure 14).

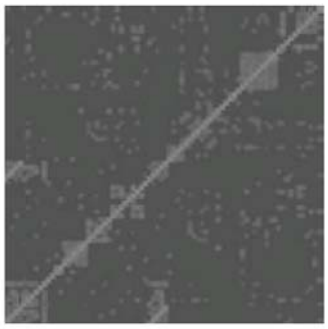


Figure 14 - Exemple de matrice de similarité (Choi, 2000 : 27)

Dans les textes courts, la mesure de similarité n'est pas fiable car elle demeure trop sensible (*e.g.* l'ajout d'une occurrence dans un court segment augmente sensiblement la mesure de similarité). L'originalité de Choi est alors de travailler sur une matrice de rang (et non directement sur la matrice de similarité). Cette matrice de rang (Figure 16) est obtenue en remplaçant chaque valeur de la matrice de similarité par son rang dans le contexte local. Pour calculer ce rang, chaque case de la matrice est comparée aux cases environnantes et obtient un score en fonction du nombre de cases possédant un score de similarité inférieur

[14] L'algorithme est disponible à : http://morphadorner.northwestern.edu/morphadorner/documentation/javadoc/edu/northwestern/at/utils/corpuslinguistics/textsegmenter/C99TextSegmenter.html

[15] *Cf.* 1.1.2.

(Figure 15). Par exemple, dans l'étape 1 de la Figure 15, la valeur 8 est entourée de 7 valeurs inférieures (5, 4, 7, 7, 6, 4 et 1), donc la valeur 8 de la matrice de similarité est remplacée par la valeur 7 dans la matrice de rang.

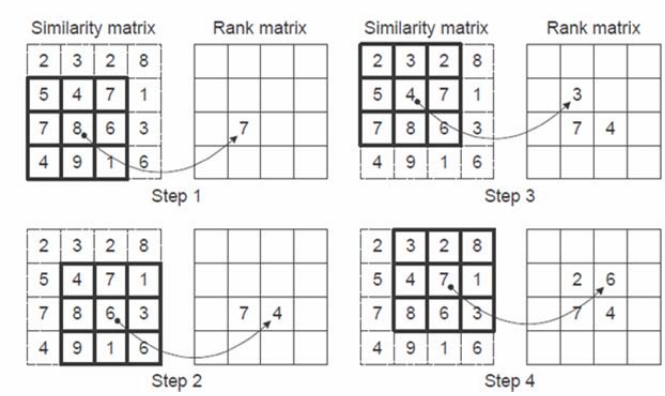


Figure 15 - Calcul de la matrice de rang à partir de la matrice de similarité

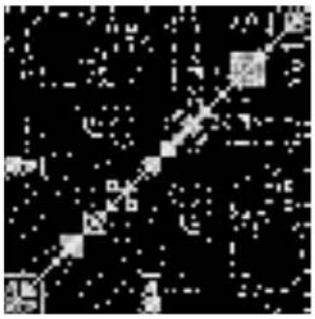


Figure 16 - Matrice de rang

La segmentation en thèmes est effectuée par une procédure d'analyse en grappes (*clustering*) – inspirée de l'algorithme de maximisation de (Reynar, 1994, 1998) – qui segmente récursivement le document selon les frontières entre les unités minimales, maximisant la similarité moyenne à l'intérieur des segments ainsi constitués. L'objectif est de trouver la configuration offrant la plus grande densité en recherchant, à chaque étape, une nouvelle frontière thématique. L'algorithme s'arrête lorsque la densité obtenue est suffisamment faible ou lorsque le nombre de frontières thématiques fixé est atteint (Figure 17).

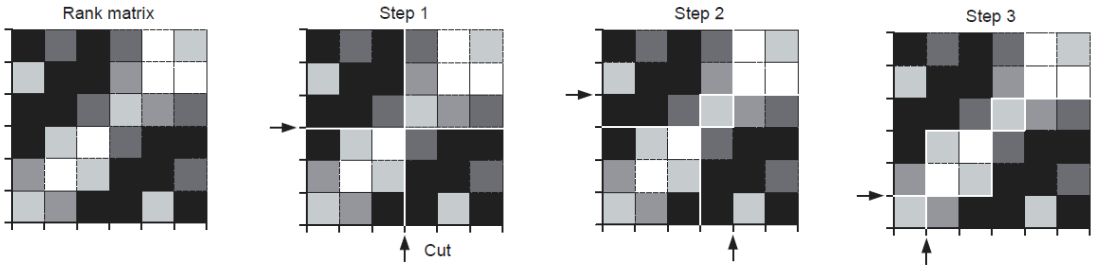


Figure 17 - Exemple de segmentation par regroupement sur la matrice de rang

Afin d'évaluer son système, (Choi, 2000) a créé un corpus artificiel de 700 documents composés d'une sélection aléatoire de phrases issues d'articles du corpus Brown[16]. Ainsi, chaque document contient 10 segments textuels composés des n premières phrases d'un article du corpus Brown sélectionné par une procédure automatique[17]. L'objectif de l'évaluation du système a consisté à

[16] http://nltk.googlecode.com/svn/trunk/nltk_data/packages/corpora/brown.zip

[17] (Choi, 2000) a fait varier n afin de récupérer soit les 3 à 5 premières phrases d'un article, soit les 6 à 8 premières phrases, soit les 9 à 11 premières phrases.

retrouver les frontières entre articles dans chaque document. A l'issue de la comparaison de *C99* avec d'autres algorithmes tels que *DotPlotting*, *Segmenter*, *TextTiling* sur ce corpus d'évaluation, *C99* obtient des performances de 2 à 7 fois supérieures aux autres systèmes. Néanmoins, l'utilisation de la matrice de similarité *via* l'indice du cosinus oblige le système à considérer la répétition stricte de mots dans une paire de phrases. De ce fait, si les deux phrases comparées ne contiennent pas de mots communs mais des synonymes ou des hyperonymes, elles ne seront pas jugées similaires, bien que thématiquement proches. Afin de pallier ce manque, (Choi *et al.*, 2001) ont optimisé l'algorithme *C99*[18] en utilisant l'Analyse Sémantique Latente (Landauer *et al.*, 1998), que nous présentons dans la section suivante[19].

1.2.2.2 L'analyse sémantique latente (LSA)

L'analyse sémantique latente (ASL ou *LSA – Latent Semantic Analysis* – en anglais) est une méthode statistique initialement créée pour la recherche documentaire. Depuis plus de 20 ans, son utilisation s'est largement étendue au domaine de la psychologie (Deerwester *et al.*, 1990 ; Kintsch, 2002), de la psycholinguistique (Landauer *et al.*, 1998), des sciences de l'éducation et du TAL (Bestgen, 2004, 2005 ; Labadié, 2008).

La LSA permet d'évaluer la proximité sémantique entre des paires de mots suivant leurs cooccurrences dans des textes, des paragraphes ou même des phrases (Piérard et Bestgen, 2005a ; Adam *et al.*, 2009 ; Bestgen, 2012) (voir Figure 18). Le point de départ d'une LSA est un tableau lexical comportant le nombre d'occurrences de chaque mot contenu dans un document[20] (étape 1). Les fréquences des termes sont pondérées suivant l'importance des termes dans le document (*i.e.* les mots apparaissant de manière constante dans tous les documents sont éliminés, car non informatifs) (étape 2). Partant du constat que, même dans de grands corpus, les cooccurrences demeurent rares, l'analyse des cooccurrences brutes constitue un problème. Pour pallier les variations aléatoires (Burgess *et al.*, 1998) résultant de la rareté des cooccurrences, la LSA décompose le tableau des fréquences des mots en valeurs singulières[21] (SVD, *Singular Value*

[18] La version améliorée de *C99* (rebaptisée *CWM*) utilisant l'analyse sémantique latente (version 1.3) est disponible à : http://www.freddychoi.co.uk.

[19] *C99* a fait l'objet de plusieurs améliorations en ce sens, notamment par l'ajout d'informations sémantiques *via* des Topic Models (*i.e.* des modèles probabilistes) pour (Riedl *et al.*, 2012 ; Du *et al.*, 2013) ou par l'utilisation des connecteurs pour (Eisenstein et Barzilay, 2008).

[20] Ici, un document peut être un texte, un paragraphe ou une phrase.

[21] La décomposition d'une matrice en valeurs singulières est une analyse factorielle où la matrice de base, composée des documents et des fréquences de mots, est décomposée en trois matrices : deux matrices orthonormales (représentant les lignes et les colonnes de la matrice de base) et une matrice diagonale. Le produit des trois matrices doit permettre de recomposer la matrice de base.

Decomposition) (étape 3) avant de le recomposer à partir d'une partie de l'information qu'il contient. Les mots caractérisant les documents laissent place à des combinaisons linéaires ou « dimensions sémantiques » sur lesquelles sont situés les mots originaux. Les dimensions sémantiques extraites sont de l'ordre de 300 (afin de faire émerger les liens entre mots) et sont non interprétables.

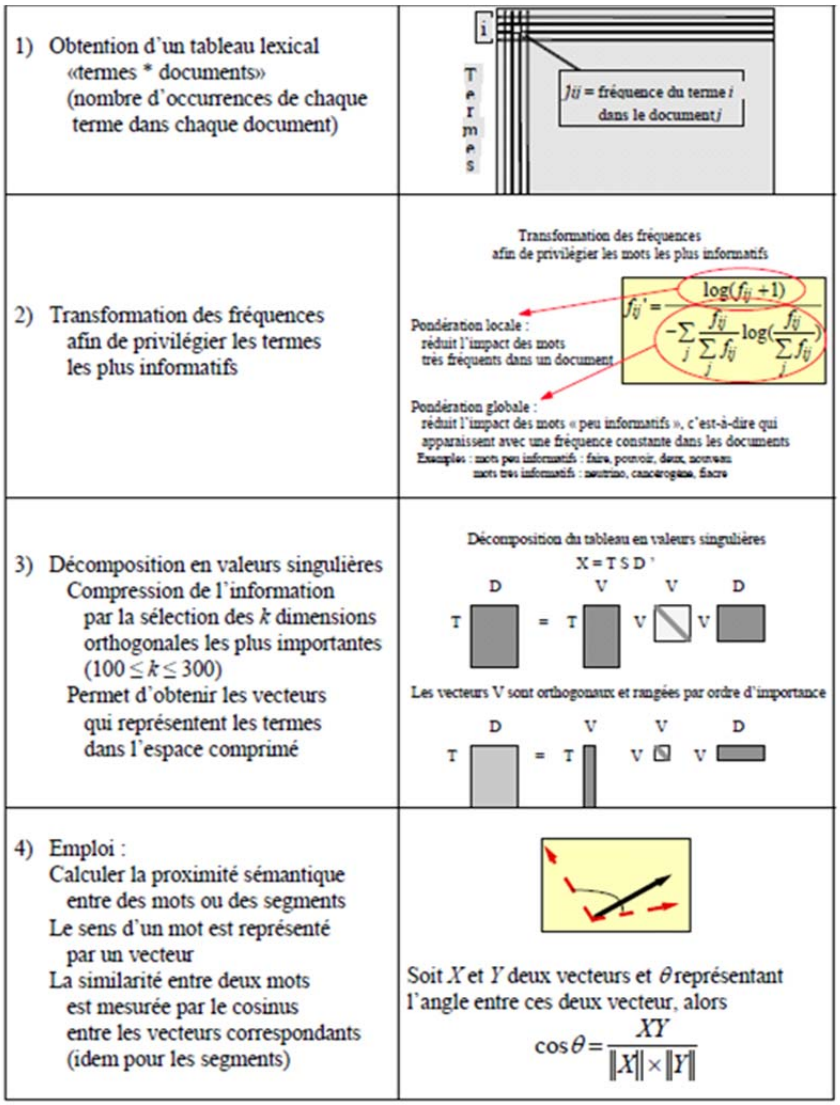


Figure 18 - Etapes de la LSA (issu de (Piérard et Bestgen, 2006b : 96))

La LSA permet ainsi d'inférer et de représenter le sens des mots suivant leur présence dans les documents. Dans l'espace sémantique, le sens de chaque mot est représenté par un vecteur. Pour mesurer la similarité sémantique entre deux mots, le cosinus entre leurs vecteurs est calculé (étape 4). Plus deux mots sont sémantiquement proches, plus leur cosinus sera élevé (la valeur maximale étant 1). Un cosinus de 0 indique une absence de similarité. Par exemple, dans (Zampa, 2003), l'autrice indique que, dans le corpus *Le Monde* de 1999, le cosinus entre « mémoire » et « souvenir » est de 0,7 tandis que celui entre « souvenir » et « ordinateur » n'est que de 0,07. Les termes « mémoire » et « souvenir » sont

donc liés sémantiquement, ce qui n'est pas le cas entre « souvenir » et « ordinateur ».

1.2.2.3 *CWM* (Choi *et al.*, 2001) : *C99* enrichi avec la LSA

Dans leur système *CWM*, (Choi *et al.*, 2001) ont donc utilisé la LSA afin d'améliorer le système *C99* initial. Dans cette nouvelle version de *C99*, la similarité entre phrases est estimée par la LSA, jugée plus qualitative que la mesure du cosinus par les auteurs, car elle permet de prendre en compte les synonymes et hypéronymes (*i.e.* la LSA n'est pas dépendante de la répétition stricte des mots). Ainsi, pour calculer la similarité entre phrases, la mesure du cosinus est alors appliquée aux vecteurs pondérés par les dimensions sémantiques (de 100 à 500 dimensions) plutôt qu'aux vecteurs bruts. Puis, les frontières thématiques sont toujours retrouvées suivant une analyse en grappes (cette étape est donc identique à la version originale de C99).

(Choi et *al.*, 2001) ont évalué *CWM* suivant le même protocole que pour C99. *CWM* a obtenu de meilleures performances que la version de base de C99 (11% d'erreurs en moins). Les auteurs ont aussi montré que la LSA était deux fois plus précise que la mesure du cosinus.

De leur côté, (Bestgen, 2004, 2006 ; Piérard et Bestgen, 2006b) ont évalué *C99* à maintes reprises et ont confirmé qu'il figurait parmi les meilleurs systèmes à l'heure actuelle. Adapté au français dans le cadre du projet Technolangue AGILE-OURAL, l'algorithme *C99* a été réévalué et s'est à nouveau révélé être le système le plus performant au regard des autres méthodes comparées (Bestgen et Piérard, 2006 ; Labadié et Prince, 2008d ; Sitbon, 2004 ; Sitbon et Bellot, 2004). Cependant, (Labadié et Chauché, 2007 ; Labadié et Prince, 2008a, 2008b, 2008e) notent que *CWM* a tendance à sous-segmenter les documents de grande taille (car il privilégie l'aspect qualitatif) et qu'il est difficile à mettre en œuvre lors de l'utilisation de larges corpus, car le nombre d'entrées dans la matrice explose (*e.g.* pour un volume de 400 000 phrases, la matrice contiendrait 1,6 * 10^{11} entrées).

1.2.3 *TextSeg*, méthode de segmentation par graphe (Utiyama et Isahara, 2001)

TextSeg (Utiyama et Isahara, 2001) se pose comme un concurrent direct de *C99* (Choi, 2000 ; Choi *et al.*, 2001). Indépendant du domaine, ce système sélectionne la segmentation optimale suivant la probabilité définie par un modèle statistique. Il permet ainsi d'obtenir des segments de longueurs variables suivant le corpus utilisé.

TextSeg[22] représente le document à segmenter sous la forme d'un graphe, où les nœuds (*e.g.* g0, g1, etc.) correspondent aux frontières thématiques potentielles et les arcs (*e.g.* e01, e14, etc.) symbolisent les segments (voir Figure 19).

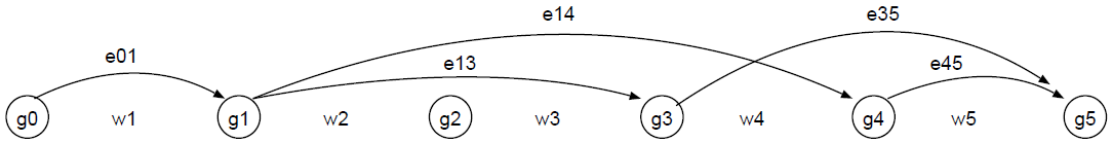


Figure 19 – Exemple de représentation par graphe des documents à segmenter (issu de (Utiyama et Isahara, 2001 : 495))

La segmentation thématique du document est trouvée en recherchant le meilleur « chemin » (*i.e.* dont le coût est minimal) dans un graphe valué[23] représentant toutes les segmentations possibles. Pour ce faire, chaque segment se voit affecter une valeur de cohésion lexicale fondée sur le calcul d'une probabilité généralisée privilégiant les segments homogènes (*i.e.* contenant de nombreuses répétitions). Ainsi, la probabilité généralisée est plus élevée lorsque les mots du segment sont répétés et plus faible lorsque les mots sont différents. La valeur de cohésion lexicale est vue comme la mesure de la capacité d'un modèle de langue[24], appris sur chaque segment, à prédire les mots des segments. Une fois la valeur de cohésion calculée, la segmentation maximisant cette valeur est alors préférée.

Afin d'évaluer leur système, les auteurs ont utilisé le corpus fourni par (Choi, 2000) ainsi que la procédure d'évaluation qui y est associée. *TextSeg* obtient des performances similaires à *C99* et demeure, par conséquent, plus performant que les autres systèmes déjà dépassés par *C99*.

TextSeg propose ainsi une approche dans laquelle le problème de segmentation thématique est ramené à trouver, parmi les segmentations possibles, celle menant aux segments les plus homogènes. D'après (Huet *et al.*, 2008), cette méthode est originale car elle se base uniquement sur la cohésion lexicale d'un segment : « À l'inverse des techniques exploitant une mesure de cohésion lexicale entre segments successifs, cette approche ne se base que sur la mesure de la cohésion au sein d'un segment ». Néanmoins, de même que *C99*, un des risques dans l'utilisation de cette méthode réside dans la sur-segmentation des documents (Simon *et al.*, 2013).

[22] L'algorithme est disponible à : https://sites.google.com/site/textseg/

[23] Un graphe valué est un graphe dans lequel chacune des arêtes se voit affecter une valeur.

[24] Un modèle de langue permet de capter les régularités linguistiques d'une langue. Il désigne une fonction de prédiction qui assigne une probabilité à un mot ou à une séquence de mots d'une langue. Par exemple, il permet de savoir si la séquence « animal, lire, cuisine » peut apparaître dans un texte donné.

1.2.4 *ClassStruggle*, méthode de segmentation par *clustering* (Lamprier *et al.*, 2008)

(Lamprier *et al.*, 2008) ont proposé une méthode de segmentation thématique par similarités plus globale que les autres approches que nous avons vues jusqu'à présent, prenant en compte non pas les mots mais les phrases. Pour segmenter thématiquement le texte, *ClassStruggle* utilise un algorithme de *clustering* évolutif (*i.e.* de regroupement) consistant à regrouper des phrases en classes (ou *clusters*) suivant leur degré de similarité. De ce fait, les phrases appartenant à la même classe auraient une forte similarité, tandis que les phrases appartenant à des classes différentes possèderaient une faible similarité. Au cours du processus, les classes évoluent suivant la position et la proximité des phrases dans le texte afin de créer des groupes ne contenant que des phrases relatives à un thème. Les frontières entre segments thématiques sont alors apposées entre les phrases appartenant à des classes différentes.

Les auteurs fournissent l'exemple d'application de la méthode *ClassStruggle* suivant (voir Figure 20) : soit un corpus composé de 10 phrases (de A à J) concaténées issues de 2 articles de journaux différents, les phrases A à G appartenant au premier article et les phrases G à J au second. L'application initiale de l'algorithme de *clustering* sur ce corpus a donné lieu à l'identification de 3 groupes de phrases C1(D,E), C2(A,B,F,G) et C3(C,H,I,J).

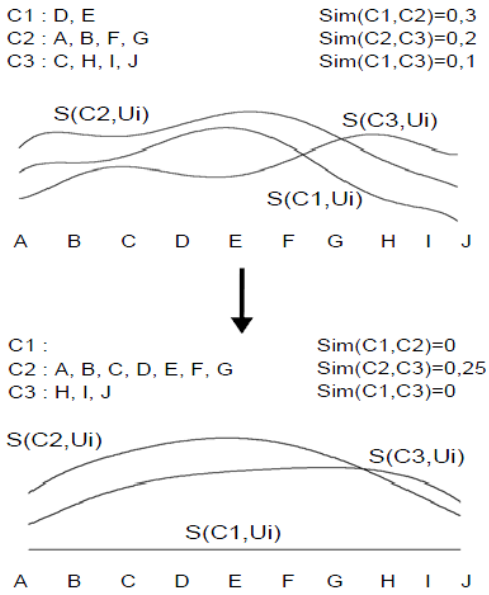


Figure 20 – Exemple d'application de *ClassStruggle* (Lamprier *et al.*, 2008 : 184)

Les potentiels d'appartenance des phrases aux 3 groupes initialement identifiés sont alors représentés par des courbes. La hauteur de ces courbes informe sur le

score d'appartenance d'une phrase à un groupe (*e.g.* couple (groupe, phrase)). De ce fait, la courbe représentant le couple (groupeC3, $phrase_n$) montre que le groupe C3 se trouve dominant à partir de la phrase H jusqu'à la fin du texte (*e.g.* H, I, J), le groupe C2 est dominant de A à G et le groupe C1 n'est jamais dominant. Il est alors vidé de toutes ses phrases. A l'issue de cette étape, la méthode appose une frontière thématique entre les phrases G et H, ce qui correspond à la zone de concaténation des deux articles.

ClassStruggle a d'abord été évalué sur un corpus artificiel composé de 350 articles concaténés issus de l'*Associated Press* (corpus TREC[25]-1, (Harman, 1993)), suivant la méthodologie de (Choi, 2000). Sur ce premier corpus, le système a obtenu globalement des performances plus élevées que d'autres systèmes tels que *TextTiling* ou *C99*. (Lamprier *et al.*, 2008) ont ensuite évalué leur méthode sur deux autres corpus spécialisés (revues informatiques), toujours issus du corpus TREC-1, composés chacun de 350 articles concaténés. Sur ces corpus spécialisés, *ClassStruggle* a perdu son avantage sur les autres systèmes : les phrases de ces corpus étant relativement similaires, le processus de regroupement a éprouvé des difficultés à segmenter thématiquement, ce qui a entraîné une diminution des performances de *ClassStruggle*. D'après les auteurs, l'utilisation de méthodes de *clustering* plus sophistiquées (*i.e.* utilisant des ressources sémantiques) permettrait de pallier ce problème.

1.2.5 Bilan

A l'issue de la présentation de quelques méthodes descendantes, nous avons pu constater que celles-ci se fondaient, pour la majeure partie, sur la mesure de la cohésion lexicale entre énoncés (*i.e.* à un niveau global) afin d'obtenir les segmentations la maximisant. De même que les méthodes ascendantes, les méthodes descendantes nécessitent des prétraitements minimaux (segmentation en phrases, lemmatisation) et s'appliquent directement sur tous types de documents.

Afin de pallier les erreurs de segmentation liées à la présence de synonymes dans les documents, nous avons vu que certaines méthodes telles que (Choi *et al.*, 2001 ; Brants *et al.*, 2002[26]) utilisent l'analyse sémantique latente et améliorent ainsi leurs performances. L'efficacité de *C99*, évaluée et démontrée maintes fois depuis sa création, nous invite à utiliser cet algorithme (dans sa version améliorée avec la LSA) dans notre projet pour segmenter les documents en sections

[25] *Text REtrieval Conference*

[26] Ces auteurs utilisent la PLSA (*Probabilistic Latent Semantic Analysis*, (Hofmann, 1999)) qui est une amélioration de la LSA incluant un modèle statistique.

thématiquement homogènes. Nous reviendrons en détails sur nos choix dans la conclusion de ce chapitre.

1.3 Méthodes hybrides

Nous venons de voir que deux familles de méthodes statistiques permettaient de segmenter automatiquement les textes en thèmes. Ces méthodes privilégient soit le repérage de changements dans le vocabulaire afin de délimiter des frontières thématiques locales (méthodes ascendantes), soit la maximisation de la cohésion lexicale entre phrases afin de les regrouper (méthodes descendantes). Certaines méthodes hybrides visent à mêler ces deux types de stratégies afin de proposer à la fois une segmentation textuelle contenant des segments thématiquement cohérents mais aussi bien différenciés les uns des autres.

1.3.1 Approches par chaînes lexicales et par similarité (Sitbon, 2004)

(Sitbon, 2004) propose des méthodes basées sur la combinaison d'approches à base de chaîne lexicales et de matrices de similarité pour optimiser la segmentation thématique :

- afin d'améliorer les résultats des méthodes par matrices de similarité, l'autrice propose de renforcer la similarité entre les phrases contenant des chaînes lexicales actives à un moment donné du texte. Pour ce faire, la méthode enrichit la matrice de similarité que l'algorithme *C99* utilise *via* un lissage de Laplace. Ce lissage consiste à ajouter, dans chaque phrase, une occurrence de chaque terme correspondant à une chaîne lexicale active,
- une autre méthode proposée pour optimiser la cohésion avant un traitement par matrice de similarité consiste à effectuer une similarité sur les chaînes lexicales actives des phrases (au lieu des mots). Ainsi, après avoir ajouté les termes des chaînes lexicales actives, (Sitbon, 2004) propose de retirer les termes qui ne correspondent pas à une chaîne lexicale active.

Ces méthodes hybrides ont été testées sur un corpus composé d'articles du *Monde* (2001) et de la *Bible*. Les résultats préliminaires obtenus montrent une amélioration des performances des outils testés. Néanmoins, ces expérimentations restent à l'état de pistes.

1.3.2 Méthode par calcul de distance thématique (Labadié et Chauché, 2007)

De leur côté, pour segmenter thématiquement les documents, (Labadié et Chauché, 2007) utilisent une représentation vectorielle des phrases fournie par l'analyseur morpho-syntaxique et conceptuel *SYGFRAN* (Chauché, 1984), ainsi que des calculs de distance entre les vecteurs des phrases. C'est en recherchant les zones de transition dans le document *via* les vecteurs sémantiques qu'est menée la segmentation thématique. L'objectif n'est donc pas de trouver la phrase qui marque un changement de thème mais plutôt les zones du texte où les thèmes changent.

La méthode a été évaluée sur le corpus DEFT'2006 (Azé *et al.*, 2006), composé de discours politiques, d'articles de loi et d'un extrait d'un ouvrage scientifique. Elle s'est classée 4ème parmi d'autres systèmes statistiques. Un des inconvénients majeurs de cette méthode est qu'elle est totalement dépendante de l'analyse en concepts fournie par *SYGFRAN*.

1.3.3 Méthode combinant cohésion lexicale et rupture lexicale (Simon *et al.*, 2013)

(Simon *et al.*, 2013) ont proposé une méthode combinant la maximisation de la cohésion lexicale d'un segment textuel à la détection de ruptures lexicales pour la segmentation thématique des documents. Les auteurs ont défini un modèle probabiliste, indépendant du domaine, basé sur la segmentation par graphe de (Utiyama et Isahara, 2001, *cf.* 1.2.3) et intégrant la rupture lexicale. Pour ce faire, le modèle considère que les segments du texte ne sont pas indépendants les uns des autres (à la différence de la méthode de base de (Utiyama et Isahara, 2001)). Une hypothèse markovienne[27] permet alors de considérer, pour chaque segment, le segment le précédant. Ainsi, à chaque nœud du graphe du document où une frontière potentielle est définie, toutes les combinaisons possibles des segments précédant la frontière sont analysées (*i.e.* les arcs). De ce fait, la rupture lexicale est calculée pour l'ensemble des paires de segments potentielles.

La méthode hybride de (Simon *et al.*, 2013) a été évaluée sur les 700 textes écrits concaténés du corpus *Brown* suivant la méthodologie de (Choi, 2000) ainsi que sur des transcriptions automatiques de 56 journaux télévisés français (dans ce

[27] Le recours à l'hypothèse markovienne permet de représenter les propriétés statistiques d'un système *via* des probabilités sans nécessiter la description de la structure complète du système.

dernier cas, un thème correspond à un reportage dans une émission). Les performances obtenues par le système hybride sur le corpus *Brown* sont plus élevées que celles obtenues par le modèle standard de (Utiyama et Isahara, 2001), surtout lorsque les segments textuels sont longs. En revanche, sur le corpus de journaux télévisés, les résultats obtenus montrent une amélioration limitée des performances de cette méthode hybride par rapport à l'utilisation de la valeur de cohésion lexicale seule (*i.e.* au modèle standard de (Utiyama et Isahara, 2001)), car ces journaux comportent peu de répétitions mais de nombreux synonymes. D'autres techniques de calcul de la rupture lexicale telles que la vectorisation[28] (Claveau et Lefèvre, 2011b) sont envisagées par les auteurs pour améliorer leur méthode.

1.4 Bilan

Quelle que soit la méthode de segmentation thématique statistique adoptée (ascendante, descendante ou hybride) et les indices utilisés pour détecter les segments thématiques (liens de répétition lexicale, de rupture, utilisation de ressources externes), l'objectif principal de la segmentation statistique est de découper un texte en segments homogènes du point de vue de ses thèmes.

Il nous semble important de souligner, avec (Khalis, 2006), que l'efficacité d'un système est fonction du type de document à segmenter (Labadié et Prince, 2008e), de la taille du document et de la variation de taille des segments à repérer (Sitbon, 2004 ; Sitbon et Bellot, 2004). Certaines évaluations consistent à comparer l'emplacement des segments thématiques fourni par les systèmes à celui déterminé par des juges humains (Passonneau et Litman, 1993 ; Labadié et Prince, 2008c). Cette dernière remarque amène à reconnaître la dimension subjective de cette évaluation (on assiste à une variabilité entre les segmentations attribuées par les juges), comme nous avons déjà pu le signaler. De son côté, (Choi, 2000) fournit un corpus d'évaluation composé d'articles différents concaténés. Bien qu'artificiel, ce corpus permet d'effectuer une évaluation automatique de plusieurs outils sur un même matériau et il s'affranchit ainsi des problèmes de subjectivité rencontrés par les segmentations manuelles. Néanmoins, comme le soulignent (Ferret, 2006b, 2007 ; Georgescul *et al.*, 2006), les systèmes ayant obtenu des performances élevées sur un tel corpus artificiel voient leurs performances chuter face à un corpus réel, notamment parce que les divers

[28] La vectorisation consiste à projeter un calcul de similarité entre deux documents dans un espace vectoriel.

critères utilisés par les juges humains pour segmenter sont difficilement reproductibles automatiquement (Bestgen et Piérard, 2006).

Parmi les méthodes statistiques présentées, peu d'entre elles ont été adaptées et évaluées sur des corpus français. Seul *C99* (Choi *et al.*, 2001) a fait l'objet de multiples évaluations qui ont, pour la majeure partie, prouvé l'efficacité de cette méthode par similarité. Ces raisons nous invitent donc à utiliser *C99* dans notre projet. Néanmoins, cette méthode statistique, bien que bénéficiant d'informations sémantiques (*via* la LSA), ne permet pas de dégager la structure hiérarchique d'un document (Pimm, 2008) permettant d'atteindre les segments thématiques les plus informatifs pour un utilisateur. De ce fait, le recours à des méthodes linguistiques utilisant des marqueurs tels que les anaphores, les connecteurs ou les chaînes de référence semble constituer une solution. Nous proposons de présenter dans la section suivante des méthodes linguistiques exploitant ces divers marqueurs de cohésion.

2 Systèmes linguistiques

Bien que les méthodes statistiques de segmentation thématique mettent en œuvre des algorithmes peu coûteux, la nature même du segment thématique n'est pas clairement définie (Widlöcher *et al.*, 2006). En effet, l'utilisation quantitative de la notion de *cohésion lexicale* facilite le traitement informatique pour segmenter thématiquement un texte, mais l'organisation thématique d'un texte fait intervenir d'autres indices de cohésion que la seule répétition : les anaphores, les chaînes de référence, les cadres de discours (*cf.* chapitre 1). Ainsi, moins nombreux que les systèmes statistiques, quelques systèmes détectant des indices linguistiques pour la segmentation thématique ont été mis en place[29]. Nous donnons un aperçu de ces systèmes dans la section suivante.

2.1 Utilisation de marqueurs discursifs (Passonneau et Litman, 1993, 1995, 1997)

(Passonneau et Litman, 1993) ont proposé trois algorithmes de segmentation thématique linéaire basés sur l'utilisation de marqueurs linguistiques :

> « Our work is motivated by the hypothesis that natural language technologies can more sensibly interpret discourse, and can generate more comprehensible discourse, if they take advantage of this interplay between segmentation and linguistic devices. » (Passonneau et Litman, 1997 : 104)

Travaillant sur des corpus oraux anglais (monologues en discours spontané), les autrices ont choisi d'utiliser les relations anaphoriques (algorithme NP-A), les connecteurs (algorithme CUE-A) et les pauses (algorithme PAUSE-A) pour segmenter automatiquement leurs transcriptions. Pour mettre en place l'algorithme des anaphores nominales, un corpus d'apprentissage annoté en relations anaphoriques a été fourni en entrée. De son côté, l'algorithme des pauses a bénéficié d'un corpus d'apprentissage annoté en pauses[30]. Suivant le type de

[29] Le faible nombre de systèmes linguistiques développés est notamment souligné par (Ferret, 2006b) et (Passonneau et Litman, 1997) : « there has been little work on examining the use of linguistic cues for recognizing or generating segment boundaries » (Passonneau et Litman, 1997 : 116).

[30] Chaque corpus d'apprentissage correspond à 10 monologues.

marqueur, chaque algorithme utilise un ou plusieurs traits pour identifier les frontières thématiques (*e.g.* le groupe nominal de la phrase précédente coréfère-t-il avec le pronom de la phrase en cours ?, la phrase est-elle initiée par un connecteur ?).

Les trois algorithmes ont été évalués en comparant les segmentations automatiques produites par les algorithmes sur 20 monologues[31] (13 500 mots environ) aux segmentations fournies par au moins 4 juges humains sur 7. Les meilleures performances ont été obtenues par l'algorithme NP-A. Néanmoins, (Passonneau et Litman, 1993) ont noté une faible corrélation entre les segmentations automatiques et les segmentations humaines. Ces faibles résultats sont à rapprocher des problèmes d'accord inter-annotateurs rencontrés en amont. En effet, bien que les autrices relèvent une concordance inter-annotateurs située entre 82% et 92% pour le placement des frontières thématiques, elles ont aussi mis l'accent sur d'importantes variations dans les taux de placement des frontières (de 5,5% à 41,3%) suivant les juges (*i.e.* certains juges ont placé beaucoup plus de frontières thématiques que d'autres).

(Litman et Passonneau, 1995) ont ensuite testé la combinaison de paires de marqueurs (*i.e.* NP-A et CUE-A, NP-A et PAUSE-A, etc.) ainsi que la combinaison des trois marqueurs en comparant les segmentations automatiques aux segmentations fournies, cette fois, par au moins 3 des 7 juges humains. Les meilleures performances ont été obtenues pour la paire (NP-A et PAUSE-A) mais aucune combinaison de marqueurs n'est parvenue à égaler les segmentations humaines. Afin d'améliorer les performances obtenues, les autrices ont effectué une analyse de leurs erreurs et elles ont utilisé des techniques d'apprentissage automatique. Evalués sur un corpus de 5 nouveaux monologues, les systèmes optimisés ont obtenu des performances proches des performances humaines, sans toutefois parvenir à s'aligner sur ces dernières.

Les travaux de Passonneau et Litman sont intéressants dans la mesure où ils mettent en évidence l'importance de l'utilisation de marqueurs linguistiques pour la segmentation thématique. Néanmoins, les algorithmes développés par les autrices nécessitent en entrée des corpus enrichis en annotations manuelles (relations anaphoriques, pauses), ce qui a pour inconvénient majeur de rendre les outils dépendants des ressources (*i.e.* de l'existence de ces ressources, de leur taille et de leur disponibilité suivant la langue[32]).

[31] Ces monologues, nommés *Pear stories*, sont des résumés d'un même film. Ils ont été collectés et transcrits par (Chafe, 1980).

[32] Par exemple, les ressources pour le français sont peu nombreuses, voire inexistantes suivant le type de marqueur et le domaine, *cf.* chapitre 5.

2.2 Utilisation d'indices de continuité et discontinuité thématique (Piérard *et al.*, 2004)

Afin de valider automatiquement la fonction de marqueurs de segmentation thématique de trois types d'expressions temporelles (ancreurs (*e.g.* « vers deux heures »), enchaîneurs (*e.g.* « ensuite ») et connecteurs (*e.g.* « et »)), (Piérard *et al.*, 2004) ont utilisé 4 indices de continuité et discontinuité thématique :

- une *mesure référentielle*, inspirée du modèle de (Kintsch et van Dijk, 1978)[33] permettant de déterminer que deux phrases sont liées (*i.e.* continuité thématique) à partir du moment où leurs propositions contiennent des arguments communs (*i.e.* des noms ou des pronoms),
- une *mesure des anaphores interphrases* grammaticales (pronoms personnels, démonstratifs et possessifs), indices de continuité thématique (vu que les anaphores sont des expressions référentielles non autonomes),
- deux mesures basées sur l'analyse sémantique latente[34] (Landauer *et al.*, 1998) : une *mesure du cosinus* (si une phrase cible introduit un changement de thème, son cosinus avec la phrase la précédant doit être inférieur à celui de la phrase qui la suit) et une *mesure de segmentation* obtenue à partir d'un paramétrage de l'algorithme de segmentation *C99* (Choi *et al.*, 2001) (identification des ruptures classées par ordre d'importance entre 10 phrases successives).

Les auteurs ont postulé que ces marqueurs temporels se distribueraient sur une échelle de continuité/rupture thématique de la manière suivante (voir Figure 21) :

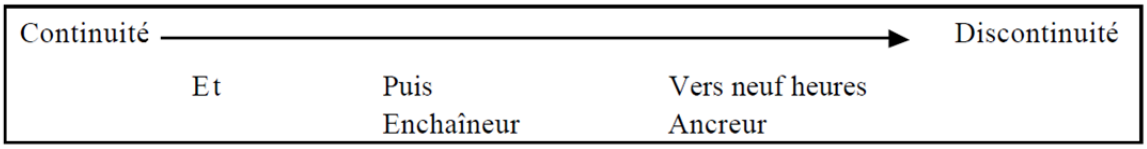


Figure 21 - Echelle de continuité/discontinuité thématique (d'après (Piérard *et al.*, 2004 : 860)

Pour évaluer l'efficacité des 4 indices de cohésion dans le signalement d'un niveau de continuité/discontinuité différent suivant le marqueur de segmentation

[33] *Cf.* chapitre 1.
[34] *Cf.* section 1.2.2.2 de ce chapitre.

thématique, (Piérard *et al.*, 2004) ont constitué un corpus de 152 et 270 phrases[35] issues de textes littéraires (roman, nouvelles et contes) provenant des bases *Frantext* et *ABU*[36]. Pour les deux premiers indices, les valeurs obtenues par le calcul automatique ont aussi été comparées à celles obtenues par un juge, afin de valider la fiabilité de ce calcul. Les valeurs moyennes de continuité obtenues pour les trois types d'expressions temporelles suivent l'échelle postulée : le connecteur « et » est associé à la continuité la plus élevée, le marqueur d'ancrage est associé à la plus forte discontinuité. De son côté, le marqueur d'enchaînement est, pour 3 des 4 mesures, plus proche du connecteur « et » que de l'ancreur. Ainsi, l'expérience menée par (Piérard *et al.*, 2004) a montré que le repérage automatique d'indices référentiels permet de valider la fonction de marqueurs de segmentation thématique d'expressions temporelles. Les auteurs soulignent que l'indice anaphorique a donné lieu à des différences entre les 3 marqueurs temporels plus nettes qu'avec les autres indices. De ce fait, la prise en compte de ce type d'indice semble constituer un moyen particulièrement fiable. Ces relations anaphoriques sont aussi exploitées par (Smolczewska et Lallich-Boidin, 2004), en complément des marqueurs dispositionnels (*e.g.* structures de listes), afin de segmenter des documents techniques en unités thématiquement homogènes.

2.3 Bilan

Les approches linguistiques présentées dans cette section montrent que l'utilisation de marqueurs de la structuration du discours permet d'identifier de manière précise les frontières thématiques des documents (Ferret, 2006b). Néanmoins, leur utilisation exclusive ne garantit pas l'obtention systématique d'une segmentation thématique fiable (ambiguïté des marqueurs tels que les connecteurs conclusifs (*i.e.* « donc ») ou absence de marqueurs suivant le genre textuel (*e.g.* les cadres de discours peuvent être absents d'un texte)). De ce fait, l'utilisation d'indices linguistiques, en complément des algorithmes statistiques, permettrait d'optimiser la segmentation thématique. C'est dans cette optique que de nombreux systèmes hybrides, combinant des méthodes statistiques et linguistiques, ont été proposés. Nous présentons plusieurs de ces systèmes dans la section suivante.

[35] Les 270 phrases sont uniquement utilisées pour l'évaluation automatique, les 152 phrases sont utilisées pour l'évaluation automatique et l'évaluation manuelle par un juge.
[36] http://abu.cnam.fr/

3 Systèmes hybrides

Les systèmes hybrides combinent une modélisation statistique des thèmes et l'utilisation d'un ou plusieurs marqueurs linguistiques, afin d'optimiser la segmentation des textes en blocs thématiquement homogènes. Certains de ces systèmes fournissent, en plus du découpage thématique, des descripteurs thématiques permettant d'assigner des thèmes et sous-thèmes aux segments délimités.

3.1 L'approche mixte de (Beeferman *et al.*, 1999)

Afin d'identifier automatiquement les frontières thématiques de flux textuels, (Beeferman *et al.*, 1999) ont proposé d'utiliser, en plus d'une méthode statistique, les connecteurs. La méthode statistique repose sur une approche probabiliste utilisant des modèles exponentiels[37] pour extraire des traits corrélés à la présence de frontières thématiques repérées dans un corpus d'apprentissage. L'idée est d'assigner, pour chaque phrase du corpus (textuel ou multimédia), une probabilité qu'il existe une frontière thématique entre cette phrase et la phrase la succédant.

Le modèle proposé utilise deux classes de traits sélectionnés automatiquement à partir d'un grand espace de traits potentiels :

- des traits thématiques utilisant des modèles de langues (*i.e.* des trigrammes) pour détecter les changements de thèmes,
- des traits identifiant des connecteurs (pouvant être dépendants du domaine) potentiellement présents dans le voisinage de frontières thématiques.

Le modèle a été entraîné sur un corpus annoté en segments thématiques de plusieurs millions de mots, composé d'articles du *Wall Street Journal* et de transmissions d'émissions de télévision. Comparé à *TextTiling*, le modèle obtient

[37] Un modèle exponentiel permet de prédire un phénomène. Par exemple, dans le roman *Pay it forward* de Catherine Ryan Hyde, une personne doit rendre service à 3 personnes le jour 1 et chacune de ces 3 personnes doit, à son tour, rendre service à 3 personnes. De ce fait, à l'issue du jour 2, les 3 personnes auront rendu service à 9 personnes. Le jour 3, les 9 personnes auront rendu service à 27 personnes, et ainsi de suite. Ce modèle algébrique prend alors la forme : $m(j)=3^{j+1}$ et permet de prédire à combien de personnes on aura rendu service au bout d'un nombre de j jours. Ce modèle algébrique est une fonction exponentielle.

de meilleures performances lorsque les traits thématiques sont utilisés avec les connecteurs. Mais, bien que performant, le modèle mixte de (Beeferman *et al.*, 1999) nécessite un large corpus d'apprentissage annoté en segments thématiques, qu'il n'est pas simple de constituer ou d'obtenir suivant la langue et le domaine.

3.2 Méthode par cohésion lexicale, réseau de collocations et cadres de discours (Ferret *et al.*, 2001)

Pour repérer les thèmes textuels pour une tâche de résumé automatique, (Ferret *et al.*, 2001) ont proposé un modèle alliant deux méthodes statistiques (l'une par cohésion lexicale et l'autre par réseau de collocations) à un système linguistique identifiant les cadres de discours (Charolles, 1997).

Dans le versant statistique, deux méthodes d'analyse thématique identifient les points de ruptures thématiques *via* un critère de distribution lexicale. La première méthode, similaire à *TextTiling* (Hearst, 1994), s'appuie sur la cohésion lexicale afin de déterminer les frontières thématiques. La seconde méthode utilise un réseau de collocations génériques (*i.e.* non spécifiques) construit automatiquement à partir d'un corpus d'articles de journaux dans (Ferret *et al.*, 1998). Ces connaissances externes permettent d'augmenter la valeur de certains mots du texte lorsqu'ils sont liés significativement dans le réseau à d'autres mots du paragraphe. La méthode permet alors de rapprocher des paragraphes contenant des mots liés.

Dans le versant linguistique, les auteurs repèrent automatiquement les introducteurs de cadres thématiques, qui indiquent quel est le thème de la proposition qui les suit (*e.g.* « *Concernant* la hausse du chômage… »). L'identification des introducteurs de cadres thématiques a été menée en utilisant la méthode d'exploration contextuelle (Desclés *et al.*, 1997 ; Minel *et al.*, 2001). Cette méthode permet de définir des règles contextuelles afin d'identifier et d'étiqueter des portions textuelles suivant des conditions telles que leur contexte droit et/ou gauche. Chaque règle contextuelle délimite un espace de recherche constitué d'un indicateur (ou marqueur déclencheur, *i.e.* un introducteur thématique) et d'indices complémentaires tels que la position de l'indicateur dans la phrase, la présence ou l'absence de certains termes ou marques typographiques (*e.g.* l'indicateur « au chapitre » ne doit pas être suivi d'un chiffre (*i.e.* « au chapitre 3 ») pour être un introducteur thématique). Pour la tâche de segmentation thématique, 7 règles de formation ont été définies afin de repérer

plusieurs configurations textuelles dans lesquelles ces marqueurs doivent apparaître (*e.g.* l'introducteur de cadre thématique est situé en position initiale, l'introducteur fait partie d'une énumération, etc.). L'évaluation des règles, menée sur un corpus composé d'articles du *Monde Diplomatique*, a montré des résultats satisfaisants pour le repérage des introducteurs de cadres thématiques, mais les règles définies pour le détachement des introducteurs gagneraient à être affinées[38]. De leur côté, les méthodes statistiques se révèlent performantes lorsque les cassures entre segments sont franches. Mais lorsque ce n'est pas le cas, les marqueurs linguistiques apportent de meilleurs résultats. De ce fait, la combinaison de méthodes statistiques et linguistiques s'avère efficace, les premières venant compléter les secondes.

L'approche hybride de (Ferret *et al.*, 2001) montre la complémentarité des méthodes linguistiques et statistiques pour la tâche de segmentation thématique. La méthode linguistique adoptée est intéressante dans la mesure où elle ne se restreint pas à identifier les introducteurs de cadres thématiques sur la seule base de leur forme graphique, elle utilise un ensemble de règles contextuelles régissant les conditions dans lesquelles ces introducteurs doivent apparaître. Néanmoins, l'utilisation de ressources externes dans la méthode statistique par réseau de collocations peut constituer un frein, car ces ressources sont difficiles à constituer (Todirascu et Gledhill, 2008).

3.3 Méthode hybride par descripteurs thématiques (Hernandez, 2004)

Dans le cadre de la modélisation d'un système de résumé dynamique automatique s'adaptant aux besoins d'un utilisateur, (Hernandez, 2004) a proposé deux mécanismes d'identification de thèmes utilisés conjointement : un système de résolution des anaphores (*SRA*) et un système de construction des chaînes lexicales (*CCL*). Dans son approche, les descripteurs thématiques sont alors les antécédents potentiels pour *SRA* et les maillons initiateurs des chaînes lexicales pour *CCL*.

SRA[39] utilise des heuristiques syntaxiques fondées sur des indices de surface pour identifier les expressions référentielles du texte (groupes nominaux simples et pronoms). Puis, pour une anaphore donnée, le système sélectionne les candidats

[38] Pour ce faire, on pourrait par exemple s'appuyer sur un étiquetage des différents types d'entités nommées (noms de personne, de lieu, d'organisation, etc.).

[39] Le système *SRA* sera présenté en détails au chapitre 5 (section 1.3.2).

antécédents suivant leur pertinence et la validation de contraintes syntaxiques (*i.e.* correspondance de la tête lexicale entre l'antécédent et l'anaphore, correspondance du genre et du nombre entre l'antécédent et l'anaphore). Le premier candidat antécédent validant le plus haut degré est alors considéré comme l'antécédent de l'anaphore.

De son côté, *CCL* utilise trois types de relations lexicales pour identifier les chaînes lexicales : les relations par variations morphosyntaxiques, les relations sémantiques et lexicales (*i.e.* répétition, synonymie, hyperonymie, hyponymie) et les relations par variation de morphologie dérivationnelle. Ainsi, l'algorithme de *CCL* parcourt le texte de terme candidat en terme candidat et recherche, pour chaque terme candidat, une similarité lexicale, sémantique ou dérivationnelle avec ses proches voisins. Pour identifier les termes candidats, le système utilise 3 ressources : la base de connaissances sémantiques de *WordNet* (Miller, 1995 ; Fellbaum, 1998), l'étiquetage morphosyntaxique et la lemmatisation fournis par l'étiqueteur TreeTagger (Schmid, 1994) ainsi que le dictionnaire de morphologie dérivationnelle CELEX (Jacquemin, 1997). Le système combine la prise en compte de la distance entre deux termes pour évaluer la validité de leur lien sémantique, et identifie les associations potentielles entre termes. L'utilisation des ressources sémantiques améliore la qualité de l'identification des chaînes lexicales en validant le rapprochement entre certains termes candidats.

Les deux systèmes ont été évalués[40] manuellement sur deux courts extraits (moins de 200 mots) du texte exemple de (Barzilay et Elhadad, 1997). Le système *SRA* obtient des performances comparables aux autres systèmes existants pour la résolution des anaphores pronominales. De son coté, le système *CCL* s'aligne sur celui de (Barzilay et Elhadad, 1997) mais il s'en démarque grâce aux ressources. En effet, la prise en compte des variantes morphosyntaxiques (*via* TreeTagger) lui permet d'identifier un nombre plus important d'entités thématiques. Néanmoins, *CCL* est dépendant de ces ressources, ce qui constitue une limite à son utilisation[41].

L'approche de (Hernandez, 2004) est intéressante car elle est une des rares méthodes à proposer des termes décrivant les thèmes des documents (les antécédents pour le système *SRA* et les premiers maillons des chaînes lexicales pour *CCL*).

[40] L'évaluation de *SRA* et *CCL* a été effectuée séparément, les deux systèmes étant pour le moment indépendants l'un de l'autre.

[41] L'auteur, qui travaille sur des corpus français et anglais avec *SRA*, reconnaît être restreint à travailler seulement sur un corpus anglais avec *CCL* en raison des ressources à sa disposition.

3.4 L'approche mixte d'(Hurault-Plantet *et al.*, 2006)

Pour résoudre la tâche de segmentation thématique du défi DEFT'2006, (Hurault-Plantet *et al.*, 2006) ont combiné une méthode par cohésion lexicale à des marqueurs linguistiques. La tâche du défi a consisté à trouver la première phrase de chaque segment thématique dans 3 corpus distincts (discours politiques, textes juridiques, ouvrage scientifique).

Dans leur approche mixte, les auteurs ont d'abord appliqué *TextSeg* (Utiyama et Isahara, 2001)[42] afin d'obtenir une première segmentation thématique. Les frontières des segments obtenus ont ensuite été corrigées en utilisant une méthode par apprentissage des marqueurs linguistiques de continuité et de rupture thématique, afin de trouver les limites exactes des segments thématiques. Pour ce faire, un modèle de langage n-grammes[43] de mots a été utilisé pour connaître la probabilité qu'un mot, situé en début ou en fin de phrase, ouvre, ferme ou soit à l'intérieur d'un segment thématique. Dans ce modèle, un marqueur de rupture thématique est discriminant s'il se trouve, dans plus d'un cas sur deux, dans une phrase de début de segment, donc s'il a une probabilité supérieure à 0,5 ; un marqueur de continuité thématique est discriminant s'il ne se situe jamais en début de segment, donc s'il possède une probabilité égale à 1. Parmi les 3 corpus, seul le corpus de discours politiques a contenu des n-grammes caractérisant le début (*i.e.* « mes chers compatriotes », « bonsoir madame ») ou la fin des segments (*i.e.* « je vous prie », « je lève mon »). Pour le corpus juridique, les auteurs ont recherché, dans une fenêtre de 10 phrases autour de la phrase frontière, la phrase la plus proche contenant l'introducteur « Article X » (car les auteurs ont remarqué que les articles juridiques de ce corpus débutaient souvent par cet introducteur). La segmentation a enfin été fusionnée avec une nouvelle segmentation obtenue par un calcul de probabilité qu'une phrase constitue un début de segment thématique. Cette dernière étape a été mise en place afin d'ajouter des segments contenant des marqueurs de rupture qui n'avaient pas été identifiés en amont par *TextSeg*.

L'évaluation de l'approche mixte d'(Hurault-Plantet *et al.*, 2006) a obtenu des performances situées entre 30% et 50% suivant le genre textuel. Bien que faibles, les résultats obtenus sont supérieurs à la moyenne des autres participants au défi.

[42] Voir la section 1.2.3 de ce chapitre pour une présentation de *TextSeg*.

[43] Un modèle n-grammes est un modèle probabiliste qui permet de prédire un mot connaissant les *n-1* mots précédents.

Néanmoins, de même que les autres approches par apprentissage, les performances de cette approche hybride dépendent du corpus d'apprentissage.

4 Discussion

Nous souhaitons discuter dans cette section d'un point, souligné notamment par (Bilhaut, 2006), concernant la définition d'un segment thématique. En effet, les différentes approches que nous avons passées en revue omettent souvent de définir cette notion, se focalisant uniquement sur le développement de leur système. Et, lorsque cette notion est définie, elle montre clairement le flou dont elle fait l'objet. Ce phénomène est relativement saillant dans la présentation de la campagne d'évaluation DEFT 2006. En effet, dans (Azé et *al.*, 2006), les organisateurs de la campagne DEFT reconnaissent, à juste titre, la difficulté pour définir un segment thématique. Ils proposent alors d'adapter cette définition au genre textuel du corpus. De ce fait, vu que le corpus d'apprentissage du défi était composé de 3 genres textuels (discours politiques, textes juridiques, ouvrage scientifique), 3 définitions différentes de la notion de *segment thématique* ont été données :

> « La définition générale d'un segment thématique est très problématique, c'est pourquoi nous avons choisi une définition différente pour chaque corpus. Nous avons privilégié la segmentation voulue par les auteurs des textes qui est aussi la plus simple à utiliser pour la préparation des corpus.
> Pour les discours politiques, la segmentation thématique est basée sur la structure thématique des discours mis en ligne sur le site de référence. Chaque discours a été divisé en paragraphes thématiques lors de leur écriture ou lors de la constitution des corpus mis en ligne par l'organisme en charge de cette tâche.
> Pour les lois de l'Union Européenne, les segments thématiques sont les lois.
> Pour l'ouvrage scientifique, les segments thématiques à retrouver sont les différentes sections, à savoir les chapitres, sections, sous-sections et sous-sous-sections. » (Azé et *al.*, 2006 : 3)

Il nous paraît difficile d'adapter la définition d'une notion « à la carte » en s'appuyant uniquement sur des marques typodispositionnelles (le segment thématique est soit un paragraphe, soit un chapitre, soit une section, soit une sous-section). Une modélisation linguistique tendrait à définir plus justement cette notion.

5 Conclusion

Dans ce chapitre, nous avons présenté divers systèmes automatiques pour la détection de thèmes. Nous avons ainsi pu relever que ces systèmes divergeaient, et ce, à plusieurs titres :

- le type d'indice utilisé pour identifier les points de rupture thématique dans le discours (*i.e.* segmentation statistique par cohésion lexicale, utilisation de marqueurs linguistiques, utilisation conjointe de ces indices),
- la tâche accomplie : détection de frontières thématiques, segmentation textuelle, segmentation textuelle et proposition de thèmes « concrets » (*i.e.* descripteurs thématiques),
- l'utilisation (ou la non utilisation) de connaissances externes (*e.g.* réseau de cooccurrences, LSA).

Nous avons constaté un fort déséquilibre entre le nombre de systèmes statistiques quantitatifs par rapport aux systèmes linguistiques, parce que les méthodes quantitatives sont plus facilement applicables (elles nécessitent de faibles prétraitements et s'appliquent à tous types de documents). Néanmoins, nous pensons, avec (Bilhaut, 2006), que l'apport des approches linguistiques est nécessaire à l'amélioration des systèmes de détection automatique de thèmes.

Au vu des résultats obtenus lors d'évaluations menées entre autres par (Choi *et al.*, 2001 ; Sitbon et Bellot, 2004 ; Piérard et Bestgen, 2006a, 2006b), nous choisissons d'utiliser l'algorithme de segmentation statistique par cohésion lexicale *C99* dans sa version améliorée avec la LSA (*i.e.* CWM) pour segmenter thématiquement les documents. L'algorithme étant disponible librement et sa récente adaptation au français (Sitbon et Bellot, 2004) en font des atouts qui, dans le cadre de notre projet, ne sont pas à négliger. De plus, la méthode utilisée est indépendante de la langue et du domaine, ce que nous recherchons pour mettre en place notre système.

Aussi, à la lumière des études linguistiques présentées, les marqueurs linguistiques que nous souhaitons identifier sont les cadres de discours et les chaînes de référence, ainsi que des marqueurs extralinguistiques (tels que la position dans la phrase ou le genre textuel), que nous mettons en relation afin de déterminer les thèmes des documents. En effet, l'efficacité de l'accumulation d'indices et leur mise en relation a été démontrée à maintes reprises (Piérard et Bestgen, 2005b, 2006a ; Bestgen, 2012) et nous choisissons de suivre cette voie. Nous adoptons

donc pour notre projet, dans la lignée de (Hernandez, 2004), une méthode hybride mêlant segmentation statistique et identification de marqueurs linguistiques, afin de détecter automatiquement les thèmes des documents et d'en proposer des descripteurs concrets.

Dans le chapitre suivant, nous nous penchons sur les deux grands types de méthodes de calcul automatique de la référence permettant d'identifier les différentes reprises d'un référent dans le discours, afin de positionner notre approche dans ce vaste domaine.

Chapitre 5

Systèmes de résolution de la référence

La résolution de la référence[1] consiste à identifier automatiquement dans un document les expressions référentielles qui réfèrent à la même entité du discours (personne, événement, entité abstraite). Il s'agit d'une tâche complexe, en linguistique comme en TAL (Mitkov, 1999 ; Stoyanov *et al.*, 2010), en raison des phénomènes entrant en jeu pour identifier les diverses mentions d'un même référent : « coreference resolution has long been recognized as a difficult task » (Zheng *et al.*, 2011 : 1114). La résolution de la référence demeure un problème ouvert en TAL (Chaumartin, 2007)[2] et la qualité des systèmes automatiques développés à l'heure actuelle pour la recherche d'information (McCarthy et Lehnert, 1995), le résumé automatique (Bergler *et al.*, 2003 ; Steinberger *et al.*, 2007), ou l'indexation des documents[3] dépend, entre autres, de leur capacité à

[1] (Popescu-Belis, 2000) distingue la *résolution de la référence* - qui consiste à trouver dans le texte toutes les expressions référentielles se rapportant au même référent – de la *résolution de la coréférence* – qui cherche à déterminer les liens établis entre deux expressions référentielles coréférentes. (Mitkov, 2000 ; Mitkov *et al.*, 2012) distinguent, de leur côté, la résolution de la coréférence de la *résolution anaphorique* qui peut comporter des cas où l'anaphore et son antécédent ne coréfèrent pas. Nous emploierons, pour notre part, l'expression « résolution de la référence », pour rester cohérente par rapport à la notion de *chaîne de référence* que nous avons adoptée avec (Schnedecker, 1997).

[2] Dans l'appel à communications de la journée ATALA « la résolution des anaphores en TAL » du 16 juin 2007, la tâche de résolution des anaphores est considérée comme un « verrou pour le TAL ».

[3] De nombreuses autres applications du TAL utilisent la résolution de la coréférence pour des tâches telles que la détection des opinions (Nicolov *et al.*, 2008 ; Hendrickx et Hoste, 2009), les

identifier les diverses mentions d'un référent au fil du texte (Weissenbacher et Nazarenko, 2007 ; Mitkov *et al.*, 2012 ; Muzerelle *et al.*, 2013).

Deux grandes familles de méthodes de résolution de la référence se sont succédées depuis plusieurs décennies :

- d'une part, les méthodes symboliques (ou *knowledge-based approaches*), utilisant des règles ou des heuristiques[4] pour identifier les antécédents d'une anaphore donnée,
- d'autre part, les méthodes par apprentissage statistique (ou *corpus-based approaches*), utilisant les informations contenues dans les corpus pour regrouper des paires antécédents-anaphores potentielles.

Dans ce chapitre, nous proposons une vue de plusieurs systèmes de calcul de la référence issus de ces deux familles de méthodes (sections 1 et 2). La section 3 sera l'occasion d'identifier certaines lacunes rencontrées par ces divers systèmes et nous permettra de positionner notre approche.

systèmes de questions-réponses (Vicedo et Ferrandez, 2006), la traduction automatique, l'extraction des événements et leurs informations associées (Chambers et Jurafsky, 2008, 2011 ; Ludovic, 2011), etc.

[4] Une heuristique est une méthode calculatoire qui tient compte des résultats précédents afin de déduire la stratégie à adopter pour fournir rapidement une solution (qui peut ne pas être optimale) face à un problème donné.

1 Systèmes symboliques

Les systèmes symboliques reposent sur l'application d'une série de règles ou de patrons contextuels issus d'analyses linguistiques pour sélectionner un antécédent potentiel pour une anaphore donnée. Ces règles syntaxiques, morpho-syntaxiques et/ou sémantiques sont souvent hiérarchisées ou sont soumises à certaines contraintes pour éviter leur recouvrement. Pour la plupart, ces règles reposent sur les théories psycholinguistiques relatives à l'activation d'une entité dans le discours (*e.g.* théorie de l'Accessibilité d'(Ariel, 1990), théorie du centrage (Grosz *et al.*, 1995), *cf.* chapitre 2).

Mais, bien que ces systèmes symboliques soient dépendants du corpus ayant servi au développement des règles, ils sont néanmoins doués d'une grande flexibilité permettant l'ajout et la modification rapide de règles (Barbu et Mitkov, 2001 ; Nouioua, 2007).

Dans cette section, nous présentons les premiers systèmes symboliques développés pour la résolution de la référence (essentiellement pour l'anglais) puis les approches à base de faibles connaissances (section 2) avant de nous focaliser sur les systèmes proposés pour le français (section 3).

1.1 Premières approches

C'est dans les années 70 qu'apparaissent les premières approches symboliques pour la résolution de la référence. Elles nécessitent en général une analyse syntaxique fine des textes, afin de filtrer le bon antécédent pour une anaphore donnée. Ce besoin accru en informations syntaxiques leur a valu d'être nommées méthodes à base de fortes connaissances ou *knowledge-intensive* (Nøklestad, 2009), relativement aux méthodes à base de faibles connaissances ou *knowledge-poor* (*cf.* section 1.2).

1.1.1 L'approche syntaxique « naïve » de (Hobbs, 1978)

Hobbs figure parmi les premiers[5] à avoir proposé un système symbolique performant conçu pour la résolution des anaphores pronominales en anglais. Son système reposait alors sur l'utilisation d'une analyse syntaxique complète de phrases annotées manuellement. Qualifié de « naïf » en raison de son approche purement syntaxique, ce système demeure encore une référence pour évaluer les performances de systèmes plus récents (Mitkov, 1999 ; Weissenbacher, 2008 ; Nand, 2012).

L'algorithme de (Hobbs, 1978) résout les anaphores pronominales (*i.e.* les pronoms *he*, *she*, *it* et *they*) et les possessifs, mais il ne gère pas les emplois impersonnels du pronom *it* ni les réfléchis. Il présuppose une analyse syntaxique parfaite du corpus sous la forme d'un arbre syntaxique (voir Figure 22). Cet arbre syntaxique doit correspondre à la structure grammaticale de la phrase.

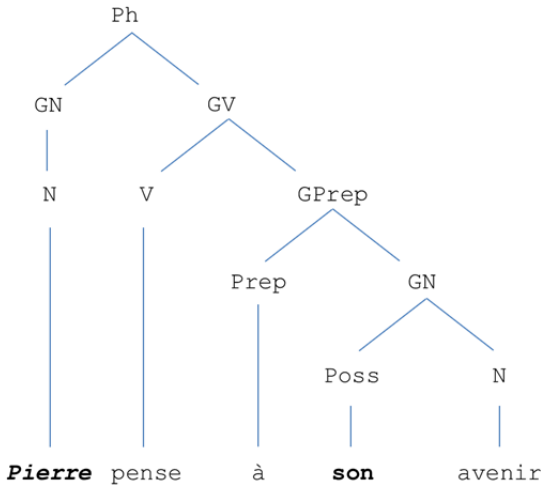


Figure 22 - Arbre syntaxique pour la phrase "Pierre pense à son avenir"

Lorsqu'un pronom est identifié dans une phrase, l'algorithme parcourt l'arbre syntaxique à la recherche d'antécédents. Le sens de la recherche dans l'arbre s'effectue d'abord de gauche à droite, en largeur, en appliquant une préférence pour l'antécédent le plus proche de l'anaphore (contrainte intraphrastique). Si aucun antécédent n'est trouvé dans la phrase en cours, les arbres syntaxiques des

[5] (Winograd, 1972) s'était déjà préoccupé de la résolution des anaphores en proposant un système de dialogue homme-machine dans un micro-monde fermé composé d'un ensemble de blocs de couleurs, formes et tailles variées. L'utilisateur demandait alors au robot d'effectuer des tâches telles que « prends **le carré rouge** et mets-**le** dans le coin gauche ». Le système était alors capable de résoudre l'anaphore pronominale pour effectuer la tâche demandée (*cf.* (Victorri, 2005) pour plus de détails sur cette approche).

phrases précédentes sont parcourus, en commençant toujours par la phrase précédente la plus proche. Une recherche en largeur est aussi menée au niveau interphrastique, en appliquant une préférence pour les candidats antécédents sujets.

Une fois les antécédents identifiés, l'algorithme leur applique des contraintes morphologiques (accord en genre et en nombre avec l'anaphore) et des contraintes syntaxiques (basées sur le principe B[6] de la théorie du liage (*Government and Binding*, (Chomsky, 1981, 1982, 1986)). Les contraintes syntaxiques permettent ainsi de vérifier que :

- un pronom (non réfléchi) et son antécédent n'apparaissent pas dans la même phrase simple (*e.g.* dans « **Paul** *le* réconforte », les expressions référentielles ne peuvent pas être coréférentes),
- l'antécédent d'un pronom (non réfléchi) précède ou c-commande[7] le pronom (*e.g.* dans « Paul likes him », « Paul » n'est pas situé au même niveau dans l'arbre syntaxique que « him », donc les deux expressions référentielles ne sont pas coréférentes).

Evalué manuellement sur 300 pronoms issus d'un corpus composé de trois genres textuels (journaux, roman, ouvrage d'archéologie), le système a obtenu des performances de 88,3%, allant même jusqu'à 91,7% en ajoutant une sélection de contraintes à l'algorithme initial. Mais, bien que ces performances soient élevées, (Hobbs, 1978) remarque que dans la moitié des cas du corpus, il n'y avait pas de compétition entre antécédents potentiels (*i.e.* il n'y avait qu'un seul candidat). L'auteur relève aussi que 90% des anaphores pronominales trouvent leur antécédent dans la phrase où elles se situent et 98% dans la phrase en cours ou la phrase précédente.

1.1.2 Le modèle du contexte d'(Alshawi, 1987)

(Alshawi, 1987) a proposé une approche plus cognitive pour résoudre la référence pronominale : le *modèle du contexte*. Inspiré notamment de la théorie du centrage[8], le modèle représente la saillance des entités du discours par une valeur

[6] Le principe B de la théorie du liage indique qu'un pronom peut avoir un antécédent si ce dernier n'appartient pas au domaine local du pronom (*i.e.* la plus petite proposition qui contient l'antécédent) ou qu'il ne le *c-commande* pas (voir note 7). Par exemple, la phrase « Marie pense que Paul l'aime » obéit à ce principe : le pronom « l' » est suffisamment éloigné de l'antécédent « Marie », donc « Marie » est l'antécédent de « l' ».

[7] La notion de *c-commande* (pour *constituent-command*) est une relation entre nœuds dans les arbres syntaxiques. Elle se définit de la manière suivante : un nœud A commande un nœud B si A ne domine pas B et B ne domine pas A et si le premier nœud branchant dominant A domine aussi B (*i.e.* si A et B sont au même niveau dans l'arbre syntaxique).

[8] *Cf.* chapitre 2, section 2.2.

numérique. Cette valeur numérique mesure la saillance des entités (nommée *activation contextuelle*) en calculant la somme des poids affectés à divers facteurs tels que la récence de l'entité dans la phrase ou dans le paragraphe, l'emphase (*i.e.* la répétition), la deixis, la fonction syntaxique[9]. Ainsi, la saillance de chacune des entités est recalculée au fil du texte, ce qui permet de sélectionner les entités les plus présentes dans la mémoire (*i.e.* des candidates à la reprise anaphorique). D'après (Victorri, 2005 : 142), l'approche d'(Alshawi, 1987) est, en ce sens, doublement originale :

> « Ce qui fait l'originalité de l'approche d'Alshawi, c'est d'une part l'utilisation de valeurs numériques, et d'autre part le fait que l'analyse n'est plus centrée sur les transitions d'une phrase à la suivante : elle prend en considération toutes les entités qui sont dans le contexte du discours et elle affecte à chacune d'elles une saillance qui se modifie tout au long du texte, que l'entité soit ou non évoquée par la phrase en cours. On ordonne ainsi, à chaque moment de la lecture, l'ensemble des entités du contexte du discours. »

Dans le *modèle du contexte* intervient un mécanisme central de persistance-dégradation (Dupont, 2002, 2003). En effet, une fonction de dégradation est affectée à chaque facteur contextuel à mesure que l'on progresse dans le texte. De ce fait, si l'entité n'est plus évoquée dans le texte, sa saillance va diminuer progressivement jusqu'à atteindre 0. En revanche, si l'entité fait l'objet d'une reprise dans la phrase suivante, sa saillance va être augmentée du score correspondant au type de la reprise (*e.g.* pronom, possessif). Ainsi, au plus une entité fait l'objet de reprises, au plus son score de saillance sera élevé.

Afin d'illustrer son modèle, Alshawi fournit de nombreux exemples (dans ses annexes B). Nous reproduisons ci-dessous une partie de l'exemple du texte A14 ainsi que l'analyse proposée menant au choix de l'antécédent pour la phrase « It is made by Plexir. » (Alshawi, 1987 : 174) :

[9] « The model of context […] provides a computational mechanism for gradually accumulating, and combining, the influence of different kind of information contributing to context (such as recency of mention, subject area, and syntactic marking of focus) as the processing of a text progresses. » (Alshawi, 1987 : 6).

```
Example text A14
…
P1010 is a terminal that is supplied by Clark.
P9000 is a green printer.
It is made by Plexir.
…

-- "It is made by Plexir";
'P9000' preferred to 'P1010' as referent for "It".
Differences:  recency 116;  emphasis  44;  association  -26;
              deixis -7; subject-area 67; processing 6.
```

Figure 23 - Exemple de choix de l'antécédent suivant le modèle du contexte d'(Alshawi, 1987)

Dans cet exemple, les deux candidats antécédents pour l'anaphore « It » sont « P1010 » et « P9000 ». Le candidat « P9000 » est retenu, au regard des scores obtenus pour la récence (116), la fonction sujet (67) et l'emphase (44). Bien que (Landragin, 2004, 2005) remette en cause les scores attribués aux différents facteurs de saillance (*e.g.* un trop fort poids attribué à la récence de l'entité) de ce modèle, il en souligne l'intérêt de par le nombre de facteurs linguistiques pris en compte pour calculer le poids de chaque entité au fil du texte.

Fort de son succès, le *modèle du contexte* a été implémenté par divers auteurs, notamment par (Lappin et Leass, 1994).

1.1.3 L'approche heuristique de (Lappin et Leass, 1994)

Afin de résoudre les anaphores pronominales de troisième personne (« he », « she », « they », « it »), les réfléchis (« himself », « herself ») et les réciproques (« each other », « one another ») en anglais, (Lappin et Leass, 1994) ont proposé une approche heuristique utilisant des mesures de saillance issues d'une analyse syntaxique profonde calculée par la *Slot Grammar* de (McCord *et al.*, 1992)[10] ainsi qu'un modèle dynamique d'état attentionnel (basé sur la théorie du centrage).

A partir d'une analyse syntaxique complète, l'algorithme RAP (*Resolution of Anaphora Procedure*) utilise plusieurs modules et filtres permettant de trier les antécédents potentiels pour une anaphore donnée :

[10] La *Slot Grammar* est un système grammatical en dépendances basé sur des règles établissant les relations syntaxiques existantes entre les divers constituants d'une phrase (*i.e.* les *slot*, comme par exemple sujet, objet direct, etc.).

- un module élimine les emplois impersonnels des pronoms « it » *via* des listes de verbes et d'adjectifs situés dans des structures telles que « it is possible that », « it is easy to », « it is time to » ;
- un filtre syntaxique exclut la coréférence intraphrastique établie entre un pronom et un groupe nominal (suivant les mêmes principes de la théorie du liage utilisés par Hobbs). Par exemple, dans « *The woman* said that *he* is funny » (Lappin et Leass, 1994 : 538), le groupe nominal « the woman » ne peut pas être l'antécédent du pronom « he » car le pronom est un argument de « be » (et « the woman » et « he » discordent morphologiquement) ;
- un filtre morphologique écarte les dépendances anaphoriques entre un pronom et un groupe nominal s'ils ne possèdent pas le même genre, le même nombre et la même personne ;
- un module de calcul dynamique de la saillance des antécédents potentiels, inspiré du modèle du contexte d'(Alshawi, 1987), classe les candidats restants suivant plusieurs facteurs (récence phrastique, fonction sujet, tête lexicale, etc.). Chaque facteur de saillance est affecté d'un poids initial suivant l'importance de ce facteur dans la procédure de résolution de la référence (voir Tableau 18). Ainsi, chacun des facteurs permet d'augmenter le score de saillance d'un antécédent potentiel.

FACTEUR	SCORE INITIAL
récence phrastique	100
emphase sur le sujet	80
emphase existentielle[11]	70
objet direct	50
objet indirect	40
tête du groupe nominal[12]	80
emphase non adverbiale[13]	50

Tableau 18 – Poids initiaux affectés aux différents facteurs de saillance, d'après (Lappin et Leass, 1994)

A chaque nouvelle phrase, un processus de dégradation des facteurs de saillance divise leur score par deux (s'ils étaient « activés » dans les

[11] Si le candidat est un prédicat nominal dans une construction existentielle, par exemple : « There are only *a few restrictions* on LQL query construction for WordSmith. » (Lappin et Leass, 1994 : 540), sa saillance augmente de 70.

[12] Si le candidat est un groupe nominal non inclus dans un autre groupe nominal (*e.g.* « le fils » dans « le fils de Jean »), sa saillance augmente de 80.

[13] Si le candidat nominal n'est pas inclus dans un groupe adverbial délimité par un séparateur alors son score est 50. Le groupe nominal (en italique) de l'exemple suivant ne bénéficie pas de ce facteur de saillance : « In *the Panel definition panel,* select the ''Specify'' option from the action bar » (Lappin et Leass, 1994 : 541).

phrases précédentes). Les facteurs dont le score atteint 0 sont alors éliminés ;

- un module de calcul de saillance de « classes d'équivalence » identifie les candidats antécédents coréférents. Chaque classe d'équivalence se voit affecter un score de saillance correspondant à la somme des facteurs de saillance des éléments de la classe. Dans une classe d'équivalence, le candidat dont le score de saillance est le plus élevé est choisi comme antécédent d'un pronom. Lorsque plusieurs antécédents potentiels obtiennent le même score de saillance, le candidat retenu est celui situé le plus proche de l'anaphore.

Aussi, les antécédents intraphrastiques sont préférés aux interphrastiques. Cette préférence est réalisée suivant trois mécanismes :

– un score de saillance supplémentaire est attribué aux candidats présents dans la phrase en cours,

– la saillance des candidats antécédents situés dans les phrases précédentes est diminuée relativement aux valeurs de saillance des candidats situés dans la phrase en cours,

– la proximité est utilisée pour résoudre les liens entre candidats antécédents possédant des valeurs de saillance égales.

Spécifiquement pour les pronoms de 3ème personne, deux contraintes vont modifier le score de saillance des candidats anaphoriques. En effet, si le candidat antécédent suit le pronom (*i.e.* cataphore[14]), une lourde pénalité va être affectée à son score de saillance (-175). En revanche, si le candidat antécédent et le pronom ont le même rôle grammatical (*i.e.* parallélisme syntaxique), alors le score de saillance de l'antécédent est augmenté (+35).

(Lappin et Leass, 1994) ont évalué leur algorithme sur des manuels techniques représentant 360 occurrences pronominales. La reconnaissance des antécédents a atteint 86% (74% en interphrastique et 89% en intraphrastique). Les auteurs ont aussi comparé leur système à celui proposé par Hobbs (après l'avoir adapté) et ils ont obtenu des performances supérieures de 4% par rapport aux performances de leur prédécesseur. Ces résultats, bien que performants, sont toutefois à nuancer. En effet, leur système traite plus efficacement les anaphores intraphrastiques, qui représentaient 80% des pronoms du corpus d'évaluation. Les auteurs ont aussi testé l'influence des différents facteurs de leur algorithme sur le traitement des

[14] La cataphore est un mécanisme d'annonce d'un élément ultérieur : l'antécédent est situé après la source. Par exemple, dans « Il est gentil, Paul », l'antécédent « Paul » est situé après le pronom cataphorique « il ».

anaphores pronominales, en les supprimant : par exemple, sans la combinaison de la dégradation de la saillance et la fonction grammaticale, les performances chutent de manière significative (respectivement 59% et 64%).

1.1.4 Bilan

Nous venons de voir que les premières approches symboliques, axées essentiellement sur la résolution des anaphores pronominales, nécessitent en amont une analyse syntaxique complète des textes ainsi que des connaissances sur le contexte. A partir d'une analyse syntaxique fine, les algorithmes utilisent une série de contraintes pour trier les antécédents potentiels pour une anaphore donnée. Certaines approches affectent un poids arbitraire à chacune des contraintes afin de départager les candidats *via* leur saillance respective, calculée de manière dynamique au fil du texte.

Ces approches, bien qu'obtenant des performances élevées, se révèlent coûteuses et complexes à mettre en œuvre et cela, à plusieurs titres :

- premièrement, vu l'apport nécessaire en ressources d'entrée (*i.e.* analyse syntaxique parfaite et complète du texte, identification des influences contextuelles sur chacun des facteurs de saillance),
- deuxièmement, vu l'arbitrarité des poids affectés à chacun des facteurs de saillance. En effet, ces poids seraient à confirmer *via* l'étude de plusieurs corpus issus de genres textuels divers (Landragin, 2004), notamment l'avantage attribué automatiquement à la récence phrastique,
- troisièmement, vu la dépendance des règles symboliques développées sur le genre du corpus utilisé. Par exemple, (Qiu *et al.*, 2004) ont implémenté et évalué l'algorithme de (Lappin et Leass, 1994) sur le corpus MUC-6[15] (contenant des articles du Wall Street Journal) et ont obtenu des performances nettement plus basses (57,9%).

Pour pallier ces difficultés, les approches robustes à base de faibles connaissances ont été proposées, notamment par (Mitkov, 1998 ; Kennedy et Boguraev, 1996).

[15] *Message Understanding Conference* (voir chapitre 8).

1.2 Approches *knowledge-poor*

Les approches proposées par la suite ont cherché à simplifier et à diminuer les ressources linguistiques utilisées par les algorithmes pour résoudre la référence. Ces approches à base d'indices de surface sont ainsi qualifiées de *knowledge-poor approaches* par (Mitkov, 1998) ou de *lightweight approaches* par (Bontcheva *et al.*, 2002). Nous présentons ci-après une sélection de systèmes largement cités dans la littérature.

1.2.1 L'approche par facteurs de (Kennedy et Boguraev, 1996)

(Kennedy et Boguraev, 1996) ont proposé un algorithme à base de faibles connaissances pour résoudre les anaphores pronominales. Cet algorithme est la version étendue et modifiée de l'approche de (Lappin et Leass, 1994). En effet, à la différence de l'algorithme de (Lappin et Leass, 1994), cet algorithme nécessite seulement en entrée un texte étiqueté morpho-syntaxiquement ainsi que des annotations relatives à la fonction grammaticale des entités lexicales. L'étiqueteur choisi par (Kennedy et Boguraev, 1996) est celui de (Voutilainen et Heikkilä, 1992) car il obtient des performances élevées (97,6%) et il traite des corpus issus de genres textuels différents. Ce dernier élément est important, dans la mesure où (Kennedy et Boguraev, 1996) ont pour objectif final de fournir un outil robuste, donc indépendant du domaine.

Une fois le corpus étiqueté, trois séries de patrons (*i.e.* des règles contextuelles) permettent d'obtenir des annotations supplémentaires sur les groupes nominaux simples, les groupes prépositionnels et adjectivaux ainsi que les emplois impersonnels du pronom « it » (lorsque « it » est le sujet de verbes tels que « seem », « appear », etc.). A l'issue de ces prétraitements, l'algorithme de (Kennedy et Boguraev, 1996) suit la logique de l'algorithme de (Lappin et Leass, 1994). Le texte est parcouru de gauche à droite en sélectionnant les antécédents potentiels pour une anaphore donnée *via* les filtres syntaxiques. Les candidats restants sont ensuite départagés suivant un score de saillance calculé à partir de facteurs contextuels, grammaticaux et syntaxiques. Chacun des facteurs possède un score arbitraire (voir Tableau 19). Parmi les 10 facteurs utilisés, deux facteurs (POSS-S et CNTX-S) ne sont pas communs avec ceux de (Lappin et Leass, 1994). Ces facteurs permettent de prendre en compte les possessifs et le contexte dans lequel un référent du discours apparaît (*i.e.* le paragraphe). Un autre point de divergence entre les deux systèmes concerne la détermination de la référence

disjointe (condition B de la théorie du liage, *e.g.* dans « **she** likes *her* », les deux pronoms ne peuvent pas être coréférents). En effet, pour traiter ces cas avec peu de connaissances, l'algorithme de (Kennedy et Boguraev, 1996) s'appuie sur des inférences sur les fonctions grammaticales des entités ainsi que sur la position des entités dans la phrase.

SENT-S:	100	iff[16] in the current sentence
CNTX-S:	50	iff in the current context
SUBJ-S:	80	iff GFUN = *subject*
EXST-S:	70	iff in an existential construction
POSS-S:	65	iff GFUN = *possessive*
ACC-S:	50	iff GFUN = *direct object*
DAT-S:	40	iff GFUN = *indirect object*
OBLQ-S:	30	iff the complement of a preposition
HEAD-S:	80	iff EMBED = NIL
ARG-S:	50	iff ADJUNCT = NIL

Tableau 19 – Facteurs de saillance et poids associés, d'après (Kennedy et Boguraev, 1996)

Le score de saillance des candidats est encore évalué et modifié (la cataphore est pénalisée, le parallélisme de fonction syntaxique ainsi que les antécédents intraphrastiques sont valorisés), puis le candidat antécédent retenu est celui dont le score de saillance est le plus élevé (ou celui situé le plus proche de l'anaphore en cas d'égalité de score entre deux antécédents potentiels).

L'algorithme, testé sur 27 textes issus de genres divers (brèves, pages web, publicités, reportages), obtient des performances de 75%. Les principales erreurs relevées concernent des informations incomplètes fournies par l'étiqueteur sur le genre de certains mots (35% d'erreurs) ainsi que des difficultés pour traiter des passages entre guillemets (14% d'erreurs). Comparativement au système développé par (Lappin et Leass, 1994), les performances obtenues sont moins élevées (10% de moins), mais cette approche à base de faibles connaissances a été évaluée sur des textes issus de genres textuels différents, ce qui n'est pas le cas de Lappin et Leass (pour rappel, ils ont uniquement évalué leur système sur des manuels informatiques). (Mitkov, 1999 : 16) souligne la valeur ajoutée de la couverture du système proposé par (Kennedy et Boguraev, 1996) sur les performances obtenues : « Evaluation reports 75% accuracy but this has to be given a "bonus" for this results span a very wide coverage : the evaluation was based on a random selection of genres. »

[16] Si et seulement si.

1.2.2 L'approche à haute précision de (Baldwin, 1997)

De son côté, (Baldwin, 1997) examine la possibilité de ne résoudre que les anaphores non ambiguës, favorisant ainsi une meilleure précision des résultats. Il remarque que les antécédents se situent habituellement dans la phrase courante ou la phrase précédente.

Son système, appelé CogNIAC, requiert peu de connaissances : un découpage du texte en phrases, un étiquetage morpho-syntaxique du texte, la reconnaissance des groupes nominaux simples et des informations de base relatives au genre et au nombre.

CogNIAC comporte six règles de base appliquée dans l'ordre suivant :

- *Unique in Discourse* : s'il n'y a qu'un seul candidat antécédent possible dans le texte, alors ce candidat est l'antécédent,
- *Reflexive* : si l'anaphore est un réfléchi, choisir l'antécédent potentiel situé le plus proche dans la phrase,
- *Unique in Current + Prior* : s'il n'existe qu'un seul candidat antécédent dans la phrase précédente et la phrase en cours, alors ce candidat est l'antécédent,
- *Possessive Pro* : si l'anaphore est un possessif et qu'un même possessif est situé dans la phrase précédente, alors ce candidat est l'antécédent,
- *Unique Current Sentence* : s'il n'existe qu'un seul antécédent potentiel dans la phrase en cours, alors ce candidat est l'antécédent,
- *Unique Subject/ Subject Pronoun* : si le sujet de la phrase précédente ne contient qu'un seul candidat antécédent et que l'anaphore est le sujet de la phrase en cours, alors ce candidat est l'antécédent.

Lors de l'utilisation du système, la résolution des anaphores s'effectue suivant l'ordre linéaire du texte. Pour chaque candidat, les règles sont appliquées dans l'ordre. Si un candidat valide une règle, aucune autre règle n'est testée (car l'anaphore est résolue). Si aucun candidat ne valide de règle pour une anaphore donnée, l'anaphore n'est pas résolue.

CogNIAC a été évalué une première fois sur 198 pronoms issus d'un corpus d'articles de journaux. Les performances obtenues sont de 74,1% : la précision[17]

[17] La précision est une mesure de qualité (voir chapitre 8 pour le calcul de cette mesure) permettant de savoir si l'étiquette est attribuée correctement à l'élément. Dans le cas de la

est très élevée (97%) mais le rappel[18] est plus faible (60%). En d'autres termes, 60% des anaphores sont résolues et parmi ces anaphores, 97% d'entre elles sont bien rattachées à l'antécédent adéquat.

Afin de résoudre « toutes » les anaphores, (Baldwin, 1997) a ajouté deux règles supplémentaires à CogNIAC :

- *Cb-Picking* : dérivée de la théorie du centrage[19], cette règle permet de résoudre des cas non pris en compte par la règle *Unique Subject/ Subject Pronoun*. S'il existe un centre rétroactif (*i.e.* le thème de l'énoncé) dans le paragraphe et qu'il constitue aussi un antécédent potentiel, alors ce candidat est l'antécédent,
- *Pick Most Recent* : choisir l'antécédent potentiel le plus récent comme antécédent de l'anaphore.

Avec l'ajout de ces deux règles, CogNIAC a perdu en précision mais a gagné en rappel, obtenant des performances de 77,9% sur un corpus composé de trois histoires contenant 298 pronoms de 3ème personne.

Le système a encore été modifié dans le cadre de la campagne d'évaluation MUC-6 (*i.e.* suppression des deux dernières règles et de la règle *Possessive Pro*, ajout d'une règle sélectionnant le sujet de la proposition, traitement des pronoms de 1ère personne dans des discours rapportés, détection des emplois impersonnels du pronom « it ») afin de pouvoir identifier de nouvelles entités en relation de coréférence dans le corpus de test (*i.e.* anaphores nominales et noms propres). Evalué sur 15 textes issus du Wall Street Journal, CogNIAC a obtenu un rappel de 75% et une précision de 73%. Mais, parmi les erreurs observées, seules 25% sont directement liées aux règles de CogNIAC.

1.2.3 L'approche *knowledge-poor* de (Mitkov, 1998)

L'approche robuste à base de faibles connaissances de (Mitkov, 1998) est l'une des plus populaires. L'auteur propose un système de résolution des anaphores (appelé MARS) dans des manuels techniques (informatique), applicable à plusieurs langues (anglais, polonais, arabe) et nécessitant le moins d'informations

résolution d'anaphore, l'objectif est de savoir si le pronom anaphorique a bien été rattaché à l'antécédent adéquat.

[18] Le rappel est une mesure de quantité (*cf.* chapitre 8). Cette mesure permet de savoir combien d'antécédents ont bien été identifiés.

[19] *Cf.* chapitre 2, section 2.2.

linguistiques possibles afin d'en faciliter l'implémentation. En effet, pour seuls prétraitements, le système requiert un étiquetage morpho-syntaxique du texte ainsi qu'un balisage des groupes nominaux simples.

L'algorithme applique une série de 10 heuristiques préférentielles basées sur des données empiriques (saillance, répétition lexicale, distance référentielle), appelées « antecedent indicators », pour sélectionner les antécédents potentiels pour une anaphore donnée. Les candidats antécédents sont recherchés parmi les groupes nominaux situés dans une fenêtre de trois phrases (*i.e.* la phrase dans laquelle apparaît l'anaphore à résoudre et les deux phrases précédentes[20]). Puis, l'accord en genre et en nombre entre l'anaphore et les candidats est vérifié[21]. Une fois cette étape accomplie, les heuristiques préférentielles sont alors appliquées : un score de saillance (allant de -1 à 2) est attribué à chaque candidat antécédent suivant sa compatibilité ou son incompatibilité avec chacune des heuristiques. La liste des heuristiques de MARS est la suivante :

- *definiteness* : si le candidat antécédent est un groupe nominal défini, alors son score est 0, si le candidat est un groupe nominal indéfini, alors son score est -1,
- *givenness* : si le candidat antécédent est le thème de la phrase précédente (*i.e.* le premier groupe nominal de la phrase) alors son score est 1, sinon son score est 0,
- *indicating verbs* : si le candidat antécédent suit un verbe « indicateur »[22], c'est-à-dire un verbe qui met en évidence sa saillance (*e.g. discuss, present, illustrate, identify, summarise, examine, describe, define, show, check, develop, review, report, outline, consider, investigate, explore, assess, analyse, synthesise, study, survey, deal, cover*), alors son score est 1, sinon son score est 0,
- *lexical reiteration* : si le candidat est répété[23] deux fois ou plus dans le paragraphe, son score est 2, s'il est répété une fois, son score est 1, sinon, son score est 0,
- *section heading preference* : si le candidat est situé dans le titre d'une section, alors son score est 1, sinon son score est 0,

[20] Les cas de cataphore ne sont pas traités. Les emplois impersonnels du pronom « it » sont éliminés *via* un filtre référentiel.

[21] Une liste d'exceptions a été mise en place par (Mitkov, 1998) car la compatibilité genre/nombre ne s'applique pas pour certains termes tels que « data » qui peut être repris par « it », bien qu'étant au pluriel.

[22] D'après (Mitkov, 1998 : 870), « Empirical evidence suggests that because of the salience of the noun phrases which follow them, the verbs listed above are particularly good indicators. »

[23] La répétition comprend la répétition de la tête lexicale (*e.g.* « the computer manual »… « the manual ») mais aussi les synonymes.

- *« non-prepositional » noun phrases* : si le candidat n'est pas inclus dans un groupe prépositionnel (*i.e.* est un complément d'objet indirect), son score est 0, sinon son score est -1,
- *collocation pattern preference* : si le candidat apparaît dans une construction identique au pronom (*e.g.* « Press *the key* down…press *it* again »), alors son score est 2, sinon son score est 0,
- *immediate reference* : si le candidat antécédent apparaît dans une construction fréquente des manuels techniques, où un premier verbe est suivi d'un groupe nominal, suivi d'une conjonction (*and/or/before/after*), suivie d'un deuxième verbe, suivi de l'anaphore (*e.g.* « press *the Power button* and hold *it* down »), alors son score est 2, sinon son score est 0,
- *referential distance* :
    - dans une phrase simple, si le candidat est situé dans la phrase précédente, son score est 1, s'il est situé 2 phrases avant, son score est 0 et s'il est situé 3 phrases avant, son score est -1,
    - dans une phrase complexe, si le candidat est situé dans la proposition précédent l'anaphore, son score est 2, s'il est situé dans la phrase précédente, son score est 1, s'il est situé 2 phrases avant, son score est 0 et s'il est situé 3 phrases avant, son score est -1,
- *term preference* : si le candidat appartient au domaine couvert par le texte (*e.g.* les manuels techniques), alors son score est 1, sinon son score est 0.

Le candidat dont le score total est le plus élevé est alors retenu comme antécédent. En cas d'égalité de score entre deux candidats, celui dont le score pour l'heuristique « immediate reference » est le plus élevé est préféré. Si cela ne permet toujours pas de départager les candidats, ce sont les heuristiques « collocation pattern » puis « indicating verbs » qui seront utilisées. Si les candidats sont toujours *ex-aequo*, le candidat choisi est le plus récent.

(Mitkov, 1998) a évalué MARS sur un corpus de textes issus de manuels techniques en anglais comportant 294 pronoms. Les performances obtenues sont élevées : 89,7%. Le système a obtenu des performances comparables à celles de (Hobbs, 1978), mais avec beaucoup moins de connaissances linguistiques. Le système MARS a aussi été adapté et évalué pour le polonais (performances de 93,3%) et l'arabe (performances de 95,2%), toujours sur des textes issus du même genre textuel. Néanmoins, Mitkov reconnaît avoir effectué l'évaluation de son système après avoir corrigé des erreurs d'étiquetage dans son corpus. Les résultats annoncés semblent donc difficilement comparables avec des systèmes évalués dans

des conditions réelles (*e.g.* avec un corpus étiqueté automatiquement sans corrections manuelles).

1.2.4 L'approche de (Bontcheva *et al.*, 2002)

De leur côté, (Bontcheva *et al.*, 2002) ont proposé deux modules à base de faibles connaissances pour traiter les anaphores et la coréférence orthographique entre entités nommées (*i.e.* noms de personnes (« Barack Obama »), noms d'organisations (« Université de Strasbourg »), noms de lieux (« Strasbourg »), unités monétaires (« euros », voir chapitre 6). Les modules ont été intégrés au système ANNIE (*A Nearly-New Information Extraction system*) lui-même inclus dans la plate-forme GATE (*General Architecture for Text Engineering*) développée par (Cunningham, 2002 ; Maynard *et al.*, 2002).

ANNIE comprend une série de modules (utilisables individuellement ou non) permettant d'effectuer divers prétraitements au texte : découper le texte en *tokens* (*i.e.* mots ou signes de ponctuation), découper le texte en phrases, étiqueter le texte, identifier des noms propres ou des indices proches des noms propres *via* des listes (villes, organisations, titres, abréviations (*e.g.* « Ltd », « Co »)), annoter les différents types d'entités nommées *via* des règles contextuelles écrites à la main.

En plus des modules de base d'ANNIE, (Bontcheva *et al.*, 2002) ont ajouté le module *orthomatcher* pour détecter la coréférence orthographique entre les noms propres et un module de résolution d'anaphores pronominales lorsque les antécédents sont des entités nommées.

Le module *orthomatcher* identifie la coréférence établie entre des noms propres, par exemple « Barack Obama » et « M. Obama ». Pour ce faire, une série de règles manuelles applicables soit à tous les types d'entités, soit à un type d'entités précis ont été définies. Parmi les règles applicables à toutes les entités, on trouve :

- *exact match* : les occurrences identiques de l'entité coréfèrent,
- *equivalent* : une liste de synonymes permet d'identifier les équivalences entre entités, par exemple « New York » et « The Big Apple »,
- *possessives* : les formes possessives des entités nommées sont identifiées,
- *spurious* : une liste d'entités « parasites » permet d'identifier des entités différentes ayant un nom similaire (*e.g.* une maison mère et sa filiale).

D'autres règles ne sont applicables qu'aux entités de type organisation et personne :

- *word token match* : fait correspondre les entités si elles comportent les mêmes *tokens*, sans tenir compte de la ponctuation ou de l'ordre des mots (*e.g.* « Barack Obama » et « Obama, Barack »),
- *first/last token match* : fait correspondre le premier/dernier *token* d'une entité avec le *token* d'une autre entité, par exemple « Barack Obama » et « Barack » ou « Barack Obama » et « Obama »,
- *acronyms/abreviations* : rattache un acronyme ou une abréviation avec son entité (*e.g.* « UNICEF » et « **U**nited **N**ations **I**nternational **C**hildren's **E**mergency **F**und »), pour les organisations uniquement,
- *prepositional phrases* : fait correspondre les variantes des noms d'organisation, par exemple « University of Strasbourg » et « Strasbourg University »,
- *multi-word name matching* : faire correspondre des entités ayant un nom long avec tous les *tokens* du nom court (*e.g.* « The Coca Cola Company » et « Coca Cola »).

Le module de résolution des anaphores pronominales repose sur les prétraitements de base fournis par ANNIE ainsi que sur le module *orthomatcher*. Le module identifie, *via* des règles contextuelles, les emplois impersonnels du pronom « it » et il détecte aussi les passages en discours rapporté. Pour résoudre les anaphores pronominales, ce module suit les mêmes étapes que les autres approches utilisant des scores de saillance : 1/ identifier le contexte de l'anaphore, 2/ rechercher les candidats antécédents qui satisfont une série de contraintes, 3/ affecter un score de saillance[24] à chaque candidat suivant une série de règles, 4/ choisir le candidat antécédent ayant le plus haut score de saillance. A partir d'une étude de corpus, (Bontcheva *et al.*, 2002) ont remarqué que de simples règles de saillance permettaient de résoudre la plupart des anaphores pronominales ayant pour antécédent une entité nommée. Par exemple, 80 à 85% des pronoms de 3ème personne (*i.e.* « he », « his », « she », « her ») réfèrent à un nom de personne ayant le même genre et le même nombre et situé dans la même phrase ou dans la phrase précédente.

Les modules ont été évalués sur le corpus ACE[25], comportant des articles journalistiques, des dépêches et des fils d'actualité. L'évaluation du module *orthomatcher* a consisté à comparer les entités nommées trouvées par le système avec les entités annotées manuellement par un annotateur. Les

[24] Aucunes données ne sont fournies par les auteurs sur les poids de saillance affectés à chacun des candidats.

[25] *Automatic Content Extraction.*

performances moyennes obtenues sont élevées (94,8%). De son côté, le module de résolution des anaphores obtient des performances allant de 78% pour les pronoms de 3ème personne (« he », « her ») à 47,6% pour les pronoms « it » anaphoriques. Ces dernières performances, moins élevées, sont notamment liées à des erreurs de reconnaissance des emplois impersonnels du pronom « it » par les règles contextuelles.

1.2.5 Bilan

Les approches à base de faibles connaissances ont cherché à alléger les ressources linguistiques nécessaires aux premières approches symboliques en utilisant des indices de surface. Les performances obtenues par les divers systèmes *knowledge-poor* s'alignent sur celles obtenues par les premières approches syntaxiques et montrent donc que la combinaison de plusieurs indices permet de retrouver le bon antécédent pour une anaphore donnée.

Parce qu'elles ne nécessitent, pour la plupart, qu'un étiquetage du texte et qu'elles reposent sur des règles contextuelles simples, ces approches robustes, bien qu'essentiellement conçues pour l'anglais, sont facilement reproductibles et transposables à d'autres langues (*cf.* Mitkov, 1998). De plus, l'architecture modulaire de ces systèmes offre la possibilité d'ajouter de nouvelles connaissances.

Néanmoins, (Salmon-Alt, 2001 ; Landragin, 2004 ; Weissenbacher, 2008) dégagent quelques lacunes des algorithmes *knowledge-poor* concernant notamment la validité scientifique des heuristiques utilisées. En effet, certains auteurs ne justifient pas les préférences accordées à certains types d'antécédents dans leurs calculs. Par exemple, une des heuristiques de (Mitkov, 1998) consiste à privilégier les candidats antécédents définis, mais nous avons vu, lors de notre étude de corpus (chapitre 3) que le premier maillon des chaînes de référence des textes juridiques étaient essentiellement des groupes nominaux indéfinis (*e.g.* « un Etat-Membre », « une décision »). Ce critère est donc étroitement lié au domaine et n'est pas forcément valable pour tout genre textuel. Une autre lacune concerne la pondération des scores, déjà présente dans les premières approches symboliques : pourquoi affecter des scores négatifs à certains candidats ?, quels critères permettent de déterminer une plage de scores allant de -1 à 2, de 50 à 100 suivant les systèmes ? Ces critères sont, pour la plupart, fondés essentiellement sur l'observation des corpus de développement. On peut donc, avec les auteurs (*e.g.* Mitkov le reconnaît lui-même), s'interroger sur la pertinence de ces scores intuitifs.

Dans la section suivante, nous nous focalisons sur les diverses approches symboliques proposées pour résoudre la référence de textes français.

1.3 Systèmes français

La plupart des approches symboliques que nous avons présentées jusqu'à présent opèrent sur des corpus anglais. Parce que notre projet porte sur le traitement de textes français, nous proposons une présentation de quelques systèmes symboliques de résolution de la référence développés pour le français. Ces systèmes, inspirés des approches syntaxiques et *knowledge-poor*, demeurent plus spécialisés que leurs aînés (dialogue homme-machine, détection d'événements, etc.) et font appel aux théories linguistiques et cognitives du discours.

1.3.1 Approches cognitives

Dans la lignée du modèle du contexte d'(Alshawi, 1987), des approches cognitives telles que (Popescu-Belis *et al.*, 1998 ; Salmon-Alt, 2001 ; Dupont, 2003) ont été développées pour la résolution de la référence en français.

1.3.1.1 Le modèle des représentations mentales de (Popescu-Belis *et al.*, 1998)[26]

(Popescu-Belis *et al.*, 1998) considèrent la résolution de la référence comme la construction de *représentations mentales*[27] des référents du discours et pas seulement l'identification des liens de coréférence établis entre antécédents et anaphores. Une représentation mentale est composée d'une série d'expressions référentielles se rapportant au même référent[28]. Elle peut être un objet concret, une personne ou un événement.

Les auteurs ont proposé un système à base de faibles connaissances qui calcule une valeur d'activation (*i.e.* sa saillance) pour chaque représentation mentale (obtenue à partir de la moyenne des valeurs d'activation des expressions référentielles qu'elle contient), modélisant la saillance de la représentation mentale présente à l'esprit d'un locuteur à un moment donné. Ainsi, pour une expression référentielle donnée, le système détermine si cette expression est à

[26] (Salmon-Alt, 2001) propose aussi un modèle cognitif des représentations mentales pour la résolution de la coréférence dans des dialogues finalisés homme-machine.

[27] Cette notion est empruntée à (Reboul *et al.*, 1997), dans le cadre du projet CERVICAL (*C*ommunication *E*t *R*éférence : *V*ers une *I*nformatique *C*ollaborant *A*vec la *L*inguistique).

[28] (Reboul et Gaiffe, 1999) utilisent la notion de *paquet* d'expressions référentielles.

rattacher à une représentation mentale existante ou non (et dans ce dernier cas, une nouvelle représentation mentale est créée). Cette décision s'effectue suivant un ensemble de règles, similaires à celles utilisées par (Lappin et Leass, 1994), permettant d'éliminer les représentations mentales auxquelles l'expression ne peut pas référer.

Le système procède en deux étapes (Popescu-Belis et Robba, 1998) :

- d'abord, il détermine l'ensemble des représentations mentales pouvant être rattachées à l'expression référentielle. Pour ce faire, trois contraintes de sélection sont testées entre l'expression référentielle à interpréter et les expressions référentielles contenues dans chacune des représentations mentales : deux contraintes morpho-syntaxiques comparent l'accord en genre et en nombre, une contrainte sémantique s'assure de la compatibilité entre les têtes lexicales et les modifieurs des groupes nominaux des expressions référentielles, *via* un dictionnaire des synonymes, hyperonymes et hyponymes spécifique au corpus utilisé,
- ensuite, il sélectionne, parmi les représentations mentales restantes, celle possédant la valeur d'activation la plus élevée et il associe cette représentation mentale à l'expression référentielle. Le calcul du taux d'activation de chaque représentation mentale s'effectue à partir d'un taux d'activation initial de 15 (exemple utilisé dans (Popescu-Belis, 1999 : 212)). Puis, si la représentation est réactivée (*i.e.* si une expression référentielle est rattachée à cette représentation mentale), un poids supplémentaire lui est attribué suivant le type de l'expression référentielle (nom propre : 40, nom commun : 20, pronom : 10). Au cours du déroulement du texte, l'activation de chaque représentation mentale diminue (4 à chaque nouveau paragraphe, 2 à chaque nouvelle phrase). L'activation est donc un facteur dynamique. S'il n'existe pas de représentation mentale compatible avec une valeur d'activation suffisante, une nouvelle représentation mentale est alors créée.

Un paramètre auxiliaire à l'activation de chaque représentation mentale permet de fixer un nombre maximal de représentations mentales à garder en mémoire (optimalement entre 15 et 20). Lorsque le score d'activation d'une représentation mentale devient nul, la représentation mentale est supprimée de cette mémoire à court terme. Cela permet d'accélérer le processus de sélection des représentations mentales pour une expression référentielle donnée.

Pour mener leur évaluation, les auteurs ont utilisé un corpus restreint composé de deux textes narratifs : *Vittoria Accoramboni* de Stendhal (2630 mots) et un extrait du premier chapitre du *Père Goriot* de Balzac (7405 mots). Le corpus a

été annoté manuellement en paragraphes et en phrases et les expressions référentielles ont été délimitées. Les performances obtenues s'alignent sur les systèmes présentés jusqu'alors : entre 65,48% pour *Vittoria Accoramboni* et 78,4% pour *Le Père Goriot*, validant l'utilité de l'activation et de la comparaison d'une expression référentielle non pas avec un seul antécédent, mais avec l'ensemble des antécédents contenus dans une représentation mentale. Les auteurs soulignent néanmoins qu'avec l'ajout de connaissances supplémentaires spécifiques au corpus, les scores obtenus seraient sensiblement améliorés.

1.3.1.2 Le modèle d'identification des entités de (Dupont, 2003)

Pour résoudre la référence, (Dupont, 2003) décrit un modèle d'identification des entités bipartite basé sur :

- une approche cognitive de la construction du sens, le *modèle des attentes*, inspiré du modèle du contexte d'(Alshawi, 1987),
- une classification des différentes expressions référentielles inspirée de la théorie de l'accessibilité d'(Ariel, 1990).

Ainsi, deux calculs distincts sont proposés respectivement : le calcul de saillance des entités d'une part et l'identification des entités d'autre part.

Le *modèle des attentes* accorde une place centrale au contexte au sens large (*i.e.* les connaissances supposées du lecteur, le cotexte et le contexte physique du texte tel que le support ou les conditions d'énonciation) pour l'interprétation. C'est notamment le contexte qui permet de résoudre sans ambiguïtés un énoncé tel que « dix vols par jour »[29], inscrit sur un dépliant publicitaire d'une compagnie aérienne. De ce fait, le *modèle des attentes* considère qu'une série d'entités *préconstruites* sont déjà présentes avant que ne débute le texte (*e.g.* dans une recette de cuisine, le lecteur s'attend à ce que des ingrédients et des ustensiles soient évoqués dans le texte). Puis, au fil du texte, la liste des entités va être modifiée suivant leur état d'activation (*i.e.* si les entités préconstruites sont effectivement évoquées ou si de nouvelles entités sont construites). Cet état d'activation est mesuré par un score de *saillance de l'entité en un point donné du texte*. Le score de saillance est attribué suivant une échelle arbitraire et il va évoluer au fil du texte, conformément au mécanisme de persistance-dégradation du modèle d'(Alshawi, 1987). Par exemple, pour chaque frontière de phrase, la saillance est multipliée par 0,75 (*i.e.* la saillance est diminuée de 25%) ce qui permet d'éliminer les entités qui ne sont plus mentionnées dans le texte. En

[29] Cet exemple est emprunté à (Victorri, 2005).

revanche, pour chaque nouvelle mention d'une entité, la valeur de saillance augmente suivant trois principaux facteurs :

- les *marques de mise en relief* (*i.e.* thématisation et focalisation) telles que les marqueurs cadratifs (« en ce qui concerne », « quant à ») ou les constructions clivées en « c'est… qui » qui vont introduire une nouvelle entité dans le discours et lui conférer une saillance forte,
- la *hiérarchie des fonctions grammaticales et des relations actancielles* : la position sujet attribue une saillance plus élevée que la position objet, le rôle d'agent attribue une saillance plus élevée que le rôle de patient,
- la *distance syntaxique* : la position sujet de la principale attribue un score de saillance plus élevé que le sujet d'une subordonnée.

Durant l'étape d'*identification des entités*, le modèle apparie les expressions référentielles aux entités *via* deux paramètres inspirés de l'échelle d'accessibilité d'(Ariel, 1990) : la *plage de saillances admissibles* et l'*importance de la concordance*. La *plage de saillances admissibles* d'un type d'expression référentielle permet de délimiter un intervalle de saillance dans lequel l'expression référentielle détermine la saillance de l'entité à laquelle elle réfère. Par exemple, pour qu'une entité soit évoquée par un pronom, sa plage de saillances doit être restreinte aux fortes saillances. *L'importance de la concordance* indique le poids à attribuer à la plage de saillances admissibles pour chaque type d'expressions référentielles. Par exemple, pour un pronom, l'importance de la concordance doit être très élevée.

Le modèle a été implémenté dans le système CALCOREF. A partir d'un texte brut, une analyse syntaxico-sémantique est effectuée (analyse syntaxique, identification des entités nommées et des emplois impersonnels du pronom « il »). Puis le texte est parcouru phrase par phrase par l'algorithme de calcul de la référence. La saillance des entités existantes est dégradée, les expressions référentielles sont parcourues suivant l'ordre linéaire du texte et elles sont associées aux entités. La saillance des entités est alors mise à jour. L'appariement des expressions référentielles avec les entités s'effectue *via* une série de critères classiques (distance entre l'expression référentielle et l'entité, compatibilité en genre et nombre, supériorité accordée à la position anaphorique, reprise de la tête lexicale, etc.) en plus de la concordance.

Une évaluation manuelle de CALCOREF a été effectuée sur les 3 premiers paragraphes (350 mots) de *La Peste* de Camus. Les performances obtenues sont élevées (87,86%). Mais, bien que le système permette d'identifier les cataphores et qu'il ne se restreint pas à la résolution des anaphores pronominales (comme la

plupart de ses prédécesseurs), les premiers résultats encourageants obtenus sont à nuancer relativement à la taille du corpus d'évaluation utilisé. Une évaluation menée sur un corpus plus large permettrait de confirmer ces performances.

1.3.2 Le système de résolution d'anaphores d'(Hernandez, 2004)

Dans le cadre de la description thématique de documents scientifiques, (Hernandez, 2004) a proposé un module d'identification des anaphores robuste, SRA (Système de Résolution d'Anaphores), inspiré des approches à base de faibles connaissances (Kennedy et Boguraev, 1996 ; Mitkov, 1998). SRA a été conçu pour fonctionner à la fois sur des textes français (Hernandez et Grau, 2003a) et anglais (Hernandez et Grau, 2003b).

Le système proposé utilise des connaissances essentiellement basées sur des heuristiques syntaxiques fondées sur des indices de surface. A partir d'un texte étiqueté morpho-syntaxiquement, le système repère tout d'abord les expressions référentielles du texte (groupes nominaux simples et pronoms) à l'aide de patrons syntaxiques, puis il sélectionne les candidats antécédents pour une anaphore donnée et il ordonne les candidats suivant leur pertinence et la validation de contraintes syntaxiques. Dans son approche, (Hernandez, 2004 : 115) considère uniquement comme anaphores les « expressions nominales introduites par un démonstratif et les pronoms personnels à la troisième personne », afin de ne traiter que des anaphores susceptibles de réaliser des liens extra-phrastiques.

La sélection des candidats antécédents potentiels s'effectue dans une fenêtre de trois phrases (environ 7 propositions) en amont de l'anaphore. Seuls sont considérés comme antécédents potentiels les groupes nominaux simples. Chacun des candidats potentiels se voit affecter deux poids ajustés empiriquement : un poids *absolu* et un poids *relatif*. Le poids *absolu* correspond aux caractéristiques du candidat (*i.e.* son type, sa fonction syntaxique, sa position dans la phrase, voir Tableau 20). Ces caractéristiques sont déterminées *via* des heuristiques, par exemple, la fonction grammaticale d'un candidat est donnée suivant sa position par rapport au verbe de la proposition et la présence éventuelle d'une préposition à sa tête (*i.e.* le candidat est sujet s'il précède le verbe de la principale). Le poids *relatif* correspond aux caractéristiques reliant l'antécédent à l'anaphore : sa position par rapport à l'anaphore (récence), le parallélisme syntaxique (voir Tableau 21).

Une fois les poids attribués à chaque candidat antécédent potentiel, les antécédents sont ordonnés suivant 5 degrés de contraintes, classés par ordre décroissant de priorité :

- correspondance de la tête sémantique et du genre/nombre du candidat avec l'anaphore,
- correspondance de la tête sémantique du candidat avec celle de l'anaphore,
- correspondance du genre et du nombre du candidat avec l'anaphore,
- poids absolu et relatif les plus élevés du candidat,
- poids absolu le plus élevé du candidat.

Le premier candidat antécédent validant le plus haut degré est alors considéré comme l'antécédent de l'anaphore.

Caractéristique	Poids absolu
sujet	65
cod	50
coi	40
cos	15
défini	15
démonstratif	15
possessif	15
thème	15
rhème	10
taille	# de tokens * 2

Tableau 20 – Poids *absolu* affecté à chaque candidat antécédent, d'après (Hernandez, 2004)

Caractéristique	Poids relatif
En fonction de sa position par rapport à l'anaphore	
dans la phrase courante	45
dans la proposition courante	35
dans la phrase précédente	30
dans la proposition précédente	20
dans le paragraphe courant	25
dans le segment courant	10
En fonction du parallélisme syntaxique avec l'anaphore	
nominal(anaphore) et sujet(anaphore) et (cod(antécédent) ou coi(antécédent))	40
pronomPersonnel(anaphore) et sujet(anaphore) et sujet(antécédent)	40
En fonction du nombre d'anaphores associées à l'antécédent	
nombre de liens anaphoriques	# d'anaphores * 2

Tableau 21 – Poids *relatif* affecté à chaque candidat antécédent

Le système SRA a été évalué manuellement sur deux extraits du texte exemple en anglais de (Barzilay et Elhadad, 1997), représentant 193 mots au total (85 mots et 108 mots). Le système résout correctement les anaphores (9 cas) lorsque le candidat antécédent est un groupe nominal et que l'anaphore est un groupe nominal démonstratif ou un pronom. Les autres types d'anaphores ne sont résolus que partiellement (3 cas sur 9). Aucune donnée chiffrée (*i.e.* performances en pourcentages) n'est fournie par (Hernandez, 2004). L'auteur affirme cependant que son système obtient des performances équivalentes à celles de (Lappin et Leass, 1994 ; Mitkov, 1998 ; Kennedy et Boguraev, 1996). Il demeure néanmoins difficile d'en juger, vu la taille restreinte du corpus utilisé. De plus, aucune évaluation sur un texte français n'est fournie par l'auteur[1], ce qui ne permet pas d'apprécier les performances du système SRA.

1.3.3 Systèmes spécialisés

Les trois premiers systèmes que nous présentons dans cette section permettent de résoudre des types de relations de référence particuliers : les anaphores infidèles chez (Salmon-Alt, 2004), les anaphores événementielles chez (Bittar, 2006), les chaînes de référence initiées par un nom propre chez (Boudreau et Kittredge, 2006). De leur côté, les deux dernières approches focalisent leur traitement sur des corpus issus de domaines spécialisés : les accidents de la route chez (Nouioua, 2007) et les élections présidentielles de 2007 chez (Adam, 2007).

1.3.3.1 Le système de résolution des anaphores infidèles de (Salmon-Alt, 2004)

(Salmon-Alt, 2004) propose un système de résolution automatique des anaphores infidèles utilisant diverses ressources linguistiques. Les anaphores « infidèles » (Kleiber, 1994) sont des groupes nominaux anaphoriques et coréférentiels dont la tête nominale est différente de celle de l'antécédent, par exemple « le médecin »… « ce docteur ». Le système récupère en entrée 78 descriptions définies (issues d'un extrait du corpus *Le Monde* de 9 000 mots environ étiqueté et analysé syntaxiquement) classées comme anaphores infidèles par 2 annotateurs humains ainsi que diverses connaissances linguistiques : fonction syntaxique, tête du groupe nominal, déterminant, genre, nombre. L'objectif est de trouver, pour chaque anaphore infidèle, le premier antécédent situé dans un contexte de 5 phrases en amont de l'anaphore qui satisfasse une série de contraintes (tête de l'antécédent, accord en genre et en nombre, distance phrastique). Pour ce faire, l'algorithme teste plusieurs configurations en cascade de ces divers paramètres

[1] Aucune évaluation n'est donnée dans l'article de (Hernandez et Grau, 2003a).

(variation du type d'antécédent (*e.g.* nom propre, groupe nominal défini), variation de la distance phrastique, accord uniquement en genre ou en nombre entre l'antécédent et l'anaphore).

Le système de résolution des anaphores infidèles obtient des performances de 31,9% pour la meilleure configuration (*i.e.* utilisation d'un dictionnaire de synonymes, accord uniquement en genre entre l'antécédent et l'anaphore, contexte de 3 phrases). (Salmon-Alt, 2004) souligne que ces résultats sont comparables à ceux obtenus par (Poesio *et al.*, 2002) pour traiter les mêmes phénomènes en anglais, tout en reconnaissant le travail restant encore à faire pour obtenir un système performant et utilisable à plus grande échelle. Notons aussi qu'en supprimant les ressources linguistiques utilisées dans cette approche (dictionnaire de synonymes, analyse syntaxique fine et annotation des entités nommées), les performances du système chutent à 15%, ce qui montre l'efficacité de l'emploi de telles ressources pour cette tâche.

1.3.3.2 Le modèle de résolution des anaphores événementielles de (Bittar, 2006)

De son côté, (Bittar, 2006) a proposé une méthode de résolution des anaphores événementielles pour le français. Les anaphores événementielles sont un cas particulier d'anaphores abstraites[2] (*i.e.* la reprise d'une phrase ou d'un verbe). Une anaphore événementielle est une reprise par un pronom (« cela », « ça », « ce », « le », « y », en ») d'un événement présent en amont dans le discours. Cet événement peut être un groupe nominal (*e.g.* « le retrait du permis de séjour ») ou une phrase (« (que) le permis de séjour a été retiré »).

Le système proposé se concentre sur les antécédents phrastiques uniquement. A partir d'un texte étiqueté morpho-syntaxiquement et découpé en propositions, le système identifie les emplois non anaphoriques des pronoms (*e.g.* « ça arrive ») afin de les éliminer. Cette étape s'effectue *via* l'utilisation de listes de verbes ayant pour sujet un pronom non anaphorique, comme « aller » (*e.g.* « ça va »), « étonner », etc. Puis, les verbes événementiels sont repérés car ils permettent d'identifier les événements sous forme phrastique. Suivant la classification de (Vendler, 1957), les verbes événementiels décrivent des achèvements (*e.g.* « réussir », « gagner »), des activités (*e.g.* « manger », « chanter ») ou des accomplissements (*e.g.* « peindre le portail », « lire le journal »). Les événements coordonnés ou « groupes événementiels » (*e.g.* « Paul a glissé et s'est cassé le

[2] L'anaphore abstraite est définie par (Amsili *et al.*, 2005 : 16) comme suit : « une anaphore dont la dénotation est une entité abstraite, au sens de (Asher, 1993), à savoir un objet sémantique dont la réalisation syntaxique privilégiée est la phrase simple ».

poignet ») sont ensuite annotés car ils sont susceptibles d'être repris par un seul pronom (*e.g.* « cela s'est produit ce matin »). Les anaphores pronominales sont alors détectées suivant les *conteneurs événementiels* (Vendler, 1957). Un conteneur regroupe les verbes qui nécessitent un sujet (« se produire », « avoir lieu ») ou un objet (« assister à », « être témoin de ») événementiel comme argument. Les conteneurs permettent de sélectionner le type d'antécédent d'un pronom. Par exemple, dans « Paul est soulagé que *l'audience soit annulée.* **Ça** s'est décidé à la dernière minute. », le conteneur « se décider » nécessite un sujet événementiel. Le pronom en position sujet de ce verbe (*i.e.* « ça ») se doit d'anaphoriser un événement afin d'assurer la cohérence du discours. Ainsi, le seul antécédent possible est « l'audience soit annulée ».

Une fois ces annotations effectuées, un algorithme de décision, inspiré des travaux de (Lappin et Leass, 1994 ; Mitkov, 1998), attribue un score à chaque candidat antécédent potentiel pour une anaphore pronominale donnée suivant une série de facteurs (distance, correspondance du temps entre l'antécédent et le conteneur événementiel sous la portée duquel est le pronom anaphorique, préférence pour un événement déjà anaphorisé, etc.). L'évaluation de chaque candidat s'effectue dans une fenêtre de 10 propositions en amont du pronom anaphorique. Le candidat événement dont le score total est le plus élevé est choisi comme antécédent. En cas d'égalité entre deux candidats, le candidat le plus proche de l'anaphore est alors préféré.

Le modèle de résolution des anaphores événementielles a été évalué manuellement sur 3 extraits de textes issus de *Frantext*[3] (49 mots, 43 mots et 91 mots). Le critère de distance joue un rôle important dans cette approche, puisque le score qui lui a été attribué s'échelonne de 10 à 0 (plus le candidat est éloigné de l'anaphore (en nombre de propositions), plus son score diminue de 1) alors que les autres critères oscillent entre -10 et 5. De ce fait, il apparaît dans les résultats obtenus que le critère de distance déséquilibre le score final de chaque candidat, ce qui a pour effet de sélectionner de mauvais antécédents. L'implémentation de ce modèle permettrait d'équilibrer ce facteur de préférence (*e.g.* en effectuant des tests à grande échelle sur l'attribution de valeur suivant l'éloignement du candidat).

1.3.3.3 L'approche « minimaliste » de (Boudreau et Kittredge, 2006)

A partir de l'étude de 3 variétés de textes (critiques de films, fusions de compagnies, manuels d'installation informatique), (Boudreau et Kittredge, 2006)

[3] www.frantext.fr

ont développé un algorithme à base de connaissances linguistiques limitées pour identifier les chaînes de référence initiées par un nom propre en français.

Le système prend en entrée un texte brut et identifie, *via* des listes de mots constituées à l'issue de l'étude des 3 corpus, différentes expressions référentielles : les pronoms, les groupes nominaux simples associés au domaine (*e.g.* noms de fonction (« acteur »), relations familiales (« père »), termes informatiques (« fichier »), termes liés au domaine des affaires (« conseil »)) ainsi que les noms propres (identifiés uniquement grâce à la présence d'une majuscule). Puis, le système crée des groupes nominaux complexes (*i.e.* groupes adjectivaux (« la petite fille »), groupes prépositionnels (« la fille de Paul »), noms propres complets (« Barack Obama »)) et attribue des fonctions syntaxiques suivant la position des expressions référentielles par rapport aux prépositions ou aux autres expressions référentielles identifiées (*e.g.* pour les appositions).

Une fois ces premiers traitements effectués, l'algorithme sélectionne les paires antécédents-anaphores suivant divers facteurs pondérés : distance entre l'antécédent et l'anaphore, répétition de la tête lexicale, appartenance de la tête lexicale au vocabulaire du domaine, fonction syntaxique (sujet>COD>COI>autre), *etc.*

(Boudreau et Kittredge, 2006) concluent qu'avec l'utilisation de ressources linguistiques limitées, leur système est capable d'identifier des chaînes de référence initiées par un nom propre dans des textes courts. Néanmoins, aucune évaluation du système n'est fournie, ce qui rend difficile l'appréciation des performances de leur système. De plus, les autrices reconnaissent que l'attribution des fonctions syntaxiques est approximative, que l'algorithme peine à délimiter les frontières des groupes complexes et que seuls quelques types d'expressions référentielles sont identifiés, ce qui engendre des erreurs dans la chaîne de traitements avant même de passer à l'étape de calcul de la référence. Ainsi, même si l'approche minimaliste proposée par (Boudreau et Kittredge, 2006) est intéressante, elle ne permet pas, en l'état, d'être utilisée à plus grande échelle.

1.3.3.4 La résolution des anaphores dans les textes d'accidents de la route (Nouioua, 2007)

Dans le cadre d'un système de compréhension automatique, (Nouioua, 2007) a proposé une heuristique pour la résolution des anaphores dans un corpus composé de 160 textes d'accidents de la route. A partir d'une analyse syntaxique de surface, diverses relations, appelées « littéraux linguistiques », sont extraites (*e.g.* verbe-sujet, verbe-objet, préposition-verbe-complément, etc.). Ces relations sont

ensuite post-traitées, notamment *via* la résolution des anaphores, afin de déterminer la cause de l'accident.

Dans le système de résolution d'anaphores, sont considérés comme référents potentiels les personnes impliquées dans l'accident mais aussi les véhicules (emplois métonymiques, *e.g.* « le véhicule a freiné »). Chaque référent est associé à une information relative à sa nature (*e.g.* « véhicule » (par défaut), « vélo », « moto », « voiture », « camion »). L'algorithme résout d'abord les anaphores simples (*e.g.* « ma », « mon » réfèrent à la référence « auteur »), puis chaque anaphore résolue est remplacée par une constante issue d'une liste prédéfinie (*e.g.* nom_de_personne_1, nom_de_véhicule) afin de propager la référence et la nature du référent. Les anaphores non résolues sont alors parcourues par l'algorithme qui dresse la liste des antécédents compatibles :

- si la liste contient un seul antécédent potentiel, il est choisi comme antécédent,
- si la liste comporte plusieurs candidats potentiels, celui qui est le plus proche de l'anaphore est choisi, en accordant une moindre importance aux candidats dont la référence est « auteur » (qui doit être présent dès le début du texte),
- si la liste est vide, un nouveau référent est créé.

La décision d'ajouter un nouveau référent s'effectue *via* l'utilisation d'indices, par exemple, la présence d'un déterminant indéfini, de noms communs tels que « adversaire » ou d'adjectifs tels que « autre ». A l'issue de la résolution de chacune des anaphores, la procédure de propagation est de nouveau déclenchée. Le système s'arrête lorsque toutes les anaphores sont résolues (procédure itérative).

Lors de l'évaluation du système (effectuée sur 87 textes des 160 textes d'accident de la route initiaux), les performances obtenues s'élèvent à 95%. L'heuristique proposée dans cette approche restreinte au domaine des accidents de la route demeure efficace. Néanmoins, comme le reconnaît l'auteur, cette approche souffre du manque de généralité lié à sa dépendance au domaine spécialisé des accidents de la route. La mise au point d'un algorithme plus général nécessiterait la création d'un nombre conséquent d'heuristiques.

1.3.3.5 La résolution de la coréférence dans des articles politiques (Adam, 2007)

Afin de créer des associations entre expressions référentielles désignant un même référent (*e.g.* « M. Hollande », « Hollande », « le président de la République Française ») pour une application de veille terminologique, (Adam, 2007) a mis en place un système de résolution de la coréférence en français.

Le système utilise en entrée un texte analysé syntaxiquement par l'analyseur syntaxique *Syntex* (Bourigault *et al.*, 2005). Le texte comporte ainsi diverses informations : catégorie grammaticale, fonction grammaticale, lien entre le nom propre et le prénom, lien entre une particule et le nom propre. Des listes de mots-clés (noms de fonctions essentiellement, par exemple « député », « ministre ») sont ensuite extraites semi-automatiquement à partir d'un corpus d'entraînement de 2,5 millions de mots (articles issus de trois journaux : *Le Monde*, *Le Figaro* et *Libération* de 2006-2007) et filtrées manuellement.

A l'issue de ces prétraitements, le module de balisage identifie les noms propres et les groupes nominaux potentiellement référentiels en effectuant 2 passages successifs sur le corpus : lors du premier passage, tous les groupes nominaux dont la tête lexicale figure parmi les mots-clés sont annotés (*e.g.* « ministre de l'intérieur »). Les noms propres sont annotés selon leur type et se voient affectés d'un genre. Lors du deuxième passage, les noms propres sont associés à un identifiant (*e.g.* M. Hollande_hollande_françois). Un module de calcul de scores d'association permet ensuite d'extraire des couples [nom propre-groupe nominal potentiellement coréférentiel], par exemple un nom propre et sa fonction associée (« Le candidat de l'UDF, François Bayrou »), *via* des patrons d'identification. Un dernier module de résolution en contexte permet de résoudre les groupes nominaux balisés en utilisant une série d'indices contextuels : compatibilité en genre entre le nom propre et le candidat, distance phrastique, fonction grammaticale. Suite au passage de ces divers modules, chaque candidat antécédent potentiel se voit attribuer un score de saillance total, permettant de le classer. Le candidat ayant le score de saillance le plus élevé est choisi comme antécédent d'un nom propre donné.

Le système a été évalué sur un corpus manuellement annoté de 46 644 mots composé d'articles issus de 3 journaux (*Le Monde*, *Le Figaro* et *Libération* de février 2007) portant sur l'élection présidentielle de 2007. Bien que le corpus d'évaluation soit homogène, que la tâche d'identification requière uniquement le rattachement de groupes nominaux à des noms propres et que les référents soient en nombre fini (*i.e.* le contexte précis des élections présidentielles amène à

trouver de nombreuses occurrences des mêmes personnalités politiques et de leurs reprises), les performances moyennes obtenues ne sont que de 67,14%. Les erreurs relevées concernent essentiellement des problèmes de reconnaissance des prénoms seuls par l'analyseur syntaxique (en effet, *Syntex* n'annote que les prénoms suivis d'un nom propre) ainsi que le traitement des abréviations pour désigner des noms propres (*e.g.* « DSK » pour « Dominique Strauss-Kahn »).

1.4 Bilan

Nous avons vu que les systèmes symboliques fonctionnaient par application de règles, par vérification de critères ou par comparaison des propriétés des expressions mises en jeu. Les diverses approches que nous avons présentées ont mis en évidence l'existence de deux groupes distincts (Salmon-Alt, 2001) :

- d'un côté, les systèmes riches en connaissances (étiquetage, lemmatisation, analyse syntaxique, identification des noms propres) et à couverture réduite, mais proches des modèles théoriques de la référence (Hobbs, 1978 ; Alshawi, 1987) ;
- de l'autre côté, des systèmes robustes et pauvres en connaissances qui exploitent des indices de surface (Mitkov, 1998 ; Bontcheva *et al.*, 2002), mais dont les fondements théoriques ne sont pas toujours justifiés concernant l'attribution des scores de saillance notamment.

Spécifiquement pour le français, les systèmes de résolution de la référence que nous avons présentés se sont inspirés de ces deux groupes d'approches, en se spécialisant sur un type particulier d'anaphores (anaphores événementielles, anaphores infidèles) et/ou sur un domaine précis (Nouioua, 2007 ; Adam, 2007). Néanmoins, les évaluations menées sont souvent réalisées de manière manuelle et elles sont effectuées sur un corpus restreint (moins de 10 000 mots). Bien qu'encourageantes, ces approches sont rarement implémentées et faiblement exemplifiées, ce qui limite, par conséquent, leur utilisation.

Dans le même ordre d'idée, plusieurs approches symboliques que nous avons présentées ont utilisé des scores dynamiques pour rendre compte de la variation de la saillance des antécédents au fil du texte (Alshawi, 1987 ; Lappin et Leass, 1994 ; Dupont, 2003). L'attribution des valeurs, nous l'avons déjà évoqué, reste arbitraire et difficile à justifier : pourquoi une entité en position sujet n'a-t-elle pas le même poids dans une phrase et la suivante ?, pourquoi la saillance d'une entité doit-elle se dégrader à chaque fin de phrase ? Bien que cette gestion dynamique des scores de saillance des entités soit tout à fait intéressante, il nous

paraît pour le moment relativement difficile d'en tenir compte, à grande échelle, dans notre calcul de la référence.

Pour combler les limites rencontrées par les systèmes à base de règles, des systèmes par apprentissage ont été mis en place. Nous proposons une présentation de plusieurs de ces systèmes dans la section suivante.

2 Systèmes par apprentissage

Avec la mise à disposition de corpus d'entraînement annotés en relations de coréférence[4] *via* les diverses campagnes d'évaluation telles que MUC[5], ACE[6], SemEval[7], TREC[8], CoNLL[9] (voir chapitre 8), de nombreuses approches par apprentissage supervisé (Aone et Bennett, 1995 ; McCarthy et Lehnert, 1995 ; McCarthy, 1996 ; Soon *et al.*, 2001 ; Ng et Cardie, 2002 ; Yang *et al.*, 2003 ; Rahman et Ng, 2009 ; Ng, 2010) puis non supervisé (Cardie et Wagstaff, 1999 ; Bean et Riloff, 2004 ; Ng, 2008 ; Haghighi et Klein, 2007 ; Poon et Domingos, 2008) ont vu le jour. La principale différence entre les deux familles d'approches relève du type d'information à fournir pour l'apprentissage : les systèmes par apprentissage *supervisé* doivent savoir si les paires de mentions sont coréférentes ou non, tandis que les systèmes par apprentissage *non supervisé* ne nécessitent pas d'information sur la coréférentialité des paires.

La plupart des systèmes par apprentissage décomposent la tâche de résolution de la référence en deux étapes (Recasens, 2010) :

- la *classification*, où le système est entraîné sur un corpus de référence pour connaître les probabilités qu'une paire antécédent-anaphore soit coréférente ou non,
- le *regroupement* (*clustering*), dans lequel les paires coréférentes identifiées lors de l'étape de classification sont assemblées pour former des chaînes de coréférence.

Ces systèmes traitent, pour une grande partie, sur la base de corpus anglophones[10] ; le français n'ayant toujours pas, à l'heure actuelle, de large corpus pour l'écrit annoté en relations de coréférence.

Dans cette section, nous présentons quelques systèmes de calcul de la référence par apprentissage supervisé et non supervisé. Pour une revue exhaustive de ces

[4] *Cf.* Annexe 1 pour des exemples de corpus annotés pour l'apprentissage.
[5] *Message Understanding Conference*
[6] *Automatic Content Extraction*
[7] *Semantic Evaluation*
[8] *Text REtrieval Conference*
[9] *Conference on Natural Language Learning*
[10] *Cf.* la section 3 de ce chapitre pour un état des lieux sur les corpus disponibles.

systèmes, nous invitons le lecteur à consulter les travaux de (Ng, 2008 ; Rahman et Ng, 2009 ; Nøklestad, 2009 ; Ng, 2010 ; Poesio *et al.*, 2010 ; Zheng *et al.*, 2011).

2.1 Systèmes supervisés

Depuis plus d'une quinzaine d'années, de nombreux systèmes par apprentissage supervisé ont été proposés pour traiter la coréférence. (Rahman et Ng, 2009) distinguent trois grandes catégories d'approches :

- les modèles *mention-pair* ou *pairwise* cherchent à savoir si deux mentions (*i.e.* une paire) sont coréférentes ou non ;
- les modèles *mention-ranking* classent l'ensemble des antécédents potentiels pour une anaphore donnée ;
- les modèles *entity-mention* ou *entity-based* déterminent la probabilité qu'une mention réfère à une *entité* (*i.e.*, une série de mentions déjà classées comme étant coréférentes) ou *cluster*.

Nous proposons ci-dessous une présentation de quelques systèmes issus de chacune des catégories d'approches.

2.1.1 Les modèles *mention-pair*

Les modèles *mention-pair* s'appuient sur un classifieur binaire (*i.e.* oui/non) comparant une anaphore avec des antécédents potentiels issus des phrases précédentes. Les modèles travaillent par paires de mentions (une anaphore avec un antécédent potentiel). Ainsi, chaque paire de mentions est annotée positivement (*i.e.* les mentions sont coréférentes) ou négativement (*i.e.* les mentions ne sont pas coréférentes) par le classifieur. Ce dernier combine alors les paires jugées coréférentes en chaînes de coréférence.

Les modèles *mention-pair* ont été proposés à l'origine par (Aone et Bennett, 1995) et (McCarthy et Lehnert, 1995), mais ce sont les modèles de (Soon *et al.*, 2001) et de (Ng et Cardie, 2002) qui demeurent les plus implémentés[11].

[11] De nombreuses améliorations de ces modèles ont été proposées, notamment en enrichissant la série de traits initiale d'informations sémantiques issues de Wikipedia, WordNet (Ponzetto et Strube, 2006) ou en incorporant des connaissances linguistiques plus complexes (*e.g.* (Uryupina, 2006, 2007) a défini une série de 351 traits ; (Versley *et al.,* 2008) ont ajouté des traits syntaxiques ; (Bengtson et Roth, 2008) ont montré qu'en utilisant une série de traits fins, un simple modèle mention-pair pouvait dépasser des modèles beaucoup plus complexes ; (Charton *et al.*, 2013) ont utilisé un module définissant le type d'un nom propre) et spécifiques à une langue

2.1.1.1 Le modèle par arbre de décision de (Soon *et al.*, 2001)

(Soon *et al.*, 2001) ont proposé un système à base d'arbre de décision[12] pour la résolution de la coréférence à partir des corpus annotés MUC-6[13] et MUC-7[14] (genres journalistiques). Dans leur modèle de classification, (Soon *et al.*, 2001) définissent plusieurs modules de détection de mentions pour déterminer les mentions coréférentes d'un document d'entraînement. L'algorithme consiste en une série de 12 traits de surface représentant deux mentions `i` (l'antécédent potentiel) et `j` (l'anaphore). Tous les traits sont des booléens (*i.e.*, vrai/faux), à l'exception du premier trait (qui est un entier) :

[1] **DIST** : distance phrastique (nombre de phrases[15] entre `i` et `j`),
[2] **I-PRONOUN** : `i` est un pronom,
[3] **J-PRONOUN** : `j` est un pronom,
[4] **STR-MATCH** : identité de la tête lexicale (*i.e.* `i` est identique à `j`, *e.g.*, « la mention » et « cette mention »),
[5] **DEF_NP** : `j` est un groupe nominal défini,
[6] **DEM_NP** : `j` est un groupe nominal démonstratif,
[7] **NUMBER** : correspondance du nombre,
[8] **SEM_CLASS** : correspondance de classe sémantique de WordNet (les classes sémantiques potentielles pour un groupe nominal sont : *personne, organisation, objet, lieu, date, heure, unité monétaire, pourcentage, femme et homme*),
[9] **GENDER** : correspondance du genre,
[10] **PROPER_NAME** : `i` et `j` sont des noms propres,
[11] **ALIAS** : `i` est un *alias* de `j` (*i.e.* ce sont deux variantes lexicales d'une même entité nommée, p.e. « F. Hollande » et « François Hollande »), ou `j` est un *alias* de `i`,
[12] **APPOSITIVE** : `j` est une apposition de `i` (*e.g.*, « (François Hollande) `i`, (président de la République Française) `j` »).

Les traits *apposition*, *alias* et *identité de la tête lexicale* sont jugés comme les meilleurs indices de coréférence entre deux mentions.

((Hoste, 2005) a travaillé sur l'allemand, (Recasens, 2010) sur le catalan, (Broscheit *et al.*, 2010) sur l'italien, l'anglais et l'allemand).

[12] Un arbre de décision contient des règles de classification qui basent leur décision sur une suite de tests organisés de manière arborescente (par exemple : si l'on souhaite savoir si l'on peut aller à la plage, on pourra effectuer 3 tests : soleil ? → vrai ; chaleur ? → vrai ; weekend ? → vrai ⇨ plage).

[13] http://cs.nyu.edu/faculty/grishman/muc6.html

[14] http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/muc_7_toc.html

[15] Les valeurs possibles sont 0 si `i` et `j` sont situés dans la même phrase, 1 si `i` et `j` sont séparés d'une phrase, *etc.*

Les exemples d'entraînement sont fournis à l'algorithme d'apprentissage pour construire un classifieur. Puis, les exemples d'entraînement sont générés par le classifieur sous la forme de vecteurs de traits (*e.g.* les mentions `i` et `j` possèdent le même genre, le même nombre, `i` et `j` sont des *alias etc.*) représentant une paire de mentions coréférentes.

Lors de l'utilisation du modèle, les paires de mentions potentiellement coréférentes sont proposées au classifieur. Ce dernier attribue une réponse binaire ou probabilisée pour valider ou invalider leur relation. Une fois les paires de mentions coréférentes validées, un processus d'assemblage regroupe les paires de mentions dans des chaînes de coréférence.

2.1.1.2 Le modèle de (Ng et Cardie, 2002)

(Ng et Cardie, 2002) ont étendu la série de 12 traits de surface définis par (Soon *et al.*, 2001) à 53 traits. Les traits lexicaux, grammaticaux et sémantiques supplémentaires testent par exemple si l'antécédent et l'anaphore sont deux pronoms identiques, s'ils sont définis tous les deux, s'ils sont imbriqués (*e.g.* « `(`le père de `(`Marie`)i )j` »), s'ils sont animés, s'ils sont des sujets grammaticaux, si l'antécédent est une apposition et pas un groupe nominal indéfini, etc.

Les auteurs ont aussi comparé deux techniques permettant de sélectionner un antécédent particulier parmi tous les antécédents classés comme candidats potentiels par le classifieur :

- la technique *closest-first*, qui sélectionne l'antécédent le plus proche de l'anaphore ;
- la technique *best-first clustering*, qui sélectionne le candidat antécédent ayant le score de probabilité le plus élevé.

L'ajout de ces connaissances (extension du nombre de traits et techniques de sélection d'un antécédent) a permis d'améliorer la précision du système initial de (Soon *et al.*, 2001) mais après une sélection manuelle de 27 traits. En effet, l'utilisation de l'ensemble des 53 traits n'était pas compatible avec la taille du corpus d'apprentissage (MUC-6 et MUC-7).

2.1.1.3 Bilan

Bien que les modèles *mention-pair* soient populaires, (Rahman et Ng, 2009) dégagent deux inconvénients majeurs :

- d'une part, parce qu'ils fonctionnent par paires de mentions, ces modèles déterminent la probabilité qu'un antécédent potentiel soit un bon candidat

pour une anaphore donnée, mais pas qu'un antécédent potentiel soit un bon candidat par rapport à tous les autres candidats potentiels (*i.e.* qu'il y ait un ordre établi entre les antécédents potentiels) ;

– d'autre part, dans ces modèles, la décision de classification est prise pour un couple de mentions données. Ainsi, ces modèles ne tiennent pas compte des décisions précédentes prises par le classifieur. De ce fait, une anaphore pourra être intégrée à une chaîne de coréférence existante sans qu'il ait été vérifié qu'elle soit effectivement compatible.

Pour pallier ces deux faiblesses, les modèles *mention-ranking* et *entity-mention* ont été proposés.

2.1.2 Les modèles *mention-ranking*

Les modèles *mention-ranking* permettent de trier les antécédents potentiels pour une anaphore donnée suivant leur score. Ces modèles ont été proposés pour combler la première faiblesse rencontrée par les modèles *mention-pair*. En effet, l'approche du *ranking* paraît plus adaptée à la résolution de la coréférence car elle permet de tester tous les antécédents potentiels et donc de déterminer lequel des antécédents constitue le meilleur candidat pour une anaphore donnée.

La notion de *ranking* est issue des algorithmes basés sur la théorie du centrage (Grosz *et al.*, 1995 ; Walker *et al.*, 1998) et sa reformulation en terme de théorie de l'optimalité par (Beaver, 2004)[16] qui utilisent notamment les rôles grammaticaux pour classer les centres rétroactifs. (Connolly *et al.*, 1994) ont été les premiers à utiliser le *ranking* pour la résolution de la coréférence par apprentissage.

2.1.2.1 Le modèle de (Connolly et *al.*, 1994)

Dans leur modèle, le problème de classification est défini pour une anaphore et deux antécédents potentiels ; l'objectif étant de choisir parmi les deux antécédents celui dont le score est le plus élevé. La classification repose sur un ensemble de traits (similaires aux modèles *mention-pair*, *e.g.*, distance, type de mention, identité de la tête lexicale, etc.) portant sur les propriétés de l'anaphore et des deux antécédents ainsi que des relations établies entre les trois mentions.

[16] La notion de *ranking* est déjà présente dans les approches symboliques telles que celle de (Lappin et Leass, 1994) où des scores de saillance sont attribués à divers facteurs.

Le classifieur est appliqué successivement aux paires de candidats potentiels, en retenant le meilleur candidat à chaque fois. Le candidat « perdant » est éliminé et une nouvelle paire d'antécédents est testée. Cette nouvelle paire est composée du candidat « gagnant » et d'un nouvel antécédent potentiel. A son tour, la nouvelle paire est testée et un des deux candidats est retenu ; le candidat retenu pouvant être le même que précédemment. Le processus se poursuit jusqu'à ce que tous les antécédents potentiels aient été testés et le dernier candidat « vainqueur » est alors choisi comme antécédent.

Parce qu'elle confronte deux antécédents potentiels pour une anaphore donnée, la méthode du *ranking* a été qualifiée de « tournament model » par (Iida *et al.*, 2003) et de « twin-candidate model » par (Yang *et al.*, 2003).

Plus récemment, (Denis et Baldridge, 2008) ont proposé un modèle de *ranking* qui compare l'ensemble des antécédents potentiels simultanément pour une anaphore donnée.

2.1.2.2 Le modèle de *ranking* de (Denis et Baldridge, 2008)

Basés sur les modèles de résolution de la coréférence de (Yang *et al.*, 2003) et de résolution de l'anaphore pronominale de (Ng, 2005), (Denis et Baldridge, 2008) ont proposé, dans leur modèle de *ranking*, une procédure d'entrainement spécifique à chaque type d'anaphore (*i.e.* pronominale, définie, etc.). Ainsi, lors de l'apprentissage, chaque anaphore est associée à un ensemble de candidats. Cet ensemble contient le « bon » candidat antécédent et d'autres antécédents non compatibles. La sélection du bon candidat dépend du type d'anaphore. Par exemple, pour l'anaphore pronominale, le bon candidat est sélectionné suivant sa distance avec l'anaphore (*i.e.*, le bon candidat antécédent est l'antécédent le plus proche de l'anaphore). Pour sélectionner les autres antécédents non compatibles, le système récupère tous les antécédents situés dans une fenêtre de deux phrases autour de l'anaphore.

Lors de l'utilisation du modèle, le classifieur teste d'abord si le candidat anaphorique est bien une anaphore. En effet, à la différence des modèles *mention-pair*, le modèle du *ranking* compare toutes les expressions référentielles les unes avec les autres et perd donc cette information. Pour ce faire, le classifieur teste la forme de la mention (*i.e.*, le type d'expression référentielle, le nombre de *tokens*), sa position dans le texte et il compare la mention aux mentions précédentes du texte. Une fois ces premiers tests effectués, le classifieur compare les traits de l'anaphore avec ceux des antécédents potentiels ainsi que leurs traits relationnels (*i.e.* les traits décrivant les relations établies entre les mentions testées). Evalué

sur le corpus ACE, le modèle de (Denis et Baldridge, 2008) obtient une précision de 80,8%.

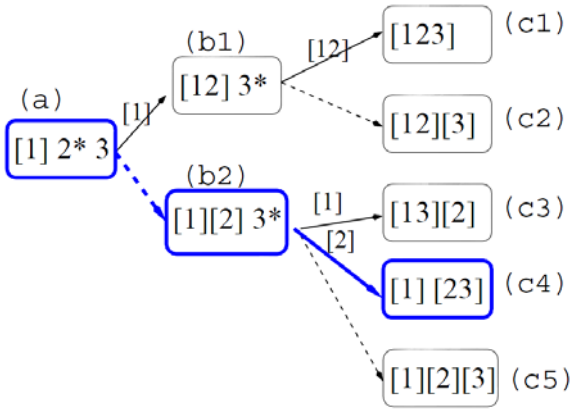
2.1.3 Les modèles *entity-mention*

De leur côté, les modèles *entity-mention* déterminent la probabilité qu'une mention réfère à un *cluster*, c'est-à dire à un ensemble de mentions coréférentes. Ces modèles ont été proposés pour combler la seconde faiblesse rencontrée par les modèles *mention-pair*. En effet, les modèles *entity-mention* sont des modèles plus complexes car ils essaient de modéliser la coréférence de manière globale (et non plus de manière locale comme cela était le cas pour les modèles *mention-pair*) afin de vérifier les liens déjà existants impliquant une anaphore donnée. Cela leur permet d'éviter de regrouper des paires antécédent-anaphore non compatibles lors de la phase de regroupement en chaînes de coréférence. Pour illustrer le problème posé par les modèles *mention-pair*, (McCallum et Wellner, 2003 : 1) utilisent l'exemple suivant : « For example, "Mr. Powell" may be correctly coresolved with "Powell," but particular grammatical circumstances may make the model incorrectly believe that "Powell" is coreferent with a nearby occurrence of "she" ». Ainsi, dans cet exemple, le modèle *mention-pair* va créer deux paires de mentions [Mr. Powell, Powell] et [Powell, she] car 1) il ne dispose pas d'information sur le genre du nom propre « Powell » et 2) il ne tient pas compte des décisions prises pour les paires précédentes impliquant l'anaphore. Cette première erreur va engendrer le regroupement des deux paires en une chaîne de coréférence [Mr. Powell, Powell, she].

Pour résoudre ce type d'erreurs, des travaux tels que (Luo *et al.*, 2004), (Daumé III et Marcu, 2005) ont proposé des modèles permettant de tester si une mention est coréférente avec un cluster (partiel) coréférentiel déjà existant.

2.1.3.1 Le modèle par arbre de Bell de (Luo *et al.*, 2004)

Le modèle de (Luo *et al.*, 2004) figure parmi les premiers modèles *entity-mention*. Ils ont choisi d'utiliser l'arbre de Bell pour représenter le processus de création des clusters à partir des mentions coréférentes. L'arbre de Bell représente l'espace de recherche pour la résolution de la coréférence et chaque nœud-feuille de l'arbre représente un potentiel cluster (voir Figure 24).



Dans la représentation en arbre de Bell, les mentions entre [ ] dénotent un cluster partiel (ou entité), les mentions actives (*active mention*) sont marquées d'une *, les flèches pleines signifient qu'une mention est liée à une entité existante (*in-focus* entity), les flèches en pointillés indiquent le début d'une nouvelle entité et les 5 résultats de coréférence potentiels sont numérotés de (c1) à (c5).

Figure 24 – Représentation en arbre de Bell de trois mentions (1, 2, 3), d'après (Luo *et al.*, 2004 : 136)

Dans l'exemple encadré en bleu de la Figure 24, un cluster partiel [1] est créé. La mention 2 est activée et elle a la possibilité de se lier à la mention 1 (nœud-feuille (b1)) ou d'ouvrir un nouveau cluster ((b2)). La mention 3 est ensuite activée et elle peut alors se lier au cluster de la mention 1 ((c3)), à celui de la mention 2 ((c4)) ou bien ouvrir un nouveau cluster ((c5)).

Le problème de résolution de la coréférence se résume alors à trouver le « bon chemin » depuis le nœud racine (a) vers les feuilles (nœuds (b) et (c)). Pour cela, le modèle *entity-mention* de (Luo *et al.*, 2004) calcule la probabilité qu'une mention active appartienne à un cluster existant. Une série de traits de base (lexicaux, syntaxiques, de distance) est alors testée pour déterminer la présence de liens entre la mention active et le cluster (*e.g.* pour le trait de correspondance en nombre, la valeur sera vraie si la mention possède le même nombre que l'ensemble des mentions du cluster, sinon, elle sera fausse).

(Luo *et al.*, 2004) ont comparé leur modèle à un modèle *mention-pair* (en adaptant les traits utilisés dans leur modèle *entity-mention*) sur le corpus ACE. Leur modèle *entity-mention* a obtenu des résultats plus faibles que le modèle *mention-pair*, notamment parce que les traits utilisés pour le modèle *entity-mention* étaient trop généraux (le modèle *mention-pair* possédait 20 fois plus de traits).

2.1.3.2 Le modèle en direct de (Daumé III et Marcu, 2005)

De leur côté, le modèle *entity-mention* de (Daumé III et Marcu, 2005) est basé sur un apprentissage en direct : un faisceau de recherche est utilisé pour corriger des décisions non optimales ou lorsqu'il devient impossible de trouver des solutions. Ainsi, l'algorithme repose sur une formule de mise à jour des paramètres de calcul au fil du texte.

Pour choisir les antécédents d'une anaphore donnée, l'algorithme regroupe les scores de l'anaphore et de chaque antécédent potentiel suivant diverses stratégies telles que « prendre le score le plus élevé parmi les mentions de la chaîne de coréférence » (*max-link*), « prendre le score moyen des mentions de la chaîne de coréférence » (*average-link*), « prendre le score de la mention la plus proche de la chaîne de coréférence » (*nearest-link*). (Daumé III et Marcu, 2005) utilisent aussi une méthode de regroupement (*intelligent-link*) spécifique au type de la mention en cours (nom propre, groupe nominal ou pronom). Par exemple, si la mention est un nom propre, elle est d'abord comparée aux autres noms propres de la chaîne de coréférence, s'il n'y en a pas, elle est ensuite comparée au dernier groupe nominal de la chaîne et s'il n'y en a pas non plus, il faut alors utiliser le score *max-link*.

Evalué sur le corpus ACE 2004 avec le score ACE[17], le modèle *entity-mention* de (Daumé III et Marcu, 2005) a obtenu des performances de 89,1% (il s'est classé au milieu des autres systèmes mis en compétition). Néanmoins, leur modèle a rencontré de nombreuses erreurs de regroupement pour les pronoms car il relie uniquement les pronoms aux autres pronoms ou aux noms propres. Ce choix est lié au corpus ACE qui contenait essentiellement des noms de personnes ou d'organisations.

2.1.4 Systèmes supervisés hybrides

Certains systèmes supervisés utilisent une approche hybride pour résoudre la coréférence (*i.e.* ils utilisent deux méthodes parmi les trois présentées ci-dessus). C'est notamment le cas de (Rahman et Ng, 2009) qui ont proposé un modèle de *cluster-ranking*. Leur système trie les clusters précédents pour une anaphore donnée plutôt que les candidats antécédents. De ce fait, la chaîne de coréférence dont le score est le plus élevé est choisie comme antécédent pour une anaphore donnée. Evalué sur le corpus ACE-2005, le modèle obtient de meilleures performances que les modèles *mention-pair*, *mention-ranking* et *entity-mention*.

De son côté, (Uryupina, 2010) a proposé un système hybride mêlant un modèle *mention-pair* et un modèle *mention-ranking* pour la résolution de la coréférence. Le système comporte tout d'abord une extension du modèle de (Soon *et al.*, 2001), reposant sur une série de 64 traits linguistiques et il utilise des classifieurs différents suivant le type de la mention (*i.e.* pronom, groupe nominal, nom propre). Puis, le système utilise les contraintes définies dans (Denis et Baldridge, 2008), à savoir la comparaison entre les traits de l'anaphore (selon son type) et

[17] http://www.itl.nist.gov/iad/894.01/tests/ace/2004/software/ace04-eval-scoring-v3.pdf

ceux des antécédents potentiels. Evalué sur le corpus SEMEVAL, le système a obtenu les meilleures performances parmi les divers systèmes en compétition.

Récemment, (Lassalle et Denis, 2013) ont proposé un modèle standard *mention-pair* auquel ils ont défini un modèle alternatif qui, suivant le type grammatical des paires de mentions, les sépare afin de ne conserver que les paires coréférentes. Pour ce faire, les auteurs ont utilisé des stratégies heuristiques standards, notamment *closest-first* et *best-first* de (Ng et Cardie, 2002) pour créer les clusters. Le système a été évalué sur la partie anglaise du corpus CoNLL-2012 Shared Task (Pradhan *et al.*, 2012) qui compte plus de 1,3 millions de mots. La meilleure version du système atteint 67,2% de performances[18], ce qui le classe parmi les meilleurs systèmes de résolution de la coréférence testés sur ce corpus.

2.1.5 Bilan

Trois grandes catégories d'approches par apprentissage supervisé se sont succédées pour résoudre la coréférence : *mention-pair*, *mention-ranking* et *entity-mention*. Même si les systèmes *mention-ranking* ou *entity-mention* ont été proposés pour résoudre les problèmes majeurs rencontrés par les modèles *mention-pair*, peu d'entre eux sont parvenus réellement à améliorer les performances obtenues par les modèles *mention-pair*. Par exemple, le modèle *entity-mention* de (Yang *et al.*, 2004 ; 2008) obtient marginalement de meilleurs résultats que le modèle de base de (Soon *et al.*, 2001) sur un corpus de documents médicaux. Les systèmes hybrides tel que celui de (Rahman et Ng, 2009) semblent être parvenus à un bon compromis.

Nécessitant moins de données d'apprentissage que les systèmes par apprentissage supervisé, des systèmes par apprentissage non supervisé ont été proposés pour résoudre la coréférence.

2.2 Systèmes non supervisés

Moins nombreux[19] que les systèmes supervisés, les systèmes par apprentissage non supervisé ont été proposés afin de réduire la dépendance des algorithmes de résolution de la coréférence en données d'apprentissage, d'une part, et afin

[18] Cette performance est le résultat de la moyenne des métriques MUC, CEAF et B^3 (voir chapitre 8).

[19] « While potentially more appealing, unsupervised learning is very challenging, and unsupervised coreference resolution systems are still rare to this date. » (Poon et Domingos, 2008 : 651).

d'étendre le domaine d'application des algorithmes à divers genres textuels (*i.e.* pour traiter d'autres genres que les genres journalistiques), d'autre part (Lang *et al.*, 2009). Pour ce faire, des algorithmes permettant de combiner une faible quantité de données annotées avec une forte quantité de données non annotées ont été mis en place. N'ayant pas accès à l'information concernant la coréférentialité entre deux mentions, ces systèmes exploitent pour la plupart des informations morpho-syntaxiques, la position ou la distance des mentions et le type de mention. Nous proposons ci-après une vue de plusieurs de ces systèmes.

2.2.1 L'approche pronominale de (Ge *et al.*, 1998)

(Ge *et al.*, 1998) figurent parmi les premiers à avoir proposé une approche non supervisée, mais cette approche est dirigée uniquement vers la résolution des anaphores pronominales (*i.e. he*, *she* et les emplois anaphoriques du *it*). Leur algorithme est composé de deux modules :

- le premier module extrait une série de traits (*e.g.* la distance entre le pronom et ses antécédents potentiels, le genre/nombre/caractère animé de l'antécédent potentiel, la tête lexicale de l'antécédent et le nombre d'occurrences d'un antécédent potentiel dans le texte (*mention-count*)) à partir d'un corpus d'apprentissage – une portion du corpus annoté Penn Treebank[20] (Marcus *et al.*, 1993) de 94 000 mots environ (soit 4 000 phrases) – permettant de déduire des statistiques. Ainsi, le module calcule les probabilités qu'un groupe nominal soit l'antécédent potentiel d'un pronom donné en attribuant un poids à chacun des traits ;
- le second module utilise ces probabilités pour résoudre les anaphores pronominales du corpus de test (*i.e.* non annoté en anaphores).

L'utilisation du trait de répétition lexicale *mention-count* permet au système de (Ge *et al.*, 1998) d'identifier plus facilement le thème d'un segment particulier du discours (*i.e.* l'antécédent), suivant un des principes de la théorie du Centrage selon lequel le thème continu constitue le candidat optimal pour une reprise anaphorique.

(Ge *et al.*, 1998) ont testé de manière incrémentale l'influence de chacun des traits dans l'amélioration de leur système. Ils ont obtenu des performances de 82,9% grâce à cette validation croisée. Appliquant une méthode non supervisée permettant d'acquérir des informations sur le genre à partir de données non

[20] http://www.cis.upenn.edu/~treebank/

annotées (corpus *Wall Street Journal* de 21 millions de mots), les auteurs ont amélioré les performances de leur système à 84,2%.

2.2.2 Approches en *clustering*

2.2.2.1 Le modèle de (Cardie etWagstaff, 1999)

Le modèle de (Cardie et Wagstaff, 1999) constitue le premier système non supervisé proposé pour la tâche de résolution de la coréférence. Dans leur approche, les auteurs considèrent la résolution de la coréférence comme une tâche de *clustering* dans laquelle les mentions sont regroupées en classes d'équivalence (*i.e.* des chaînes de coréférence). Le système ne nécessite pas de données d'apprentissage et il est indépendant du domaine.

D'abord, les mentions du texte sont extraites puis elles sont représentées par une série de 11 traits (les mots contenus dans la mention, la tête nominale, la position de la mention dans le texte, le type de pronom, la définitude/l'indéfinitude, la syntaxe, la catégorie grammaticale des expressions référentielles, le genre, le nombre, le caractère animé et la classe sémantique) et une distance est enfin calculée entre chaque paire de mentions. Si la distance entre deux mentions est inférieure au rayon r de *clustering* (*e.g.* une valeur seuil), alors elles sont considérées comme coréférentes et font donc partie du même *cluster*.

Lors de l'utilisation du système, les mentions sont comparées deux à deux en commençant par la fin du texte. Si les mentions sont jugées compatibles, elles intègrent le même *cluster*. Deux classes d'équivalence peuvent être fusionnées si elles ne comportent pas de mentions incompatibles.

(Cardie et Wagstaff, 1999) ont évalué leur système sur le corpus MUC-6 et ont obtenu des performances de 53,6% (ils se sont classés dans la moyenne des systèmes mis en compétition). Les erreurs rencontrées par le système proviennent en partie d'un manque de reconnaissance fine des noms propres, des rôles thématiques et grammaticaux ainsi que l'absence de trait permettant de filtrer les emplois impersonnels du pronom « *it* ». (Wagstaff, 2002) a amélioré le système initial en ajoutant plus de contraintes linguistiques pour juger de la compatibilité ou de l'incompatibilité des mentions au cours du processus de *clustering*.

2.2.2.2 Le modèle de (Haghighi et Klein, 2007)

Le système de (Haghighi et Klein, 2007) a été considéré comme le meilleur système non supervisé pour la résolution de la coréférence, parvenant non loin des

performances obtenues par les modèles supervisés (Ng, 2008 ; Poon et Domingos, 2008).

Pour résoudre la coréférence, (Haghighi et Klein, 2007) utilisent un modèle bayésien[21] non paramétrable, basé sur un processus hiérarchique de Dirichlet (Teh *et al.*, 2006). D'après (Clark et Gonzàles-Brenes, 2008), un processus de Dirichlet est souvent expliqué en utilisant la métaphore du restaurant chinois (voir Figure 25) : supposons un restaurant chinois possédant une infinité de tables, chacune des tables pouvant accueillir une infinité de clients et servant le même plat à tous les clients de la table. Les tables sont circulaires afin que l'ordre des clients installés ne gêne pas le processus. Un premier client entre dans le restaurant et s'assoit à une table. Lorsqu'un second client entre dans le restaurant, il a alors la possibilité de s'asseoir à la table du premier client ou bien de s'installer à une table vide et ainsi commander un menu pour la table entière. Si l'on généralise le processus, le n-ième client entrant dans le restaurant (cas (a)) peut soit s'asseoir à une table déjà occupée, avec une probabilité proportionnelle au nombre de clients déjà installés sur cette table (cas (b)), soit s'installer à une table vide (cas (c)).

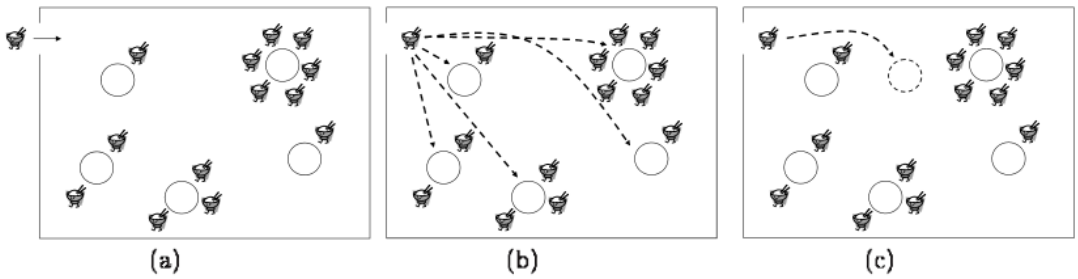


Figure 25 - Métaphore du restaurant chinois, d'après (Caron, 2006 : 84)

L'utilisation du processus de Dirichlet pour la résolution de la coréférence offre la possibilité de ne pas préciser le nombre de *clusters* attendus (*i.e.* le nombre de chaînes de coréférence).

Au cœur du système de (Haghighi et Klein, 2007) se trouve un modèle mixte comportant une série de traits linguistiques tels que le type de la mention (*i.e.* nom propre, pronom ou groupe nominal), la tête lexicale et la saillance de la mention. Pour améliorer la reconnaissance des pronoms, un modèle pronominal a aussi été ajouté, comportant des traits grammaticaux et sémantiques (genre, nombre, type sémantique (*e.g.* organisation, lieu, personne)) ainsi que la récence.

[21] Un modèle bayésien permet de créer des inférences : en fonction des informations observées, le système calcule la probabilité des données non observées. Par exemple, si un patient a le nez pris, on pourra calculer les probabilités qu'il ait un rhume mais aussi qu'il ait mal à la gorge et qu'il ait la fièvre (symptômes non observés) et en déduire des examens complémentaires à effectuer.

Evalué sur le corpus MUC-6, le système a obtenu des performances de 70,6%, ce qui le classe quasiment au même niveau que les systèmes supervisés.

(Poon et Domingos, 2008) et (Ng, 2008) ont modifié le modèle de (Haghighi et Klein, 2007) pour corriger les principaux manques rencontrés par le modèle. En effet, les appositions n'étant pas prises en compte dans le modèle initial de (Haghighi et Klein, 2007), (Poon et Domingos, 2008) ont étendu le modèle pour les identifier. Aussi, plutôt que de modéliser la saillance des pronoms, (Poon et Domingos, 2008) ont imposé une priorité à la distance (en nombre de mentions) entre le pronom et ses antécédents potentiels.

De son côté, (Ng, 2008) a apporté des modifications au modèle initial de (Haghighi et Klein, 2007), améliorant ses performances de 8 à 16%. Par exemple, le module basé initialement sur la tête lexicale du groupe nominal vérifie à présent que les deux candidats sont strictement similaires, ou sont des *alias* ou sont apposés (ces traits supplémentaires permettent d'éviter le rapprochement de paires incompatibles telles que « le grand immeuble » et « le petit immeuble »). Aussi, le calcul de saillance des mentions a été exclusivement réservé aux pronoms, afin d'améliorer la précision du système.

2.2.3 L'approche par rôle contextuel de (Bean et Riloff, 2004)

De leur côté, (Bean et Riloff, 2004) ont proposé un système d'apprentissage non supervisé utilisant le rôle contextuel pour évaluer les antécédents potentiels d'une anaphore donnée. Dans leur approche, un rôle contextuel représente le rôle joué par un antécédent dans une action ou un événement (*e.g.* dans « Paul a fait griller le steak. Il était bien saignant. », la résolution correcte de l'anaphore « il » dépend des connaissances suivant lesquelles seul le steak peut être saignant).

Utilisant des techniques issues de l'extraction d'information, le système représente et apprend les rôles contextuels des antécédents à partir d'un corpus de grande taille non annoté (le corpus se compose du corpus MUC-4 et d'articles de *Reuter's* de 1996-1997). Chaque patron contextuel représente le rôle joué par une expression référentielle dans le contexte environnant. Les exemples d'entraînement sont générés automatiquement en détectant des couples anaphores-antécédents facilement identifiables *via* des heuristiques lexicales (*e.g.* les noms propres et leurs variantes, les descriptions définies) et syntaxiques (*e.g.* appositions, propriétés d'une mention (p.e. « *Paul* est *le président*… »)). Puis, à partir des exemples d'entraînement, le système calcule des statistiques permettant

de prédire qu'un antécédent et une anaphore puissent être coréférents. Pour ce faire, 11 modules contextuels (lexicaux, syntaxiques, sémantiques, genre, nombre, distance) sont testés. Par exemple, pour le module lexical, le système :

- identifie le patron contextuel permettant d'extraire le candidat antécédent,
- assemble ce patron avec le patron contextuel identifiant l'anaphore,
- vérifie si les deux patrons sont des cooccurrents (*e.g.* s'ils sont synonymes, comme dans « le meurtre de X » et « X est assassiné »).

Si tel est le cas, le système prédit que l'anaphore et l'antécédent sont coréférents.

A partir de ces prédictions, le système crée des patrons d'équivalence lui permettant d'associer plusieurs mentions pour créer des chaînes de coréférence. Le système a obtenu des performances moyennes de 62%, révélant un impact plus important du rôle contextuel sur la résolution des pronoms (+13%) que sur celle des groupes nominaux définis.

2.2.4 L'approche par hypergraphe de (Lang *et al.*, 2009)

Pour résoudre la coréférence, (Lang *et al.*, 2009) ont proposé un algorithme de partitionnement par hypergraphe qui répartit les mentions directement en classes d'équivalence (*i.e.* en chaînes de coréférence). Un hypergraphe est un graphe spécifique dans lequel une arête peut connecter plus de deux sommets (Berge, 1989). Par exemple, dans l'hypergraphe du plan de Fano (voir Figure 26), on compte 7 sommets (0, 1, 2, 3, 4, 5, 6) et sept arêtes (615, 523, 504, 643, 103, 206, 124).

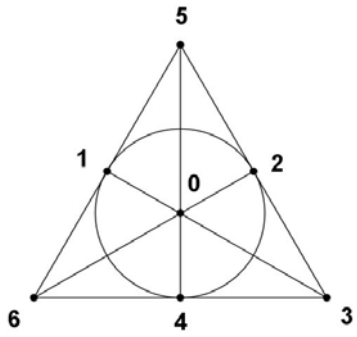


Figure 26 - Hypergraphe du plan de Fano[22]

Dans l'approche de (Lang *et al.*, 2009), pour modéliser la résolution de la coréférence, les mentions sont considérées comme des sommets. Si un groupe de mentions possède une propriété commune spécifique, alors il est couvert par une arête. De cette manière, l'hypergraphe permet de récupérer de l'information sur plusieurs mentions simultanément.

(Lang *et al.*, 2009) ont défini 12 arêtes dans leur modèle : des arêtes de base (identité des mentions, identité de la tête lexicale des mentions, genre, nombre, type de nom propre, *alias*, apposition), ainsi que des arêtes sélectionnant un pronom et les mentions le précédant (dans la phrase précédente, dans les deux phrases précédentes ou les trois phrases précédentes). Chacune des arêtes possède un score arbitraire (allant de 0 à 30) suivant son niveau de pertinence dans l'identification d'une relation de coréférence. Par exemple, l'arête *alias* possède un score de 30 (arête pertinente) tandis que l'arête sélectionnant les mentions dans une fenêtre de trois phrases possède un score de 5.

Lors de l'utilisation du système, les mentions sont d'abord placées dans un même espace. Puis, les mentions couvertes par une arête sont regroupées dans des clusters (*e.g.* les mentions m1 et m3 sont des *alias* (ces mentions forment un premier *cluster*), les mentions m1 et m5 ont le même nombre (ces mentions forment un autre *cluster*)). Le processus s'arrête lorsqu'une valeur seuil est atteinte ou lorsqu'il ne reste plus qu'une seule mention dans l'espace initial. Seules les mentions contenues dans des arêtes possédant un score élevé sont alors retenues. Les mentions issues d'un même *cluster* sont alors jugées coréférentes.

Evalué sur le corpus ACE-2, le système de (Lang *et al.*, 2009) obtient de meilleures performances que les systèmes non-supervisés le précédant (+1,97%), mais n'atteint pas les performances des systèmes supervisés (-2,57%).

[22] http://commons.wikimedia.org/wiki/File%3AFano_plane.jpg

2.2.5 Bilan

Le bref état des systèmes de calcul de la référence basés sur des méthodes par apprentissage, qu'elles soient supervisées (Soon *et al.*, 2001 ; Ng et Cardie, 2002) ou non supervisées (Cardie et Wagstaff, 1999 ; Haghighi et Klein, 2007), nous a permis de prendre connaissance de la diversité des méthodes proposées pour résoudre la coréférence. Nous avons aussi pu constater l'étendue de la présence des systèmes supervisés (modèles *mention-pair*, *mention-ranking* et *entity-mention*) face aux systèmes non supervisés, bien que ces derniers ne nécessitent pas de données d'apprentissage sur les relations de coréférence. La raison de cette différence relève essentiellement de la complexité dans la mise en œuvre des méthodes non supervisées relativement aux performances obtenues (Poon et Domingos, 2008).

Mais, bien que les résultats obtenus par les systèmes non supervisés soient encore peu satisfaisants par rapport à ceux des systèmes supervisés (Ng, 2008), les récents résultats obtenus par les systèmes par apprentissage non supervisé tels que (Haghighi et Klein, 2007, 2010[23]) pourraient mettre en partie en doute l'efficacité de l'utilisation des données annotées par les systèmes supervisés (Recasens, 2010).

Néanmoins, comme le suggère (Ng, 2010), la comparaison des divers systèmes de résolution de la coréférence serait facilitée si tous ces systèmes étaient librement accessibles, s'ils utilisaient les mêmes corpus et les mêmes mesures d'évaluation.

2.3 Discussion

Pour combler les limites rencontrées par les systèmes à base de règles (poids attribués à la main et impossibilité de formuler des inférences), les systèmes par apprentissage supervisé ou non-supervisé ont été proposés pour résoudre la référence. Mais, bien que performantes, ces méthodes d'apprentissage ont aussi montré leurs limites, que (Lee *et al.*, 2013 : 2) résument ainsi :

> « Supervised machine learning systems rely on expensive hand-labeled data sets and generalize poorly to new words or domains. Unsupervised systems are increasingly more complex, making them hard to tune and difficult to apply to new problems and genres as well. »

[23] Même si le système de (Haghighi et Klein, 2010) n'est pas totalement non-supervisé.

Afin de tenter de pallier ces faiblesses et d'améliorer les performances pour la résolution de la coréférence, des systèmes hybrides mêlant les connaissances issues des méthodes par apprentissage et des règles symboliques ont été mis en place. Par exemple :

- pour la résolution de la coréférence en allemand, (Hartrumpf, 2001) a combiné des règles syntactico-sémantiques à des règles statistiques issues d'un corpus annoté en coréférence (articles issus de *Süddeutsche Zeitung*). Les règles syntactico-sémantiques permettent d'identifier les éventuelles relations de coréférence entre antécédents et anaphores, le corpus annoté est utilisé pour hiérarchiser d'autres paires possibles *via* des probabilités. Le système hybride CORUDIS (COreference RUles with DIsambiguation Statistics) a obtenu des performances de 66% qui s'alignent sur celles des systèmes anglais (60%),
- à partir du corpus MUC-7 annoté en relations de coréférence, préalablement analysé syntaxiquement et annoté en entités nommées, (Uryupina, 2006, 2007) a largement étendu le nombre de traits utilisés par les méthodes par apprentissage et par règles. L'autrice a proposé une série de 351 traits syntaxiques, sémantiques, lexicographiques et discursifs (saillance). L'évaluation de la méthode a montré une amélioration des performances de différents algorithmes de l'état de l'art uniquement en utilisant un corpus d'apprentissage très large,
- pour résoudre les anaphores en anglais (*i.e.* pronom « it » uniquement), (Weissenbacher et Nazarenko, 2007 ; Weissenbacher, 2008) ont proposé une modélisation reposant sur les réseaux bayésiens. L'approche probabiliste utilisée permet de regrouper, dans une même représentation, des connaissances linguistiques (*i.e.* annotation manuelle des emplois anaphoriques du pronom « it » et de leurs antécédents) et des indices de surface (*i.e.* des heuristiques). Le classifieur recherche d'abord l'antécédent le plus saillant qui précède l'anaphore grâce à une série d'indices inspirés des approches *knowledge-poor*. Puis ces indices sont complétés grâce à l'utilisation des annotations effectuées sur le corpus d'entraînement. Evalué sur un corpus de 697 occurrences de « it » issues de résumés d'articles de recherche génomique, leur approche obtient un taux de succès de 61% mais elle rencontre de nombreuses erreurs liées à un calcul erroné de la saillance des antécédents,
- (Lee *et al.*, 2013) ont récemment proposé un système de résolution de la coréférence qui combine un algorithme par apprentissage (du type *entity-mention*) avec une série de règles inspirées du modèle à haute précision de (Baldwin, 1997). Le système procède en une succession de 10 filtres

relativement fins pour identifier les expressions référentielles coréférentes (noms propres et *alias*, noms propres et pronoms, groupes nominaux identiques, groupes nominaux ayant la même tête lexicale, appositions, acronymes) du corpus d'entraînement. Un module de résolution des anaphores pronominales est ensuite utilisé pour résoudre les anaphores restantes. L'algorithme de (Lee *et al.*, 2013) a été évalué sur 3 corpus différents (ACE-4, MUC-6 et OntoNotes[24]). Les performances obtenues sont équivalentes, sinon supérieures aux systèmes de l'état de l'art. L'utilisation du module symbolique après le module par apprentissage a permis de gagner 9.5 points, ce qui prouve l'impact de l'utilisation des techniques hybrides pour la résolution de la coréférence. Le système de (Lee *et al.*, 2013) a été implémenté dans des systèmes de résolution de la coréférence en anglais, chinois et arabe.

Ainsi, l'utilisation de méthodes hybrides symboliques-statistiques semble être un bon compromis pour améliorer les performances de la résolution des relations anaphoriques et de coréférence. Néanmoins, ces méthodes hybrides sont encore dépendantes en ressources annotées, ce qui peut constituer un frein pour certaines langues comme le français. De plus, nous avons vu que la plupart des méthodes symboliques utilisent des règles contextuelles relativement dépendantes du domaine et que les méthodes par apprentissage reposent sur l'utilisation de corpus de grande taille, préalablement annotés en relations de coréférence et homogènes dans leur contenu. De ce fait, les outils développés pour la résolution de la référence sont, pour la plupart, dépendants de la quantité et de la qualité des données fournies pour l'apprentissage, mais aussi du domaine. Nous proposons dans la section suivante de revenir sur ces limites.

[24] http://www.ldc.upenn.edu/Catalog/docs/LDC2011T03/OntoNotes-Release-4.0.pdf

3 Calcul de la référence : lacunes

Malgré les efforts fournis ces dernières années pour améliorer les performances des systèmes de résolution de la référence, force est de constater, à l'instar de (Recasens, 2010), que les systèmes actuels n'obtiennent pas de résultats satisfaisants : « Despite the wealth of existing research, current performance of coreference resolution systems has not reached a satisfactory level. » (Recasens, 2010 : vii).

Nous pensons que ces systèmes souffrent de plusieurs lacunes, parmi elles, l'absence de prise en compte de tous les types d'expressions référentielles intervenant dans le calcul de la référence (section 1) et la dépendance au domaine ou à un genre textuel donné (section 2). Plus spécifiquement pour le français, l'absence d'un corpus de référence large annoté en relations de coréférence limite fortement l'utilisation des techniques d'apprentissage automatique qui participeraient à l'amélioration des systèmes symboliques de calcul de la référence (section 3).

3.1 Non exhaustivité des expressions référentielles annotées

A l'issue de la présentation des différentes méthodes symboliques pour la résolution de la référence, nous avons pu observer que la plupart des systèmes se focalisaient sur la résolution des anaphores pronominales (*e.g.* Hobbs, 1978 ; Lappin et Leass, 1994 ; Kennedy et Boguraev, 1996 ; Mitkov, 1998). Ainsi, l'objectif de ces systèmes est de rattacher un pronom avec un antécédent qui est souvent un groupe nominal défini. Rares sont les systèmes présentés permettant de constituer des paires coréférentes contenant par exemple un nom propre avec une description définie (*e.g.* « François Hollande … le président de la République Française ») ou un groupe nominal indéfini avec un groupe nominal défini (*e.g.* « un Etat-membre … l'Etat-membre »), bien que ces paires soit relativement courantes en corpus. De ce fait, il est important de mettre l'accent sur la distinction entre les systèmes de calcul des anaphores (*i.e.* où il existe une relation de dépendance interprétative du pronom par rapport à son antécédent par exemple) des systèmes de calcul de la coréférence (*i.e.* où l'interprétation de l'élément repris est indépendante de son antécédent). De plus, il faut distinguer les systèmes identifiant uniquement des paires antécédent-anaphore (*i.e.* ayant

pour objectif le rattachement d'une anaphore donnée avec son antécédent uniquement) de ceux identifiant toutes les expressions référant à la même entité du discours (*i.e.* construisant des chaînes de référence). La terminologie anglaise est, à cet égard, relativement confuse.

D'autre part, parmi les approches par apprentissage que nous venons de présenter, il apparaît que plusieurs systèmes, bien qu'annonçant résoudre la coréférence, ne se limitent en réalité qu'à la résolution de certains types d'expressions référentielles. Cela peut être causé par le type de corpus utilisé ou par la nature de la tâche effectuée.

Concernant le type de corpus, (Recasens, 2010) montre que les corpus ACE sont restreints en types d'expressions référentielles. En effet, dans les corpus ACE-2004/2005 par exemple, la coréférence est réduite aux relations établies entre 7 types d'entités : personnes, organisations, entités géopolitiques, lieux, constructions humaines, véhicules et armes. Ceci s'explique par le fait que le corpus ACE, de même que le corpus MUC, ont été constitués à l'origine pour servir les tâches d'évaluation en extraction d'information, ils n'ont donc pas été créés spécifiquement pour la résolution de la coréférence (*i.e.* la résolution de la coréférence n'en est qu'un moyen). De ce fait, les systèmes par apprentissage qui utilisent uniquement ces corpus comme base d'apprentissage des relations de coréférence (*e.g.* pour ACE : (Daumé III et Marcu, 2005 ; Denis et Baldridge, 2008 ; Lang *et al.*, 2009 ; Rahman et Ng, 2009)) ne seront pas à même d'identifier tous les types d'expressions référentielles coréférentes si on les soumet à un autre corpus.

Aussi, certaines tâches proposées dans les campagnes d'évaluation, telle que la tâche de suivi d'entité (*Entity Detection and Tracking*) présente notamment dans le cadre des campagnes ACE, ne consiste pas en la résolution de la coréférence à proprement parler, mais elle cherche plutôt à reconnaître les variantes possibles d'une même entité (Ludovic, 2011), c'est-à-dire ses variations orthographiques, ses *alias*, etc.

Pourtant, la résolution automatique de la coréférence consiste à identifier dans le texte toutes les expressions référant à la même entité du discours, soit les noms propres, les pronoms, les groupes nominaux (définis, démonstratifs, indéfinis), les possessifs, etc. De ce fait, l'utilisation de ces systèmes ne permet que de résoudre partiellement la coréférence et ne permet pas d'identifier réellement les chaînes de référence dans les textes, tel que nous souhaitons l'effectuer dans notre système de détection de thèmes.

3.2 Limitation du genre textuel traité

Un autre problème majeur, à nos yeux, pour la résolution de la référence, tient à la limitation au nombre et au type de genres textuels traités par les systèmes de calcul de la référence. En effet, (Recasens, 2010), reprise par (Muzerelle *et al.*, 2013) montre que la plupart des corpus de référence (de plus de 200 000 mots) utilisés pour la mise en place des systèmes de résolution de la référence ne comptent, pour la majorité, que des textes issus des genres journalistiques et ce, quelle que soit la langue (voir Tableau 22).

Langue	Corpus	Genre
allemand	TüBa-D/Z	*News* = journaux d'information radio-diffusés (transcription de l'oral)
anglais	ARRAU, Switchboard, ACE, SemEval OntoNotes	*News*, weblog, forum, chat, récit oral, conversation téléphonique
arabe	ACE	*News*
catalan	AnCora-Ca	*News*
chinois (mandarin)	ACE, OntoNotes	*News*
espagnol	ACE, Ancora-Es	*News*
italien	I-CAB	*News*
japonais	NAIST Text	*News*
néerlandais	COREA	*News*, oral, texte encyclopédique
tchèque	PDT	*News*

Tableau 22 – Etat actuel des corpus annotés manuellement en coréférence de plus de 200 000 mots, d'après (Recasens, 2010 : 10) repris dans (Muzerelle *et al.*, 2013 : 557)

Pourtant, notre étude des chaînes de référence menée sur des genres textuels divers (voir chapitre 3) a révélé des particularités liées au matériau linguistique contenu dans les chaînes, suivant le genre (nombre moyen de maillons d'une chaîne suivant le genre, catégorie grammaticale du maillon initiateur d'une chaîne, distance moyenne entre les maillons, etc.). Cette importance du genre textuel dans la mise en place d'un système de calcul de la référence a été signalée par (Tutin *et al.*, 2000) :

> « Pour mettre au point des stratégies de reconnaissance d'anaphores efficaces et robustes et élaborer des stratégies de génération valides et réalistes, il paraît donc indispensable de s'appuyer sur des productions réelles, prenant en compte le genre textuel. » (Tutin *et al.*, 2000 : 2)

La limitation du genre du corpus d'apprentissage constituerait alors un frein aux performances des algorithmes développés, qui, face à un genre textuel « inconnu »

(*p.e.* un article de loi), éprouveraient des difficultés pour identifier correctement toutes les mentions d'un même référent. A notre connaissance, aucune étude contrastive n'a été effectuée en ce sens pour évaluer et adapter les performances des outils existants face à des textes issus de genres divers.

3.3 Absence de corpus de référence large et libre annoté en coréférence pour le français écrit

Il y a plus de 10 ans déjà, (Salmon-Alt, 2002 : 164) signalait que « les ressources françaises annotées en relations anaphoriques et accessibles librement pour la recherche sont insuffisantes en taille et hétérogènes en ce qui concerne les phénomènes annotés ainsi que les schémas d'annotation ». L'autrice dressait la liste des corpus français comportant des annotations en relations de coréférence (voir Tableau 23).

Auteurs	*Bruneseaux et al. 1997*	*Popescu-Belis 1999*	*Clouzot et al. 2000*	*Tutin et al. 2000*	*Salmon-Alt 2001*	*Trouilleux 2001*
Corpus	*Père Goriot*, (Balzac)	*Vittoria Accoramboni* (Stendhal)	*Le Monde Diplomatique*	différents genres	corpus *Ozkan 1994* (dialogues)	*La Tribune* (finance)
Nb de mots	30.000	10.000	95.000	1.000.000	11.000	45.000
Expressions annotées	GN pour personnages, lieux et objets principaux	tous les GN	pronoms personnels 3ième personne	expressions anaphoriques sauf descriptions définies	tous les GN	pronoms personnels 3ième pers. (sujet et objet), pronoms possessifs 3ième pers.
Nb d'expressions	3359	638	1316	?	1344	886
Liens annotés	coréférence, anaphores associatives	coréférence	coréférence	coréférence, anaphores associatives	anaphores associatives en « *autre* »	coréférence, anaphores associatives
Schéma d'annotation	compatible MATE	MUC - compatible MATE	TEI + format propriétaire	format propriétaire	TEI + compatible MATE	format propriétaire
Accès libre pour recherche	oui	oui	négociable ?	non	oui	non

Tableau 23 – Ressources françaises annotées en relations anaphoriques (issu de (Salmon-Alt, 2002 : 166))

Parmi ces 6 ressources, seul le corpus CRISTAL-GRESEC (Tutin *et al.*, 2000)[25] est de taille conséquente (1 million de mots) et comporte des documents issus de divers genres textuels (vulgarisations scientifiques, analyses économiques,

[25] http://catalog.elra.info/product_info.php?products_id=634

bulletins juridiques). Néanmoins, seules certaines relations anaphoriques sont annotées (les descriptions définies sont exclues) et ce corpus n'est pas disponible librement pour la recherche.

Si l'on se reporte une nouvelle fois au Tableau 22 ci-dessus, force est de constater qu'il n'existe toujours pas, à l'heure actuelle, de corpus large (supérieur à 200 000 mots) annoté en coréférence pour le français écrit, susceptible d'être utilisé comme corpus d'apprentissage automatique (Muzerelle *et al.*, 2013). En effet, parmi les corpus annotés en relations de coréférence en français, on compte le corpus ANANAS[26] (Annotation anaphorique pour l'analyse sémantique de corpus) de (Salmon-Alt, 2002) d'un million de mots issus de genres divers (genres journalistiques, revues scientifiques, compte-rendus administratifs et bulletins juridiques), qui, de même que le corpus FReeBank[27] (Salmon-Alt *et al.*, 2004) de près de 100 000 mots annotés en coréférence, ne sont malheureusement plus disponibles (Hernandez et Boudin, 2013).

De son côté, le corpus DeDé[28] (Gardent et Manuélian, 2005) ne compte que 48 000 mots. Les annotations sont focalisées sur les descriptions définies et démonstratives et le corpus est composé des différents genres issus du journal *Le Monde*.

Récemment, le projet ANNODIS (Péry-Woodley *et al.*, 2009, 2011 ; Afantenos *et al.*, 2012) fournit un corpus de grande taille (687 000 mots) issu de divers genres textuels (brèves, articles encyclopédiques, articles de recherche, rapports). Même si les annotations des structures énumératives sont privilégiées, une partie du corpus (666 000 mots) est aussi annotée en « chaînes topicales » qui sont définies comme des « segments de texte regroupant des phrases reliées par un référent commun » (Péry-Woodley *et al.*, 2011 : 74). Ainsi, 572 chaînes topicales sont-elles annotées dans le sous-corpus. Néanmoins, ces chaînes topicales ne sont pas synonymes de chaînes de (co)référence, puisque peuvent être inclus dans les chaînes topicales des exemples ou des commentaires. Ces chaînes topicales correspondent donc à des portions de texte (qui occupent une taille moyenne de 187 mots dans le corpus) contenant des indices (*i.e.* des expressions référentielles) plutôt qu'à la succession d'expressions référentielles coréférentes (*i.e.* les maillons) à proprement parler. Les liens entre les maillons d'une chaîne ne sont donc pas explicitement marqués dans le corpus. De ce fait, en l'état, le corpus n'est pas directement exploitable pour un système de calcul de la référence par apprentissage.

[26] http://www.inalf.fr/ananas/
[27] http://web.archive.org/web/20090514021120/http://freebank.loria.fr/index.php#etat
[28] http://www.cnrtl.fr/corpus/dede/plus_dede.php

Si l'on regarde du côté des corpus oraux annotés en relations de coréférence, le corpus CO2[29] (Schang *et al.*, 2011) propose, depuis juin 2013, un corpus de 35 000 mots annotés en coréférences et anaphores associatives. Il s'agit d'un extrait de 3 interviews sociolinguistiques issu du corpus ESLO[30]. Cependant, même si ce corpus est annoté en relations de coréférence, il se limite uniquement aux relations établies pour des groupes nominaux ou pronominaux : les possessifs ne sont ainsi pas identifiés comme expressions référentielles (*i.e.*, dans le guide d'annotations, on trouve cet exemple annoté : « Le disque se vend bien (...) Ses acheteurs se comptent par centaines ». Dans cet exemple, seul le possessif « ses » devrait être surligné). De plus, les annotations ne sont pas l'objet principal du projet CO2, elles sont un moyen permettant de comprendre les mécanismes complexes de la référence pour la parole spontanée (Muzerelle *et al.*, 2012). Ce sont justement ces spécificités de l'oral, notamment des disfluences (pauses, hésitations, reformulations, répétitions) ainsi que des chevauchements de paroles, qui rendent difficile l'exploitation, en l'état, d'un tel corpus en vue d'un apprentissage par des algorithmes de résolution de la référence de l'écrit.

Dans la même lignée, en novembre 2013 a été rendu disponible le corpus ANCOR[31] (Muzerelle *et al.*, 2013), annoté en relations de coréférence, de grande taille (418 000 mots), contenant entre autres le corpus C02. Il s'agit aussi d'un corpus de parole (dialogues par téléphone, en présentiel ou entretiens) qui souffre donc des mêmes problèmes liés aux spécificités de l'oral que le corpus C02.

Malgré les efforts fournis dans la communauté francophone ces dernières années, il ne paraît pas envisageable, à l'heure actuelle, d'utiliser les corpus disponibles.

[29] http://www.info.univ-tours.fr/~antoine/parole_publique/CO2/index.html
[30] http://www.lll.cnrs.fr/eslo-1
[31] http://tln.li.univ-tours.fr/Tln_Corpus_Ancor.html

4 Conclusion

Dans ce chapitre, nous avons présenté divers systèmes de calcul de la référence à base symbolique ou par apprentissage. Nous avons relevé que ces systèmes différaient en plusieurs points :

- la technologie utilisée pour implémenter le système (règles symboliques, apprentissage statistique),
- le domaine d'étude et le genre textuel des documents,
- le niveau de résolution proposé (résolution des anaphores pronominales, résolution des anaphores nominales, résolution de la cataphore, résolution de la coréférence) et le type des catégories prises en compte (identification des pronoms personnels de troisième personne uniquement (pour la résolution des anaphores pronominales), possessifs, démonstratifs),
- l'utilisation (ou la non utilisation) de connaissances externes (pré-annotation des emplois impersonnels du pronom *il,* relations ontologiques, synonymes).

Les diverses lacunes mises au jour concernant les systèmes de résolution de la référence, d'une part (restriction des expressions référentielles annotées et fréquente limitation du corpus d'apprentissage aux genres journalistiques) et l'absence d'un corpus de référence large et libre de la coréférence en français écrit, d'autre part, nous mènent à ne pas pouvoir utiliser, à l'heure actuelle, les techniques d'apprentissage pour la résolution de la référence. De plus, le cadre industriel dans lequel est inscrit notre projet ne nous permet pas d'utiliser d'analyseur syntaxique ni de ressources linguistiques complexes. De ce fait, dans notre système de détection automatique de thèmes, le module d'identification automatique des chaînes de référence utilisera des techniques à base de règles et des heuristiques, dans la lignée des approches *knowledge-poor* (voir Partie III).

Néanmoins, les systèmes symboliques n'excluant pas l'apprentissage (Ludovic, 2011), lorsqu'un corpus de référence annoté en coréférence pour le français sera disponible, nous pourrons utiliser les techniques d'apprentissage pour améliorer notre système.

PARTIE III
ATDS-Fr[1], système de détection automatique de thèmes

La détection des thèmes est un problème assez courant en traitement automatique des langues (TAL) et elle fait l'objet de nombreux travaux concernant l'extraction d'informations ciblées (p.e. l'identification des entités nommées) ainsi qu'en veille numérique (p.e. les citations entre brevets (Blanchard, 2009 : 6)[2]. Face au nombre toujours croissant des documents numériques, la méthode de recherche « plein texte » employée par les moteurs de recherche actuels semble s'essouffler. Une méthode innovante basée sur la détection automatique des thèmes contenus dans les documents constitue une solution pour améliorer les performances des moteurs de recherche.

Même si les techniques de segmentation statistiques utilisant des chaînes lexicales (Hearst, 1997), (Choi *et al.*, 2001) se révèlent efficaces pour identifier les zones de rupture thématique (des zones de texte moins cohésives), elles ne proposent pas de liste des thèmes associés aux segments délimités. Pour effectuer cette tâche, des systèmes hybrides mêlant des méthodes statistiques à des connaissances linguistiques tels que ceux de (Ferret *et al.*, 2001 ; Hernandez, 2004), ont été mis en place (*cf.* chapitre 4). Nous nous situons dans cette lignée.

Dès lors, pour détecter les thèmes des documents nous combinons dans notre modèle les résultats de segmentation thématique fournis par l'algorithme

[1] ATDS-Fr (*Automatic Topic Detection System for French*), système de détection automatique de thèmes pour le français.

[2] La problématique de détection de thèmes est courante, même si nous avons vu que la finalité de cette tâche diffère selon le système (recherche de frontières thématiques, attribution de thèmes suivant une liste préétablie, etc., *cf.* chapitre 4).

statistique *C99* enrichi de la LSA[3] (*i.e. CWM*, Choi *et al.*, 2001, *cf.* chapitre 4) à des informations linguistiques relatives à la cohésion (et la cohérence) textuelle (cadres de discours de (Charolles, 1997) et marqueurs référentiels). Alors que (Hernandez, 2004) a proposé une méthode de détection de thèmes basée sur la résolution d'anaphores (système *SRA*[4]), nous nous appuyons aussi sur l'analyse des chaînes de référence pour identifier les thèmes des documents[5].

Dans le chapitre 6, nous présentons l'architecture hybride statistique – linguistique d'ATDS-Fr adoptée pour mener à la détection automatique des thèmes. Dans le chapitre 7, nous nous focalisons sur un des modules de notre méthode linguistique, le module d'identification automatique des chaînes de référence (*RefGen*). Nous présentons l'outil *RefGen* que nous avons mis en place ainsi que ses sous-modules *RefAnnot* (pour l'annotation des diverses expressions référentielles) et *CalcRef* (pour le calcul de la référence). Dans le chapitre 8, nous procédons à l'évaluation des modules de *RefGen*.

[3] Même si la version améliorée de *C99* a été rebaptisée *CWM*, nous nommerons toujours cet algorithme *C99* dans la suite de ce mémoire.

[4] *SRA* : Système de Résolution d'Anaphore, voir chapitre 5 section 1.3.2.

[5] En effet, (Hernandez, 2004 : 115) se limite à identifier uniquement les « expressions nominales introduites par un démonstratif et les pronoms personnels à la troisième personne ».

Chapitre 6

Description du système de détection automatique de thèmes (ATDS-Fr)

Dans ce chapitre, nous présentons le système automatique ATDS-Fr développé pour détecter les thèmes dans les documents. ATDS-Fr permet d'effectuer, en parallèle à la recherche « plein texte » classique des moteurs de recherche, une recherche par thèmes. Pour cela, l'indexation par mots-clés est doublée d'une indexation des documents par thème. Comme nous le verrons, l'architecture d'ATDS-Fr est une architecture hybride combinant des méthodes statistiques et linguistiques pour mener à la détection des thèmes. Cette méthode mixte s'inscrit dans les préoccupations actuelles du TAL, telles que discutées lors de l'atelier MIXEUR « Méthodes mixtes pour l'analyse syntaxique et sémantique du français » (Retoré et *al.*, 2013) de la conférence TALN 2013. Nous présenterons une vue générale du système avant de détailler chacune des deux composantes d'ATDS-Fr. A partir du découpage statistique, nous proposons des combinaisons

entre les catégories de marqueurs linguistiques sélectionnés (marqueurs lexicaux cadratifs et chaînes de référence) pour mener à la détection automatique des thèmes que nous visons.

1 Architecture générale du système

Le système de détection automatique de thèmes ATDS-Fr combine un module statistique et un module symbolique[1] identifiant automatiquement des indices linguistiques (voir Figure 27).

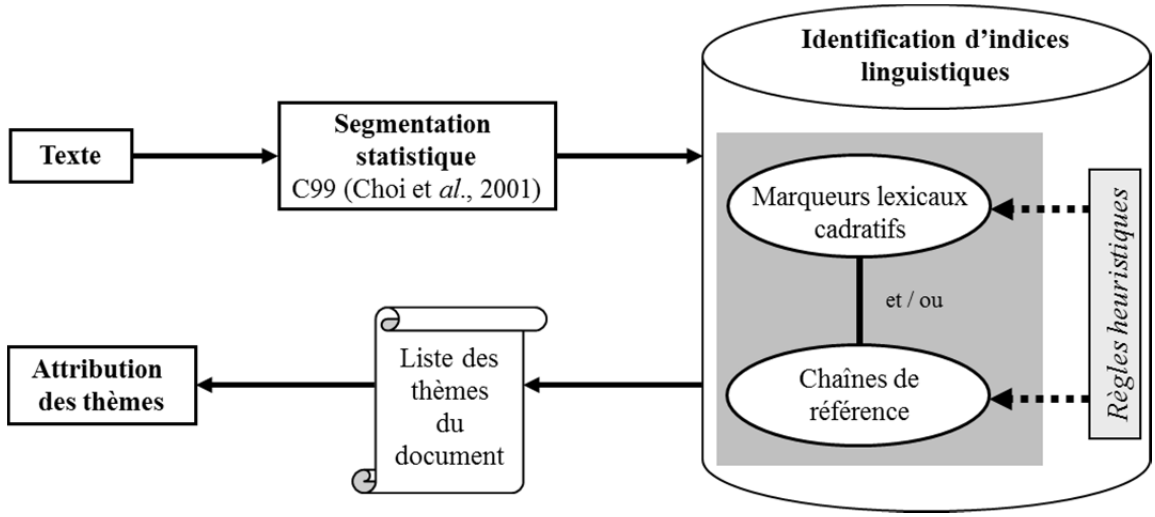


Figure 27 - Architecture globale du système ATDS-Fr

Au départ, le texte brut est extrait à partir de documents provenant de divers formats (PDF, Office ou HTML). A l'issue de l'extraction du texte, les documents sont découpés en segments thématiquement homogènes (*e.g.* des segments proches d'un point de vue thématique) par l'algorithme de segmentation par cohésion lexicale *C99* (Choi *et al.*, 2001)[2]. Cependant, bien que cet outil de segmentation statistique fournisse un découpage des textes en segments thématiquement homogènes, en fonction de la distribution et des fréquences de mots, il n'indique pas quels sont les thèmes de chaque segment.

Pour déterminer les thèmes des segments, nous nous appuyons sur des marqueurs linguistiques de cohésion et de cohérence qui constituent des instructions données

[1] Un module symbolique est un système d'analyse à base de règles.

[2] Se reporter au chapitre 4 pour une présentation de l'algorithme *C99* (Choi, 2000) et de sa version améliorée *CWM* (Choi *et al.*, 2001).

au lecteur sur la structure des documents. Nous utilisons différents marqueurs lexicaux cadratifs : les introducteurs de cadre du discours de (Charolles, 1997) et les marqueurs référentiels constituant les chaînes de référence (Corblin, 1995a, 1995b ; Schnedecker, 1997 ; Cornish, 1999).

Afin d'identifier automatiquement ces indices linguistiques, plusieurs classes de règles heuristiques, définies pour chaque catégorie de marqueurs, sont appliquées séparément ou simultanément, pour détecter les thèmes possibles de chaque segment. L'absence de certains marqueurs peut être compensée par l'application des autres catégories de marqueurs (comme nous le verrons dans la section 4). La sortie de ce module se présente sous la forme d'une liste de thèmes associés à chaque segment.

L'architecture générale d'ATDS-Fr exposée, nous présenterons les modules statistique et linguistique du système dans les sections suivantes.

2 Le module statistique

Le module statistique est le premier traitement automatique appliqué au texte extrait des documents. L'objectif de cette étape est de découper le texte en segments thématiquement homogènes du point de vue de leur contenu. La technique utilisée s'appuie sur la cohésion lexicale. Comme nous avons pu le mettre en évidence au chapitre 4 lors de la présentation et de l'évaluation des divers outils statistiques de segmentation thématique, notre choix s'est porté sur l'algorithme *C99* enrichi de la LSA (Choi *et al.*, 2001) qui obtient les meilleurs résultats par rapport à l'état de l'art.

A partir du texte brut, l'outil *C99* segmente de manière linéaire le corpus en fonction des proximités sémantiques (*e.g.* le fait que des segments textuels possèdent un matériau lexical similaire) établies entre les unités. A l'issue de ce traitement statistique, *C99* ajoute des doubles tirets (==========) pour séparer chaque segment thématique. Un exemple de sortie, issu de notre corpus de rapports publics (traitant de la satisfaction des usagers), est présenté ci-après (voir Figure 28). Dans cet extrait, *C99* a marqué deux ruptures thématiques : la première, après le titre « la mesure de la satisfaction des clients », la seconde avant le titre de la section I « les réformes de l'Etat ont longtemps présumé les attentes des usagers ».

La première zone thématique s'articule autour de la notion de « satisfaction » qui est mesurée dans le secteur privé (puisque l'on parle de « clients ») mais pas dans les services publics (où l'on emploie plutôt le terme d' « usagers »). Cet écart entre les préoccupations des deux secteurs privé/public tend à disparaître, ce qui amène à présent les services publics à s'intéresser à la mesure de la satisfaction de leurs usagers.

La seconde zone thématique explique les raisons qui ont bloqué la prise en compte des attentes des usagers des services publics.

LA MESURE DE LA SATISFACTION DES USAGERS, UN ENJEU RECENT
==========
La mesure de la satisfaction des clients trouve naturellement sa place dans la finalité des entreprises exerçant dans des secteurs concurrentiels.
Pour conserver des parts de marché ou les accroître, celles-ci doivent connaître la manière dont se forme la satisfaction des clients à l' égard de leurs produits ou de leurs services.
A l'inverse, la sanction d'une insatisfaction des clients intervient de façon directe, par une diminution du chiffre d'affaires.

La place accordée à la satisfaction et donc à la fidélisation des clients prend une importance de plus en plus sensible à la faveur de la mutation d'une société de production de biens de masse à une société de production de services.
Les usagers des services publics ne sont pas, bien sûr, toujours assimilables à des clients. Ils n'ont souvent pas le choix de refuser un produit ou un service.
A la différence des entreprises dont le mode principal d'action est la conviction, les services publics peuvent également faire usage de la coercition.
En outre, d'autres principes que celui de la recherche de la satisfaction individuelle de tel ou tel usager inspirent l'action publique, au premier rang desquels figure celui de l'égalité devant le service public.
De ce fait, la recherche de la satisfaction des usagers se présente différemment selon le type de service public.
On constate ainsi sans surprise que les services publics dont le mode d'action privilégié est la conviction, accordent une importance plus grande à la recherche de la satisfaction de leurs publics.
Il n'empêche que, globalement, c'est bien dans le cadre d'un rapprochement des méthodes des gestions publique et privée (développement des démarches "qualité", du contrôle de gestion et de la mesure des performances) qu'est réapparu l'intérêt porté à la satisfaction de l'usager.
==========
I. LES REFORMES DE L ' ETAT ONT LONGTEMPS PRESUME LES ATTENTES DES USAGERS
De façon paradoxale, la réforme de l'Etat, qui affirme placer l'usager au centre des préoccupations de l'administration, ne s'est pas appuyée en règle générale sur la recherche des attentes des usagers.
Elle a porté sur des aspects juridiques ou procéduraux sans pour autant que les usagers soient mis en situation d'exprimer leurs attentes ou leurs réactions devant des projets d'amélioration.
En d'autres termes, les attentes des usagers à l'égard des services rendus par l'administration ont été le plus souvent présumées par les décideurs publics.
Plusieurs raisons expliquent ce paradoxe : une pression insuffisante des citoyens, la complexité des relations entre l'administration et les usagers, l'existence de dispositifs de régulation et de médiation.

Figure 28 - Extrait découpé en segments thématiques

Cependant, bien que le module statistique propose un découpage du texte en segments homogènes du point de vue de leur thématique, il n'indique ni ne fournit la liste des thèmes abordés dans les documents. Pour cibler les thèmes saillants du discours (Porhiel, 2001), nous identifions automatiquement des marqueurs linguistiques de cohésion. Cette étape constitue le versant linguistique de notre méthode de détection de thèmes.

3 Le module linguistique

Une étape cruciale de l'extraction d'information est la localisation des énoncés contenant de l'information pertinente. Pour ce faire, nous avons sélectionné deux types de marqueurs linguistiques de cohésion : des marqueurs lexicaux cadratifs (une sélection de cadres de discours de (Charolles, 1997)) et des expressions référentielles constituant les chaînes de référence (Chastain, 1975 ; Corblin, 1995a, 1995b ; Schnedecker, 1997). Nous présentons chacun des marqueurs linguistiques sélectionnés et motivons les raisons de nos choix.

3.1 Identification automatique des marqueurs lexicaux cadratifs

Dans le cadre de notre étude[3], nous repérons les « marqueurs de surface » (Hernandez, 2004) essentiellement lexicaux (nous ne repérons pas, par exemple, les marques de paragraphe). Par exemple, dans

> **Sur le plan interministériel**, une première impulsion vient d'être donnée : le guide méthodologique « Services publics : s'engager sur la qualité du service », publié par la délégation interministérielle à la réforme de l'Etat en février 2001, souligne la nécessité de faire intervenir le regard des usagers aux différents stades d'une démarche « qualité ». Ainsi, l'évaluation de la satisfaction des usagers apparaît-elle d'emblée comme un enjeu interne à l'organisation, la conduite du changement s'appuyant sur les observations des usagers. (corpus *La Documentation Française*).

les adverbiaux cadratifs (Charolles, 1997) s'inscrivent dans cette perspective puisqu'ils sont marqués lexicalement dans la phrase (à la différence du sujet[4]).

Dans notre approche, les marqueurs lexicaux cadratifs englobent différentes unités linguistiques susceptibles d'introduire ce que l'on nomme les *champs thématiques*

[3] Parmi les études sur les adverbiaux cadratifs (Charolles, 1997 ; Charolles et Péry-Woodley, 2005), de nombreux travaux ont porté sur les cadres à portée vérifonctionnelle (univers de discours) ; essentiellement sur les adverbiaux temporels et spatiaux comme « En 1980 », « en Italie » (Virtanen, 1992), (Piérard et Bestgen, 2006a ; 2006b) car ces indices sont facilement identifiables de manière automatique (ce sont des locutions adverbiales figées).

[4] (Porhiel, 2005a : 58) donne cet exemple : dans « En ce qui concerne vos vacances, il vous faudra les retarder de quelques jours », le cadre thématique est marqué lexicalement, alors que le sujet ne l'est pas dans « Vos vacances commenceront quelques jours plus tard ».

(*e.g.* « concernant X »), les *espaces de discours*[5] (*e.g.* « d'une part/d'autre part ») qui fonctionnent souvent en série, et les *domaines qualitatifs* qui apportent des précisions sur le contenu des propositions (*e.g.* « en dépit de X »)[6], selon la terminologie de (Charolles, 1997). Nous avons choisi de traiter ces trois types de cadre de discours car leur fonction consiste à introduire le thème du segment ou à apporter des précisions « relatives à ce thème », ce qui n'est pas le cas des cadres véridictionnels qui ont pour fonction de limiter le cadre de vérité d'un groupe propositionnel (un secteur d'activité (« au cinéma »), un énoncé (« à vrai dire »)). Les adverbiaux cadratifs temporels et spatiaux ne sont pas pris en compte dans notre démarche puisque les dates et événements ne sont pas identifiés par notre système automatique.

Parmi les cadres dits de discours thématiques, nous traitons ceux que (Porhiel, 2005a) nomme « marqueurs de thématisation ». Il s'agit d'unités linguistiques détachées en tête de phrase, composées d'un introducteur de cadre thématique (une préposition simple ou composée, comme « pour », « pour ce qui concerne » (Porhiel, 2004)) et d'un complément souvent nominal, par exemple « à propos de Pierre », « au sujet des préparatifs ». Les marqueurs de thématisation sont à la fois des thèmes phrastiques et des thèmes textuels (Porhiel, 2005b) (voir chapitre 1). Ce sont des éléments de reprise, ce qui sous-entend que les référents évoqués sont déjà connus des locuteurs (on trouve souvent des déterminants : « à propos de *la* crise … », « concernant *la* réunion de ce matin …»).

L'apport de l'étude de (Porhiel, 2004) se situe au niveau de la description linguistique des introducteurs thématiques et des moyens qu'elle donne pour les identifier sans ambiguïté. Elle en décrit différentes propriétés :

- *syntaxiques* :
  - externes, telles que leur détachement en tête de phrase ou le fait que le complément soit introduit par un déterminant défini (*e.g.* « en ce qui concerne *les* dépenses annuelles… ») ;
  - internes : les introducteurs thématiques sont plus ou moins figés, c'est-à-dire qu'ils acceptent des insertions (adverbiales, adjectivales) avant ou après le mot base (*e.g.* « sur *l'épineux* sujet de la crise », « à propos *notamment* de la crise »).
- *référentielles* : par exemple, le référent est toujours connu ; ce qui sous-entend que le référent est exprimé *via* un syntagme nominal anaphorique,

[5] Notons que (Charolles, 1997), inclut les marqueurs d'intégration linéaire tels que « *Premièrement [...]. Deuxièmement [...]* » (Turco et Coltier, 1988) dans les espaces de discours.

[6] Les domaines qualitatifs n'initient ni des univers, ni des champs thématiques.

– *discursives et textuelles* : un introducteur thématique partitionne l'information (*e.g.* en différentes séquences) et favorise la cohérence thématique des textes, par exemple :

> Rapport entre les taux maximaux de résidus et les limites analytiques par analyse
>
> **Pour ce qui est des** substances dont l'administration aux animaux de boucherie est interdite, la limite de détection de la méthode d'analyse doit être suffisamment basse pour que les taux de résidus auxquels on pourrait s'attendre après utilisation d'une substance illégale soient détectés avec une probabilité d'au moins 95 %.
>
> **Pour ce qui est des** substances à limite maximale de résidus, la limite de détermination de la méthode majorée de trois fois l'écart type que la méthode indique pour un échantillon correspondant au taux maximal de résidus ne doit pas dépasser celui-ci. (corpus *Acquis Communautaire*)

ainsi que plusieurs fonctions pour le locuteur et l'interlocuteur (pour un état complet de ces propriétés, voir (Porhiel, 2004 : 31-34)).

Au terme de son étude, (Porhiel, 2004) fournit une double liste des introducteurs de cadre thématique qui inclut :

– des prépositions : *à l'égard de, à propos de, au chapitre (de), au niveau (de), au plan (de), au point de vue (de), au sujet de, chapitre, comme, concernant, côté, dans le cas de, du côté de, du point de vue (de), en ce qui a trait à, en ce qui concerne, en ce qui regarde, en ce qui touche (à), en fait de, en matière de, en parlant de, niveau, point de vue, pour, pour ce qui a trait à, pour ce qui concerne, pour ce qui est de, pour ce qui regarde, pour ce qui relève de, pour ce qui touche (à), quant à, question, relativement à, s'agissant de, sur, sur le chapitre (de), sur le plan (de), sur le sujet de, touchant (à) ;*

– des adverbes constitués à la base, qui semblent partiellement figés (anaphores résomptives) : *à ce propos, à ce chapitre, à ce sujet, sur ce chapitre, sur ce sujet.*

Les marqueurs lexicaux cadratifs incluent donc les introducteurs thématiques étudiés par (Porhiel, 2004 : 29) ainsi que les exemples d'espaces de discours portant sur l'organisation du discours (*d'une part, d'autre part, premièrement/deuxièmement, d'un côté/de l'autre*) et de domaines qualitatifs (*en dépit de X, à l'insu de X*) fournis par (Charolles, 1997) car ils donnent une « instruction de segmentation » (Ho-Dac, 2003). (Charolles, 1997) précise que les introducteurs de champs thématiques et de domaines qualitatifs sont dépendants

des contenus propositionnels présentés en amont et se retrouvent moins souvent à l'initiale du discours (à la différence des cadres spatiaux et temporels). En cela, ils ont une portée généralement limitée. Par exemple, le locuteur pourra utiliser l'expression « en ce qui concerne Paul » que si le référent « Paul » est déjà connu de l'interlocuteur.

L'analyse automatique de ces structures constitue en soi un problème complexe, qui a fait l'objet d'études spécifiques, telles que (Ferret *et al.*, 2001) et (Couto *et al.*, 2004) pour les cadres thématiques, (Jackiewicz, 2002) pour les cadres organisationnels et (Bilhaut *et al.*, 2003), étude dans laquelle un système dédié à l'analyse des cadres spatiaux-temporels a été mis en place. Néanmoins, une implémentation basée uniquement sur le repérage des marqueurs lexicaux cadratifs paraît trop risquée pour mener à la détection des thèmes, car ce premier type d'indices n'est pas forcément présent dans chaque texte. La combinaison de ces marqueurs à d'autres marqueurs tels que les marqueurs référentiels constitue un moyen pour mener à la détection des thèmes. Cette combinaison entre des marqueurs de cohésion « descendants » (les cadratifs) et « remontants » (les marqueurs référentiels), permettra un traitement automatique efficace des thèmes (Charolles et Péry-Woodley, 2005).

3.2 Identification automatique des chaînes de référence

Dans notre approche robuste à base de peu de connaissances, nous travaillons sur les relations s'établissant entre des expressions coréférentes dans un même paragraphe, par exemple : « Barack Obama…il…il ». De plus, nous traitons les situations de coréférence directe (Manuélian, 2003) où les groupes nominaux coréférents possèdent la même tête nominale (par exemple, « le diplomate français / ce diplomate »). Cela sous-entend donc que seules les reprises d'un référent « identifié » (une personne, un objet, un événement, un groupe de personnes) sont considérées ici. Par exemple, face à « Les supporters du Quinze de France », le système sera capable d'identifier uniquement des reprises concernant le groupe en entier (*e.g.* « ils », « les supporters », « ces supporters », « les supporters du Quinze de France »).

L'identification automatique des chaînes de référence s'effectue par l'annotation des diverses expressions référentielles contenues dans le texte (susceptibles de constituer des maillons) mais aussi grâce à l'utilisation de ressources externes telles que des listes de noms de fonctions (*e.g.* « président », « ministre »,

« sénateur »).

3.2.1 Expressions référentielles annotées

Notre système annote diverses expressions référentielles potentiellement incluses dans des chaînes de référence. Sont ainsi identifiés les pronoms[7] (« elle »), les groupes nominaux simples (« la décision », « l'arrêté »), les descriptions définies (« Président de la République »), les démonstratifs (« ce »), les possessifs (« son »), les entités nommées (« Enki Bilal », « France », « Université de Strasbourg »). Cette étape sera détaillée au chapitre 7, lors de la présentation du module d'identification automatique des chaînes de référence *RefGen*.

En suivant la remarque formulée par (Ehrmann, 2008 : 23), nous identifions différents types d'entités nommées (les noms de lieux, de personnes, d'organisations, de fonctions) car *« les entités nommées semblent constituer une source d'information non négligeable pour un système s'attachant à calculer des liens de coréférence »*. Par exemple, ((Zhou *et al.*, 2005), cités par (Ehrmann, 2008 : 23)) montrent que l'utilisation du type d'entité nommée (personne, organisation, etc.), permet une amélioration de leur système de 8,1 points pour la tâche de résolution de la coréférence. Dans la même lignée, pour combler (en partie) les lacunes rencontrées par les systèmes d'identification de la coréférence, (Dimitrov *et al.*, 2005) ont développé un système permettant de résoudre les anaphores pronominales lorsque l'antécédent est une entité nommée.

En France, lors de la dernière journée ATALA consacrée aux entités nommées[8], le « suivi d'entités nommées (suites coréférentielles) » a été considéré parmi les nouveaux challenges-phares. Toujours pour traiter le français, dans sa thèse, (Tardif, 2010) propose un algorithme pour la résolution de la coréférence entre entités nommées (dans des journaux). Nous inscrivons notre démarche dans cette lignée.

Nous annotons aussi les emplois impersonnels du pronom *il (e.g.* « il pleut », « il faut écouter »), pour éviter que ces pronoms non anaphoriques soient pris en compte par notre système de calcul de la référence. Ce faisant, nous allons dans le sens de (Danlos, 2005 : 390) : *« Un système de résolution des anaphores doit être capable de repérer les occurrences des pronoms impersonnels avant de s'attaquer*

[7] L'identification des pronoms non anaphoriques (« il pleut », « il faut partir ») font l'objet d'une annotation particulière dans le chapitre 7, section 3.3).

[8] L'intitulé de la journée ATALA était : « Reconnaissance d'Entités Nommées - Nouvelles Frontières et Nouvelles Approches », Paris, 20 juin 2011, http://tln.li.univ-tours.fr/Tln_Colloques/Tln_REN2011.html

aux pronoms anaphoriques et aux autres anaphores. ». L'utilité de cette distinction entre pronom anaphorique et pronom impersonnel est aussi signalée par (Ehrmann, 2008 : 23) : *« pour résoudre la coréférence du pronom neutre* it *en anglais, il peut être utile de discriminer parmi ses antécédents possibles ceux référant à une personne ou non, afin d'empêcher les rattachements malheureux. »*

L'étape d'annotation des diverses expressions référentielles est suivie de l'étape du calcul de la référence, permettant de rattacher les divers maillons. Cette étape calculatoire sera présentée au chapitre suivant. A la suite de ce calcul, chaque maillon de la chaîne de référence se voit automatiquement ajouter un attribut (`coref`) suivi du numéro de la chaîne de référence[9] (voir Figure 29). Nous proposons ci-dessous un exemple de résultat final, pour l'énoncé : « **Martine Aubry** est candidate à la primaire socialiste. **Elle** l'a annoncé depuis **sa** ville de Lille. **Elle** promet, si **elle** est élue à l'Elysée, d'[10]engager le redressement de la France ». Dans cet exemple de sortie XML de notre module d'identification des chaînes de référence *RefGen*, chaque maillon de la chaîne de référence (*i.e* {« Martine Aubry », « Elle », « sa », « Elle », « elle »}) possède un attribut `coref="1"` (surligné en jaune dans l'exemple).

```
<?xml version="1.0" encoding="UTF-8" ?>
  <!DOCTYPE segments (View Source for full doctype...)>
- <segments>
- <seg lang="fr">
- <s id="ttlfr.1">
  <w Ner="NER#1, Pers#1" ana="Np" chunk="Np#1" lemma="Martine" coref="1">Martine</w>
  <w Ner="NER#1, Pers#1" ana="Np" chunk="Np#1" lemma="Aubry" coref="1">Aubry</w>
  <w ana="Vaip3s" lemma="être">est</w>
  <w ana="Ncfs" chunk="CNP#1, Np#2" lemma="candidat">candidate</w>
  <w ana="Spa" chunk="CNP#1, Pp#1" lemma="à">à</w>
  <w ana="Da-fs" chunk="CNP#1, Pp#1, Np#3" lemma="le">la</w>
  <w ana="Ncms" chunk="CNP#1, Pp#1, Np#3" lemma="primaire">primaire</w>
  <w ana="Af-fs" chunk="CNP#1, Pp#1, Np#3, Ap#1" lemma="socialiste">socialiste</w>
  <c>.</c>

  </s>
  </seg>
- <seg lang="fr">
- <s id="ttlfr.2">
  <w ana="Pp3fs" lemma="il" coref="1">Elle</w>
  <w ana="Pp3" lemma="le">l'</w>
  <w ana="Vaip3s" chunk="Vp#1" lemma="avoir">a</w>
  <w ana="Vmps-s" chunk="Vp#1" lemma="annoncer">annoncé</w>
  <w ana="Sp" chunk="CNP#1, Pp#1" lemma="depuis" pattern="119">depuis</w>
  <w ana="Ds3fs" chunk="CNP#1, Pp#1, Np#1" lemma="son" coref="1">sa</w>
  <w ana="Ncfs" chunk="CNP#1, Pp#1, Np#1" lemma="ville">ville</w>
```

[9] Nous expliquerons les divers attributs présents dans le code XML des sorties de *RefGen* dans le chapitre 7.

[10] Dans notre approche, les expressions implicites et les marques d'accord ne sont pas identifiées comme maillon de chaîne de référence. D'autres approches telle que celle menée par le groupe de travail COREF (laboratoire Lattice, Paris) annote toutes les expressions (expressions référentielles et « indices » (marques d'accord, expressions implicites)). Notons que cette dernière approche est une approche manuelle qui utilise le logiciel d'annotation et d'analyse de corpus écrit *Analec* (Victorri, 2011 : http://www.lattice.cnrs.fr/Telecharger-Analec).

```
<w ana="Spd" chunk="CNP#1, Pp#2" lemma="de">de</w>
<w Ner="loc#1" ana="Np" chunk="CNP#1, Pp#2, Np#2" lemma="Lille">Lille</w>
<c>.</c>
</s>

</seg>
- <seg lang="fr">
- <s id="ttlfr.3">
<w ana="Pp3fs" chunk="Vp#1" lemma="il" coref="1">Elle</w>
<w ana="Vmip3s" chunk="Vp#1" lemma="promettre">promet</w>
<c>,</c>
<w ana="Cs" lemma="si">si</w>
<w ana="Pp3fs" lemma="il" coref="1">elle</w>
<w ana="Vaip3s" chunk="Vp#2" lemma="être">est</w>
<w ana="Vmps-s" chunk="Vp#2" lemma="élire">élue</w>
<w ana="Spa" lemma="à">à</w>
<w ana="Da-fs, Da-ms" lemma="l'">l'</w>
<w ana="Np" chunk="Np#1" lemma="Elysée">Elysée</w>
<c>,</c>
<w ana="Spd" chunk="Vp#3" lemma="de">d'</w>
<w ana="Vmn" chunk="Vp#3" lemma="engager">engager</w>
<w ana="Da-ms" chunk="CNP#1, Np#2" lemma="le">le</w>
<w ana="Ncms" chunk="CNP#1, Np#2" lemma="redressement">redressement</w>
<w ana="Spd" chunk="CNP#1, Pp#1" lemma="de">de</w>
<w ana="Da-fs" chunk="CNP#1, Pp#1, Np#3" lemma="le">la</w>
<w Ner="loc" ana="Np" chunk="CNP#1, Pp#1, Np#3" lemma="France">France</w>
<c>.</c>
</s>
</seg>
```

Figure 29 - Exemple de sortie annotée en chaîne de référence par *RefGen*

Néanmoins, certaines relations anaphoriques, bien que participant aux chaînes de référence, n'ont pu être identifiées par notre système par manque de connaissances externes et constituent des limites.

3.2.2 Limites

Parce que nous utilisons une méthode robuste utilisant peu de connaissances externes (Mitkov, 2001), nous ne sommes pas en mesure de traiter les cas d'anaphores plurielles, comme dans : « Barack et Michèle Obama … le couple présidentiel … Michèle » où les deux noms propres « Barack » et « Michèle » sont d'abord instanciés par le groupe nominal « le couple », puis extraits par le nom propre « Michèle ». En effet, le « complexe référentiel » (Schnedecker, 2006 : 316) initial composé de deux noms propres conjoints coordonnés par « et » doit être identifié comme un seul nouveau référent[11]. De plus, ce complexe est réinstancié par le groupe nominal « le couple » qui fait appel à des connaissances sur le monde qui pourraient être formulées ainsi : « un couple est formé de deux

[11] Nous suivons (Kleiber, 1986 : 77), cité par (Schnedecker et Bianco, 1995 : 87) qui pose qu'avec la séquence [Np1 *et* Np2], « ce ne sont pas deux nouveaux référents qui sont en fait introduits, mais bien un seul référent, en l'occurrence l'ensemble des référents constitués par la coordination ».

entités ». Sans connaissance, notre système va « ouvrir » une nouvelle chaîne de référence et ne cherchera pas à rattacher « le couple » avec « Barack et Michèle Obama ». De plus, la réinstanciation d'un élément de l'ensemble [Barack + Michèle] par le nom propre « Michèle » aura aussi pour conséquence l'ouverture d'une nouvelle chaîne de référence dans notre système.

De même, l'absence de connaissances externes ne nous permet pas de traiter les cas de coréférence indirecte, comme l'hyperonymie dans « Le disjoncteur évite les surintensités. *Cette protection* est destinée aux matériels et non pas aux personnes. » et l'hyponymie, p.e. « Pour être couvert en cas d'intrusion, vous devez installer un système de sécurité. *Cette alarme* devra être conforme aux normes en vigueur. » ainsi que les cas d'anaphores associatives (relation méronymique ou partie – tout, fonctionnelle, locative, actancielle, (Kleiber, 2001 ; Charolles, 1995a, 1999)), comme « Les nageurs étaient mécontents. *L'eau* était extrêmement froide. ».

4 Détection automatique de thèmes

Une fois le texte découpé en segments thématiques et les divers marqueurs linguistiques identifiés dans chaque segment (marqueurs lexicaux cadratifs et chaînes de référence), vient la phase de détection de thèmes à proprement parler. L'objectif est ici d'indiquer à l'utilisateur le ou les thèmes traités dans le document et de lui présenter la zone de texte qui les traite. En comparant la liste des thèmes et sous-thèmes de chaque document, il sera aussi possible de proposer à l'utilisateur des documents traitant des « mêmes » thèmes (et sous-thèmes) que le document consulté.

4.1 Méthodologie

En suivant (Hernandez, 2004), nous posons l'hypothèse que le locuteur organise son discours suivant une logique d'imbrication de thèmes ; les thèmes emboités étant secondaires au thème englobant (*e.g.* un thème central et des sous-thèmes) (Legallois, 2004, 2011).

Pour chaque segment thématique, plusieurs configurations sont possibles à partir des marqueurs linguistiques identifiés :

- un seul type de marqueur linguistique est présent dans le segment thématique (un marqueur lexical cadratif ou une chaîne de référence),
- les deux types de marqueurs linguistiques sont présents dans le segment thématique.

Plusieurs cas de figure peuvent alors se présenter :

a) deux ou plusieurs chaînes de référence possèdent le même premier maillon : les chaînes de référence font référence au même référent qui constitue alors le thème du document.

b) le premier maillon d'une chaîne de référence coïncide avec le thème introduit par le marqueur lexical cadratif (*e.g.* dans « Concernant Paul », l'introducteur de cadre thématique « Concernant » introduit le thème « Paul ») : le thème est renforcé par les deux marqueurs donc le thème est un thème du document.

c) seule une chaîne de référence est présente dans le segment : dans ce cas, le premier maillon de la chaîne de référence constitue un sous-thème du document.

d) seul un marqueur lexical cadratif est présent dans le segment thématique : dans ce cas, le thème introduit est un sous-thème.

e) deux ou plusieurs marqueurs lexicaux cadratifs sont présents dans le segment thématique : chaque thème introduit est un sous-thème du document.

Une fois les thèmes et sous-thèmes identifiés, les thèmes et sous-thèmes de chaque document sont comparés deux à deux pour supprimer la présence éventuelle de doublons. Une liste de « descripteurs » thématiques est alors définie pour chaque document.

4.2 Application

Nous proposons d'illustrer nos propos à partir de l'extrait suivant issu de notre corpus de rapports publics (voir Figure 30). L'extrait choisi porte sur la mesure de la satisfaction des usagers et compte 685 mots. La segmentation statistique fournie par *C99* est schématisée par des tirets « == » et l'identification automatique des marqueurs linguistiques de surface est signalée par un italique. Chaque maillon d'une chaîne de référence est en gras et porte le numéro de sa chaîne en indice[12] . Rappelons que, suivant la définition que nous avons adoptée, une chaîne de référence doit comporter au moins trois expressions référentielles.

==========

C. LES EXPLICATIONS POSSIBLES

Plusieurs raisons peuvent expliquer les constats ci-dessus.
En premier lieu, la réforme de l'Etat ne **s'**est pas en général faite en France sous la pression d'une opinion publique mécontente de son administration en général. Les sondages d'opinion révèlent ainsi un niveau élevé de satisfaction globale vis-à-vis des services publics. *Quant* [***aux* attentes globales**]**__1**, [elles]**__1** [s']**__1** expriment de façon récurrente : plus grande rapidité dans le traitement des dossiers, meilleure explication des droits et des obligations, simplification des démarches et des procédures…

Par ailleurs, la relative faiblesse des mouvements consuméristes français les porte peu à être présents dans le secteur des prestations publiques. Lorsqu'elles les abordent, [**les associations consuméristes**]**__2** [**s'**]**__2** intéressent plus aux performances (durée de

[12] La représentation des maillons des chaînes de référence présentée ici est conforme au format de sortie utilisé dans les campagnes d'évaluation telles que SemEval (http://stel.ub.edu/semeval2010-coref/).

délivrance des titres, longueur des procédures judiciaires, risques sanitaires dans les hôpitaux…) et réagissent à des dysfonctionnements, souvent signalés par [leurs]__**2** membres.

En deuxième lieu, la complexité [**des relations**]__**3** entre les usagers et les administrations ne facilite pas la prise en compte de la satisfaction des premiers d'une façon globale. Le « service public à la française » recouvre un vaste champ de prestations et de produits variés, dont les deux extrêmes sont le secteur purement régalien et la production de services en situation réelle ou potentielle de concurrence.

Dans certains services, les usagers sont en contact régulier avec le service public (abonnés de France Télécom, d'EDF-GDF ou de théâtres, garagistes pour les cartes grises…), ce qui permet de réfléchir en termes de fidélisation ; dans d'autres, [**les relations**]__**3** sont plus ponctuelles (demandeurs de cartes nationales d'identité ou de passeports par exemple). Dans d'autres enfin, [**les relations**]__**3** sont répétées dans un espace de temps restreint (demandeurs d'emploi avec l'Agence nationale pour l'emploi par exemple).

Plus fondamentalement, cette diversité se retrouve dans l'ambiguïté du statut de [**l'usager**]__**4** : est-[**il**]__**4** un assujetti, un bénéficiaire, un client, un citoyen, un contribuable ? Sans doute tout à la fois, ce qui peut induire de [**sa**]__**4** part des attentes contradictoires... L'identification même [**des usagers du service public**]__**5** soulève des difficultés : qui sont-[**ils**]__**5** dans le secteur pénitentiaire et dans celui de la police (délinquants, victimes, personnels pénitentiaires, citoyens) ? dans le secteur scolaire (élèves, parents, employeurs potentiels) ?

L'IGAS observe par exemple qu'il n'est pas aisé d'identifier les représentants [**des usagers**]__**5** dans des services comme les hôpitaux ou les cliniques. "La notion de représentation [**des usagers de l'hôpital**]__**5** renvoie à des définitions différentes de [**l'usager**]__**4** : patient représenté par une association spécialisée de malades ; citoyen malade potentiel représenté par une association relevant de la société civile ; consommateur ; militant engagé dans un travail social."

En troisième lieu, des formes de régulation et de médiation plus ou moins institutionnalisées transmettent les attentes et perceptions [**des usagers**]__**5** à l'égard des services rendus par les administrations : comités d'usagers, associations, élus locaux et nationaux, mais les agents au contact [**des usagers**]__**5** qui perçoivent [**leurs**]__**5** attentes sont également en mesure de les traduire en réforme. Par exemple, la décision de la préfecture de police d'ouvrir la possibilité de déposer plainte dans des commissariats d'arrondissement (cf. annexe 1) autres que ceux dans lesquels avait été commis le délit n'a pas été précédée d'une enquête révélant l'insatisfaction du public : les cadres et agents de la police nationale avaient en effet perçu cette source d'insatisfaction des victimes.

Au-delà de ces dispositifs de médiation qui transmettent les attentes et insatisfactions [**des usagers**]__**5**, [**les sondages d'opinion sur les services publics**]__**6** constituent des guides, certes frustes, pour la conduite de la réforme de l'Etat. [**Ils**]__**6** révèlent, de façon constante, des souhaits en matière de délais de réponse, de temps d'attente, de facilité des démarches et de lisibilité des formulaires et des procédures. [**Leur**]__**6** exploitation a permis d'orienter certaines réformes administratives, quoiqu'à un niveau très agrégé, qui ne permet pas de distinguer les souhaits en fonction des services.

C'est sans doute pour cet ensemble de raisons que, très globalement, les réformes récentes ont rejoint les quatre préoccupations majeures exprimées par les citoyens

(également [**usagers**]__**5**) : des services publics plus proches, des procédures plus simples et transparentes, un accueil plus personnalisé et, pour les plus démunis, une capacité d'écoute et de conseil.

==========

(corpus *La Documentation Française*)

Figure 30 – Extrait issu du corpus de rapports publics

A l'issue de l'identification des marqueurs lexicaux cadratifs (*en premier lieu, quant à, en second lieu, en troisième lieu*) et des chaînes de référence (six chaînes), en suivant les critères présentés plus haut, la liste des thèmes et des sous-thèmes de l'extrait est la suivante :

- thèmes :
    - « les attentes globales », car le premier maillon de la chaîne de référence coïncide avec le thème introduit par le marqueur lexical cadratif « quant à » (règle b)),
    - « la complexité des relations entre les usagers et les administrations ». Ici, le complément « les relations » constitue le premier maillon d'une chaîne de référence. Ce mode de présentation des premiers maillons d'une chaîne de référence est propre au genre textuel en présence (le rapport public). Dès lors, il nous est possible d'appliquer la règle b), même si le complément est un marqueur de niveau « inférieur ».

- sous-thèmes :
    - « la réforme de l'état » (règle d)),
    - « les associations consuméristes » (règle c)),
    - « des relations » (règle d)),
    - « l'usager » (règle c)),
    - « des usagers » (règle c)),
    - « des formes de régulation et de médiation plus ou moins institutionnalisées » (règle d)),
    - « les sondages d'opinion sur les services publics » (règle c)).

Nous pouvons remarquer que les chaînes de référence 4 et 5 portent sur la notion d'« usager ». A un niveau conceptuel, on pourrait se poser la question du rattachement de ces deux réseaux formés (l'usager au singulier et l'usager au pluriel) en un seul thème englobant aux contours flous (Berrendonner, 1994 ; Johnsen, 2010). Ce regroupement permettrait de placer la notion

d'usager au rang de thème du document (et non plus de sous-thème), puisqu'il représente 3 des 9 paragraphes de l'extrait étudié.

5 Bilan

Au terme de ce chapitre, nous avons décrit le système de détection automatique de thèmes ATDS-Fr du point de vue de son fonctionnement général. Les deux modules (statistique et linguistique) de notre méthode hybride ont été illustrés. Dans un premier temps, l'utilisation du système de segmentation statistique *C99* de (Choi *et al.,* 2001) permet de découper les documents en segments thématiques. Puis, dans un second temps, pour chaque segment thématique, l'identification automatique des marqueurs lexicaux cadratifs et des chaînes de référence utilisés en complément permet de définir des thèmes et des sous-thèmes qui constituent alors des descripteurs de documents.

Dans le chapitre suivant, nous présentons de manière détaillée *RefGen*, le module d'identification automatique des chaînes de référence que nous avons mis en place pour mener à la détection automatique des thèmes des documents.

Chapitre 7

RefGen, un module d'identification automatique des chaînes de référence

Dans ce chapitre, nous présentons *RefGen*, le module d'identification automatique des chaînes de référence d'ATDS-Fr. Nous utilisons les chaînes de référence

comme marqueurs linguistiques de cohésion participant à l'identification des thèmes des documents (Cornish, 1995 ; Schnedecker, 1997).

RefGen utilise des propriétés des chaînes de référence liées au genre textuel (identifiées lors de l'étude de corpus (voir chapitre 3)) ainsi que des informations morphosyntaxiques pour identifier les premiers maillons potentiels des chaînes de référence. Une série de contraintes sémantiques, lexicales et syntaxiques permettent de déterminer les autres maillons des chaînes de référence.

Après avoir décrit l'architecture générale de *RefGen*, nous en présentons chacun des modules : le module d'étiquetage morphosyntaxique, le module d'identification des expressions référentielles (*RefAnnot)* et le module de calcul de la référence (*CalcRef).*

1 Architecture de *RefGen*

RefGen est le module d'identification automatique des chaînes de référence d'ATDS-Fr. Ce module fonctionne de manière incrémentale[1] (voir Figure 31). A partir du texte brut segmenté par *C99* (Choi *et al.*, 2001), chaque segment est étiqueté par l'étiqueteur TTL (Ion, 2007)[2]. Puis, *RefGen* procède en deux étapes pour identifier automatiquement les chaînes de référence.

La première étape est assurée par le module d'annotation *RefAnnot*. Ce module identifie plusieurs types d'expressions référentielles qui représentent des candidats potentiels au poste de premier maillon d'une chaîne de référence : les groupes nominaux complexes (*e.g.* « le ministre des affaires étrangères »), les entités nommées (*e.g.* « Philippe de la Clergerie », « IBM », « Marseille »). Les emplois impersonnels du pronom *il* (*e.g.* « il pleut ») sont aussi annotés pour ignorer ces emplois non anaphoriques dans le calcul des chaînes de référence et pour éviter ainsi les fausses paires antécédent-anaphore.

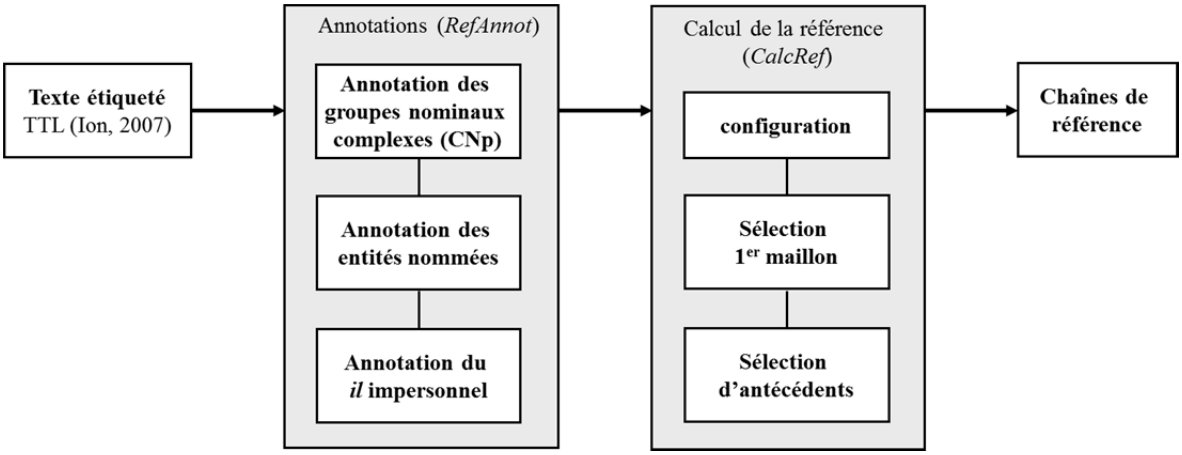


Figure 31 - Architecture du module *RefGen*

La seconde étape est assurée par le module *CalcRef*. Ce module comporte l'algorithme d'identification automatique des chaînes de référence de *RefGen*. Le calcul de la référence s'effectue en trois phases. A partir du texte enrichi en annotations, le texte est configuré suivant les paramètres des chaînes de référence

[1] Dans un système incrémental, les différents traitements sont appliqués les uns à la suite des autres.

[2] Nous présentons l'étiqueteur TTL dans la section 2. Pour une explication plus détaillée de l'outil, voir (Ion, 2007).

dépendants du genre textuel (identifiés lors de l'étude de corpus au chapitre 3). Puis, *CalcRef* associe un score à chaque expression référentielle (ce score est basé sur l'échelle d'Accessibilité de (Ariel, 1990)) pour sélectionner les candidats potentiels au poste de premier maillon d'une chaîne de référence. Ensuite, le module identifie les anaphores et les antécédents possibles. Pour sélectionner les paires antécédent-anaphore valides, *CalcRef* vérifie une série de contraintes lexicales, morphosyntaxiques et sémantiques.

L'identification automatique des diverses expressions référentielles contenues dans les chaînes de référence nécessite au préalable une phase d'étiquetage du corpus (Le Pesant, 2008 : 194). En effet, l'étiquetage morpho-syntaxique joue le rôle de filtre, permettant notamment de désambiguïser les expressions référentielles intervenant dans les chaînes de référence. Cette phase d'étiquetage est assurée par l'étiqueteur morpho-syntaxique TTL (Ion, 2007). Nous motivons le choix de cet outil et revenons sur notre participation à l'adaptation de TTL au français dans la section suivante.

2 Etiquetage avec TTL

Plusieurs étiqueteurs (ou *taggers*) sont couramment utilisés en TAL. Pour le français, on peut citer parmi eux Treetagger[3] (Schmid, 1994), Brill (Brill, 1994), Cordial Analyseur[4] (Synapse), VISL[5] (Bick, 2001). Malgré la relative fiabilité de ces outils qui obtiennent une précision de l'ordre de 97% d'annotations correctes en moyenne, le problème de l'étiquetage reste d'actualité. On trouve de nouvelles implémentations d'étiqueteurs existants (par exemple Febril[6] (Seddah *et al.*, 2010) qui est une nouvelle version de Brill) ainsi que de nouveaux outils comme Hybrid Tagger (Sigogne, 2010) ou MElt$_{fr}$[7] (Denis *et al.*, 2010). La plupart de ces étiqueteurs identifient les catégories lexicales, les lemmes et utilisent des ressources (dictionnaires, corpus d'apprentissage) (Poudat, 2004). Le jeu d'étiquettes utilisé est restreint aux catégories lexicales et à quelques sous-catégories grammaticales (pronoms personnels, démonstratifs, réfléchis, etc.). Les propriétés telles que le genre et le nombre sont souvent indisponibles.

Pour les raisons que nous venons d'évoquer, nous avons utilisé l'étiqueteur TTL (Ion, 2007) qui possède un étiquetage morphosyntaxique fin en format MULTEXT[8] (Ide et Véronis, 1994) et est disponible pour le français, l'anglais et le roumain. Nous présentons TTL et nous revenons sur notre participation à l'adaptation de cet outil pour traiter le français.

2.1 Présentation de TTL (Ion, 2007)

TTL (*Tokenizing, Tagging and Lemmatizing free running texts*) est un étiqueteur lexical probabiliste indépendant du langage. C'est une extension du catégoriseur de (Brants, 2000) utilisant des modèles de Markov cachés. Développé en *Perl* par (Ion, 2007), TTL se compose de différents modules de traitement : segmentation en phrases `(s)`, découpage en *tokens*[9] `(w)`, étiquetage `(ana)`, lemmatisation

[3] *Treetagger* est disponible à : http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

[4] *Cordial* : CORrecteur D'Imprécisions et Analyseur Lexico-sémantique

[5] *VISL* : Visual Interactive Syntax Learning

[6] Febril est disponible à : http://www.computing.dcu.ie/~dseddah/downloads/FEBRIL.tar.gz

[7] *Melt* est disponible à : https://gforge.inria.fr/projects/lingwb/

[8] La convention Multext est une notation standardisée des propriétés morphosyntaxiques de chaque catégorie lexicale. Elle est disponible à : http://aune.lpl.univ-aix.fr/projects/multext.

[9] Les *tokens* comprennent les mots, les nombres et les signes de ponctuation.

`(lemma)`[10] et analyse syntaxique partielle (identification de *chunks* simples : groupes nominaux `(Np)`, adjectivaux et adverbiaux `(Ap)`, prépositionnels `(Pp)` et verbaux `(Vp)`). TTL récupère en entrée du texte en format UTF-8 (sans BOM) et fournit une sortie en format XML. Voici un exemple de sortie annotée par TTL pour la phrase « Les cadrans solaires ont joué un rôle déterminant » (voir Figure 32) :

```
<seg lang="fr">
- <s id="ttlfr.12">
  <w lemma="le" ana="Da-mp" chunk="Np#1">Les</w>
  <w lemma="cadran" ana="Ncmp" chunk="Np#1">cadrans</w>
  <w lemma="solaire" ana="Af-mp" chunk="Np#1,Ap#1">solaires</w>
  <w lemma="avoir" ana="Vaip3p" chunk="Vp#1">ont</w>
  <w lemma="jouer" ana="Vmps-s" chunk="Vp#1">joué</w>
  <w lemma="un" ana="Da-ms" chunk="Np#2">un</w>
  <w lemma="rôle" ana="Ncms" chunk="Np#2">rôle</w>
  <w lemma="déterminant" ana="Af-ms" chunk="Np#2,Ap#2">déterminant</w>
```

Figure 32 – Exemple de sortie en XML étiquetée par TTL

Le jeu d'étiquettes utilisé est défini par le projet MULTEXT (Ide et Véronis, 1994) qui fournit une description complète et compacte des propriétés morphosyntaxiques (MSD, *Morpho-Syntactic Descriptors*)[11] comme le genre, le nombre, le temps, le mode, la personne. L'identification des noms propres (sans le typage de la catégorie) est possible à l'aide de règles heuristiques simples (annotation des mots débutant par une majuscule, sans les abréviations). Dans l'exemple de la Figure 32, l'étiquette `Ncmp` de « cadrans » signifie « nom commun masculin pluriel ».

TTL réalise un étiquetage à deux niveaux (Ceauşu, 2006) pour obtenir un meilleur traitement des mots inconnus : un premier étiquetage est appliqué en utilisant un jeu d'étiquettes simplifié (94 étiquettes issues de MULTEXT), puis les lemmes sont recherchés à l'aide d'un dictionnaire (de 600 000 mots environ, contenant la forme du mot, son lemme et son étiquette morphosyntaxique (voir Figure 33)) et des contextes.

[10] En cas d'ambiguïté, le système fournit une probabilité pour le lemme proposé, mise entre parenthèses.

[11] La liste des étiquettes morphosyntaxiques de MULTEXT pour le français utilisée par TTL est disponible en Annexe 2.

Forme	Lemme	Etiquette
soutenir	=[12]	Vmn
soutenons	soutenir	Vmip1p
soutenais	soutenir	Vmii1s
soutenais	soutenir	Vmii2s
soutenait	soutenir	Vmii3s
soutenables	soutenable	Af-fp
soutenables	soutenable	Af-mp

Figure 33 – Extrait du dictionnaire de TTL

Si le lemme est inconnu, des règles heuristiques dépendantes de la catégorie lexicale permettent de le déduire.

TTL est disponible en ligne comme service Web sur la plate-forme Weblicht[13].

2.2 Adaptation de TTL au français

Nous avons participé à l'adaptation de l'étiqueteur TTL (Ion, 2007) – développé à l'origine pour le découpage et la lemmatisation du roumain et de l'anglais – au français[14]. Pour construire le modèle de langue, TTL nécessite des données d'entraînement validées par un humain. Cette phase préparatoire du catégoriseur a consisté à fournir un corpus d'apprentissage d'un million de mots et formes en français (composé de textes issus de l'*Acquis Communautaire* (Steinberger *et al.*, 2006) et de l'*Est Républicain* (2003))[15], étiqueté et lemmatisé correctement. Pour ce faire, le corpus d'apprentissage a d'abord été étiqueté avec TreeTagger[16] (Schmid, 1994) puis nous avons appliqué Flemm[17] (Namer, 2000) pour étiqueter avec le jeu d'étiquettes fin de MULTEXT (voir un exemple en Figure 34 et Figure 35).

[12] Le signe « = » signifie que le lemme est identique à la forme ; dans notre exemple, le lemme de « soutenir » est « soutenir ».

[13] La plate-forme Weblicht est consultable à : https://weblicht.sfs.uni-tuebingen.de/. Un code d'accès est nécessaire (disponible sur simple demande).

[14] La procédure d'adaptation de TTL au français est détaillée dans (Todirascu *et al.*, 2011).

[15] Précisément, le corpus de référence compte 901 645 *tokens* (506 412 pour l'*Acquis Communautaire* et 395 233 pour l'*Est Républicain*) et le corpus d'apprentissage compte 886 563 *tokens* (498 889 pour l'*Acquis Communautaire* et 397 674 pour l'*Est Républicain*).

[16] La liste des étiquettes utilisées par Treetagger en français est disponible en annexe 3.

[17] Flemm est disponible à : http://www.univ-nancy2.fr/pers/namer/Outils.htm

Forme	Etiquette	Lemme
Les	DET:ART	le
cadrans	NOM	cadran
solaires	ADJ	solaire
ont	VER:pres	avoir
joué	VER:pper	jouer
un	DET:ART	un
rôle	NOM	rôle
déterminant	ADJ	déterminant

Figure 34 - Etape 1 : étiquetage avec TreeTagger pour « les cadrans solaires ont joué un rôle déterminant »

Forme	Etiquette	Lemme
Les	DET(ART):Da3-p---	le
cadrans	NOM:Nc-p--	cadran
solaires	ADJ:A---p--	solaire
ont	VER(aux:pres):Vaip3p--3	avoir
joué	VER(pper):Vmps-sm--	jouer
un	DET(ART):Da3ms---	un
rôle	NOM:Nc-s--	rôle
déterminant	VER(ppre):Vmpp----1	déterminer

Figure 35 - Etape 2 : étiquetage avec Flemm

En raison du grand nombre de propriétés prises en compte par Flemm, les sorties contiennent de nombreuses ambiguïtés : la forme d'un verbe peut compter jusqu'à 8 étiquettes MSD différentes. Par exemple, Flemm propose 5 étiquettes différentes pour « elle songe » (l'étiquette qui convient dans cet exemple est en gras) (voir Figure 36)[18] :

```
songe VER(pres):Vmip1s songer ||
songe VER(pres):Vmip3s songer ||
songe VER(pres):Vmmp2s songer ||
songe VER(pres):Vmsp1s songer ||
songe VER(pres):Vmsp3s songer
```

Figure 36 – Exemple d'étiquetage ambigu avec Flemm

Plusieurs erreurs systématiques d'étiquetage ou de lemmatisation de Flemm ont été observées (voir annexe 4). Le verbe « avoir », par exemple, est souvent considéré comme un auxiliaire (au lieu d'un verbe principal) lorsqu'il n'est pas suivi directement par un nom (*e.g.* « il a sans doute intérêt »).

Nous avons corrigé les erreurs d'étiquetage ou de lemmatisation récurrentes par l'application de scripts de corrections automatiques (en *perl*) et par quelques

[18] Dans Flemm, les différentes possibilités d'étiquettes pour un même mot sont séparées par des ||.

corrections manuelles pour les cas ambigus[19]. Dans l'exemple de la Figure 37, le genre est absent pour le nom commun et l'adjectif (tirets rouges). Le script `repair_det.pl` (voir annexe 5) permet de récupérer et de propager automatiquement l'information du genre du déterminant vers le nom et l'adjectif qui le suivent.

Forme	**Etiquette**	**Lemme**
Les	DET(ART):Da3mp---	le
cadrans	NOM:Nc-p--	cadran
solaires	ADJ:A---p--	solaire

Application de `repair_det.pl` :

Forme	**Etiquette**	**Lemme**
Les	DET(ART):Da3mp---	le
cadrans	NOM:Ncmp--	cadran
solaires	ADJ:A--mp--	solaire

Figure 37 – Exemple de correction automatique par script Perl (propagation du genre)

Nous avons enrichi le dictionnaire existant d'une liste de noms propres, de sigles et d'abréviations issus de notre corpus d'apprentissage. TTL a ensuite été entraîné à l'aide de ce corpus corrigé. Ainsi, nous avons pu bénéficier d'une première version de TTL en français[20]. Nous avons ensuite procédé à l'évaluation de TTL.

2.3 Tests de TTL

Dans (Todirascu *et al.*, 2011), deux tests ont été menés pour l'étiquetage et la lemmatisation de la version française de TTL :

– *Le premier test* de TTL a consisté à comparer les sorties fournies par TTL par rapport aux annotations manuelles d'un extrait du corpus de référence[21]. Le corpus de test comptait 15 091 *tokens* (7532 pour l'*Acquis Communautaire* et 7559 pour *l'Est Républicain*). TTL a obtenu une précision moyenne de 98,01% pour l'étiquetage (97,92% pour l'*Acquis Communautaire* et 98,1% pour *l'Est Républicain*) et de 98,16% pour la

[19] Ce travail a été confié en partie à Mirabela Navlea dans le cadre de son stage de master 2 recherche « linguistique et informatique » à l'Université de Strasbourg en 2008 (les règles ont été définies dans le cadre du stage), puis poursuivi par Amalia Todirascu et nous-même.

[20] La partie technique de l'apprentissage de TTL a été menée par R. Ion (Laboratoire RACAI, Romanian Academy, Roumanie).

[21] Le corpus de test est distinct du corpus utilisé pour l'apprentissage de TTL.

lemmatisation (98,36% pour l'*Acquis Communautaire* et 97,96% pour *l'Est Républicain*). Les erreurs récurrentes de TTL que nous avons pu relever sont notamment la confusion entre les participes passés et les adjectifs (*e.g.*, « il est *essoufflé* » et « *Essoufflé*, il a du mal à respirer »), les erreurs de lemme pour les formes ambiguës, l'absence de désambiguïsation de l'étiquette pour les déterminants « l' » et « les ». Pour cette dernière erreur, TTL propose par exemple la double étiquette « `Da-fs,Da-ms` » (déterminant féminin singulier ou déterminant masculin singulier) pour « l' » dans « L'une des manières… » (voir Figure 38) :

```
<seg lang="fr">
- <s id="ttlfr.273">
  <w lemma="L'" ana="Da-fs,Da-ms">L'</w>
  <w lemma="un" ana="Da-fs" chunk="Np#1">une</w>
  <w lemma="de_le" ana="Dg-fp" chunk="Np#1,Pp#1">des</w>
  <w lemma="manière" ana="Ncfp" chunk="Np#1,Pp#1">manières</w>
```

Figure 38 – Exemple d'ambiguïté du genre conservée dans l'étiquette de TTL pour « L'une des manières… »

– dans *le second test* de TTL, les sorties de TTL ont été comparées avec celles de TreeTagger (Schmid, 1994)[22]. Pour ce faire, un modèle d'étiquetage pour TreeTagger a été créé à partir d'un extrait du corpus d'entraînement de 370 000 mots environ. Puis, nous avons constitué un corpus d'évaluation composé d'une partie du corpus d'entraînement (textes issus de l'*Acquis Communautaire* et de l'*Est Républicain*), d'un extrait du roman *les Trois Mousquetaires* de Dumas et de revues d'informatique (80 000 mots) pour comparer les sorties des deux étiqueteurs. Les performances obtenues par TTL sur ce corpus de genres textuels variés sont de l'ordre de 97,14% en moyenne pour l'étiquetage et 98,05% pour la lemmatisation (voir Tableau 24), soit une différence de 2 à 3 % par rapport aux performances de TreeTagger. Concernant la lemmatisation, TTL obtient des précisions supérieures à 98% pour l'*Acquis Communautaire*, *Les Trois Mousquetaires* et *l'Est Républicain*. Les performances de TTL sont moins élevées pour l'extrait de corpus de *revues d'informatique* (97% pour l'étiquetage et 97,45% pour la lemmatisation), mais elles sont encore plus basses pour TreeTagger (94,85% pour l'étiquetage et 93% pour la lemmatisation). Cette baisse des performances rencontrée pour les deux étiqueteurs s'explique par la présence de nombreux termes spécialisés dans ce sous-corpus (*i.e.* des mots inconnus

[22] La validité de ce test est discutable, dans la mesure où l'outil TreeTagger a été utilisé comme base d'étiquetage du corpus d'apprentissage de TTL. Néanmoins, lorsque ce test a été mené (en 2011), aucun autre étiqueteur pour le français n'était disponible (par exemple, l'étiqueteur MeLT_{fr} (Denis et Sagot, 2010) n'était pas distribué avec la possibilité d'apprendre un nouveau modèle).

des étiqueteurs) ainsi que par des erreurs de segmentation (liées notamment à une absence de distinction typographique entre le titre et le début d'un paragraphe).

Corpus	**Etiquetage TTL**	**Lemmatisation TTL**	**Etiquetage TreeTagger**	**Lemmatisation TreeTagger**
Acquis Communautaire	97,31%	98,74%	95,6%	96,86%
Les trois Mousquetaires	97,01%	98,01%	95,81%	96,6%
L'Est Républicain	97,22%	98%	96,12%	97%
Revues d'informatique	97%	97,45%	94,85%	93%
Moyenne	**97,14%**	**98,05%**	**95,60%**	**95,87%**

Tableau 24 – Comparaison des précisions de TTL et de TreeTagger pour l'étiquetage et la lemmatisation du corpus de genres variés

Les principales erreurs communes rencontrées par les deux étiqueteurs concernent l'identification d'agrégats, de déterminants ou d'adjectifs. Les erreurs fréquentes de TreeTagger touchent surtout la reconnaissance des noms propres, essentiellement lorsque ces noms constituent des entrées lexicales. De son côté, TTL étiquette des auxiliaires comme des verbes principaux, alors que TreeTagger évite ces erreurs. Et, lorsque la forme du verbe est ambiguë, TTL propose souvent le subjonctif (bien que la majorité de ces emplois concerne l'indicatif), ce qui n'est pas le cas de TreeTagger.

A l'issue de ces tests, nous avons mis en place des patrons de correction des erreurs les plus fréquentes de TTL, afin d'éviter, tant que faire se peut, de propager ces erreurs dans la suite de la chaîne de traitements de *RefGen*. Une évaluation de TTL est proposée au chapitre 8, section 2.2.

2.4 Patrons de correction

Nous avons défini une série de patrons de correction des erreurs les plus fréquentes de TTL. Les patrons de correction ont été intégrés à l'entrée du module d'annotations *RefAnnot* (voir section 3) et ils suivent donc son formalisme. Par exemple, pour corriger l'erreur fréquente de TTL relative à la catégorie du verbe (principal ou auxiliaire), comme dans « il n'y **a** pas de modèle », le patron de correction est le suivant (voir Figure 39) :

```
- <pattern id="5">
   <one value="il" />
   <one lemma="ne" />
   <one lemma="y" />
   <one lemma="avoir" ana="Vaip3s" />
   <all chunk="Ap" />
   <all chunk="Pp" />
  </pattern>

- <actions>
  - <action kind="set" tag="ana" value="Vmip3s">
      <filter lemma="avoir" ana="Vaip3s" />
    </action>
  </actions>
```

Figure 39 – Exemple de patron de correction d'erreur de TTL sur la catégorie du verbe (principal ou auxiliaire)

Le patron 5 permet de modifier l'étiquette de verbe auxiliaire (`"Vaip3s"`) attribuée à « avoir » en verbe principal (`"Vmip3s"`). Le patron se divise en deux parties : la première partie décrit le contexte (`<pattern> … </pattern>`) tandis que la seconde décrit l'action à effectuer (*e.g.*, remplacer (`set`) la valeur de l'étiquette `"ana"` de « avoir »).

Le patron de correction 14 (voir Figure 40) permet, lui, de modifier la catégorie de « pas », dans une phrase telle que « il n'y a qu'**un pas** » :

```
- <pattern id="14">
   <one lemma="un" />
   <one lemma="pas" ana="R" />
  </pattern>

- <actions>
  - <action kind="set" tag="ana" value="Ncms">
      <filter lemma="pas" />
    </action>
  </actions>
```

Figure 40 – Exemple de patron de correction d'erreur de TTL sur la catégorie de « pas »

Dans ce dernier patron, l'étiquette erronée d'adverbe (`"R"`) attribuée à « pas » est remplacée par l'étiquette `"Ncms"` (nom commun masculin singulier).

A l'issue de l'étiquetage avec TTL dans sa version française et de l'application de patrons de correction, le texte étiqueté et lemmatisé est traité par le module d'annotations *RefAnnot*.

3 Annotations (*RefAnnot*)

A partir de la sortie fournie par TTL, le texte passe dans le module d'annotations *RefAnnot*. L'objectif de ce module est d'annoter certains types d'expressions référentielles susceptibles d'être présentes en début des chaînes de référence (car plus informatives) : les groupes nominaux complexes (Royauté, 1999) et les entités nommées. Nous exploitons ces annotations pour rechercher les relations de coréférence.

RefAnnot annote également les emplois impersonnels du pronom *il* pour que le module de calcul de la référence ne les prenne pas en compte et évite ainsi les fausses paires antécédent-anaphore.

Les patrons d'identification de *RefAnnot* sont traités de manière séquentielle. Les règles s'appliquent en respectant les deux heuristiques suivantes :

- principe du plus long patron (*longest match*) : les règles les plus longues sont appliquées en premier,
- pour un même type d'annotation (groupes nominaux complexe, entités nommées, il impersonnel) une règle ne peut pas s'appliquer à l'intérieur d'une séquence déjà reconnue par une autre règle.

Les sorties de ce module sont disponibles en format XML et en format HTML[23] pour une meilleure lisibilité des résultats[24].

3.1 Annotation des groupes nominaux complexes

A l'issue du découpage en *chunks* simples fournis par TTL (groupes nominaux `(Np)`, adjectivaux et adverbiaux `(Ap)`, prépositionnels `(Pp)` et verbaux `(Vp)`), nous avons mis au point une base de 122 patrons symboliques pour identifier les groupes nominaux complexes (CNp). Les patrons sont ordonnés par nombre de *chunks*, pour reconnaître d'abord les séquences les plus longues et éviter les inclusions.

[23] Le format HTML sera préféré pour illustrer les exemples qui suivront.

[24] L'architecture Java du module *RefAnnot* a été développée par Damien Obringer, ingénieur développement à RBS.

Un groupe nominal complexe est un groupe nominal modifié par deux groupes prépositionnels au plus ou bien un groupe nominal modifié par une proposition relative (une proposition simple contenant un prédicat et un complément d'objet direct ou indirect), comme « l'utilisation des fonds publics », « le rapport qui présente les mesures prises contre le changement climatique ». Les groupes nominaux complexes sont plus informatifs que les groupes nominaux simples : « le ministre des affaires étrangères » est plus informatif que « le ministre ». De plus, les groupes nominaux complexes introduisent souvent une nouvelle entité du discours. Par exemple, le patron 72 (voir Figure 41) permet de regrouper une séquence formée d'un groupe nominal simple suivi de deux groupes prépositionnels. L'action de ce patron consiste alors à ajouter une étiquette « CNp » à chaque élément de cette séquence :

```
- <pattern id="72">
    <all chunk="Np" />
    <all chunk="Pp" />
    <all chunk="Pp" />
  </pattern>

- <actions>
  - <action kind="addleft" tag="chunk" value="CNp" grouped="true">
    </action>
  </actions>
```

Figure 41 – Exemple de patron d'identification des groupes nominaux complexes

Ainsi, avec le patron 72, le groupe nominal simple (`Np#8`) « une élévation » et les groupes prépositionnels « du niveau moyen global » (`Pp#4`) et « de la mer » (`Pp#5`) sont regroupés en un seul groupe nominal complexe « une élévation du niveau moyen global de la mer » (`CNP#3`) (voir Figure 42) :

ANA	CHUNK	LEMMA	PATTERN	VALUE
Da-fs	**CNP#3** | Np#8	un	72	une
Ncfs	**CNP#3** | Np#8	élévation	72	élévation
Dg-ms	**CNP#3** | Pp#4 | Np#9	de+le	72	du
Ncms	**CNP#3** | Pp#4 | Np#9	niveau	72	niveau
Ncms	**CNP#3** | Pp#4 | Np#9	moyen	72	moyen
Af-ms	**CNP#3** | Pp#4 | Np#9 | Ap#4	global	72	global
Spd	**CNP#3** | Pp#5	de	72	de
Da-fs	**CNP#3** | Pp#5 | Np#10	le	72	la
Ncfs	**CNP#3** | Pp#5 | Np#10	mer	72	mer

Figure 42 – Exemple d'annotation de groupe nominal complexe

A la suite de cette première série d'annotations, *RefAnnot* identifie un autre type d'expressions référentielles : les entités nommées (noms de personnes, d'organisations, de lieux, de fonctions).

3.2 Annotation des entités nommées

L'extracteur d'entités nommées de *RefAnnot* identifie diverses catégories d'entités nommées[25] grâce à un système de règles symboliques, des listes de mots et des indices de surface.

Dans notre approche, nous adoptons la définition des entités nommées formulées par (Ehrmann et Jacquet, 2006 : 64) :

> « les entités nommées correspondent traditionnellement à l'ensemble des noms propres présents dans un texte, qu'il s'agisse de noms de personnes, de lieux ou d'organisations, ensemble auquel sont souvent ajoutées d'autres expressions comme les dates, les unités monétaires, les pourcentages et autres ».

Comme la plupart des systèmes de reconnaissance des entités nommées, notre extracteur d'entités nommées identifie la triade noms de personnes, noms d'organisations (entreprises, sociétés) et noms de lieux (villes, pays, continents).

Dans sa thèse, (Ehrmann, 2008, chapitre 5), suite à une discussion sur le statut d'entités nommées pour les expressions telles que les descriptions définies, s'interroge sur la notion d'entité nommée relativement aux chaînes de référence :

> « Si un système informatique doté d'un modèle particulier annote le nom propre *Jacques Chirac* puis la description définie *le Président de la République* et relie ensuite les deux expressions, l'une étant le titre ou la fonction de l'autre, alors on est en droit de dire que ces deux expressions coréfèrent dans le modèle. » (Ehrmann, 2008 : 173)

Dans notre modèle, nous souhaitions identifier les liens établis entre des noms de personnes et des noms de fonctions. Ainsi, aux trois catégories générales d'entités nommées sont aussi extraits les noms de fonctions (p.e. « ministre des affaires étrangères »). Les noms de fonctions sont utilisés comme indices pour déterminer le type d'une entité nommée ou pour trouver différentes reprises d'un même

[25] Bien que les systèmes de reconnaissance des entités nommées développés soient à présent nombreux et pour la plupart disponibles librement, nous souhaitions conserver les informations morphosyntaxiques fines obtenues avec TTL et les annotations en groupes nominaux complexes. Pour cette raison, nous avons préféré développer notre propre extracteur d'entités nommées.

référent (un nom de personne et sa fonction), comme par exemple les expressions en italique dans :

> *Barack Obama* a promu lundi l'entrepreneuriat, dans le but de doper la création d'emplois. *Le président des Etats-Unis* a également dit qu'il demanderait au Congrès de rendre permanentes les exemptions d'impôt sur les revenus du capital pour les PME, et de voter d'autres allégements fiscaux pour ce secteur. (20 minutes, 31/01/2011)

Les noms de personnes et d'organisations sont souvent les acteurs principaux des textes et sont reliés au thème principal d'un paragraphe. De ce fait, les noms de personnes et d'organisations réfèrent à une entité particulière et l'identifient. Ils constituent de bons candidats au poste de premier maillon d'une chaîne de référence.

Nous avons ainsi constitué de manière semi-automatique[26] des listes de noms de fonctions à partir de nos corpus de genres variés (analyses politiques, éditoriaux, lois, rapports publics, roman)[27] :

- *des grades* : officier, lieutenant, colonel, général, capitaine, commandant, maréchal, … (Cruse, 1986 : 186-187)
- *des fonctions politiques* : président, ministre, procureur, sénateur, député, maire, …
- *des titres* : ambassadeur, archevêque, cardinal, comte, roi, prince, duc, empereur, évêque, pape, imam, …
- *des métiers*[28] : directeur, enseignant, professeur, routier, docteur, écrivain, éditeur, comédien, cinéaste, …

L'extracteur d'entités nommées, qui compte 156 règles, procède en deux phases :

- il délimite d'abord les bornes de l'entité (pour les cas où l'entité nommée contient plus qu'un mot ou lorsque des noms propres possèdent des particules issues du lexique courant, comme « M. Chirac », « lycée Couffignal », « Benoît XVI ») en ajoutant un attribut « ner=" NER " »,
- puis il attribue un type « pers », « org », « loc », « fonc », pour catégoriser le nom de personne, d'organisation, de lieu ou de fonction.

[26] Les listes de noms de fonctions ont été en partie repérées *via* des scripts Perl utilisant l'annotation des entités nommées de type 'personne'.

[27] Nous avons complété nos listes à l'aide de noms de fonctions issus de sites Web (http://www.juritravail.com, http://fr.wikipedia.org).

[28] Appelés aussi des professions, ou encore noms de statut (Swart *et al.*, 2007) ou noms de rôle (Fauconnier, 1984 ; Riegel, 1985).

Afin d'identifier les bornes des entités nommées et de les typer, le système utilise des preuves internes et externes (McDonald, 1996) :

- une *preuve externe* est le contexte d'apparition de l'entité nommée (les noms situés avant et après l'entité nommée). Par exemple, le « mot-clé » « entreprise » dans « l'entreprise RBS compte 150 employés » constitue un indice permettant de catégoriser l'entité « RBS » en tant que nom d'organisation.
- une *preuve interne* fait partie intégrante de l'entité nommée : le prénom « Jacques » dans « Jacques Chirac » indique que l'entité nommée est de type « personne », « Inc. » dans « Microsoft Inc. » permet de typer l'entité en tant qu'organisation.

L'extracteur d'entités nommées possède un formalisme propre en XML contenant deux parties : la partie *condition*, où sont listés les patrons regroupés par type (organisation, personne, lieu, fonction) et la partie *action*, où sont précisés les attributs à ajouter (*e.g.* étiquette, valeur à affecter) aux entités nommées identifiées. Par exemple, le patron ci-dessous (patron 721, voir Figure 43) permet d'annoter l'entité nommée « Zundapp » comme organisation (`"Org"`) dans le contexte « la firme allemande Zundapp ». Cette règle utilise la preuve externe « firme » (présente dans la liste de noms communs `@list@OrgNcExt`) qui désigne une organisation, pour typer « Zundapp ». Pour annoter seulement l'entité nommée « Zundapp » en tant qu'organisation (et pas toute la séquence correspondant au patron 721), un filtre est utilisé sur l'étiquette du nom propre (`filter ana="Np"`). L'action consiste alors à ajouter à droite de l'étiquette « `NER` », le trait « `Org` » (`action kind="addright" tag="NER" value="Org"`) (voir Figure 44).

```
- <pattern id="721">
   <one lemma="@list@OrgNcExt" />
   <one ana="Af.*" />
   <one ana="Np" />
  </pattern>

  - <action kind="addright" tag="NER" value="Org">
      <filter ana="Np">
      </filter>
    </action>
```

Figure 43 – Exemple de règle identifiant une entité nommée de type organisation

NER	ANA	CHUNK	LEMMA	PATTERN	VALUE
	Da-fs	Np#1	le		La
	Ncfs	Np#1	firme	726	firme
	Af-fs	Np#1 | Ap#1	allemand	726	allemande
NER#1 | Org#1	Np	Np#2	Zundapp	619 | 726	Zundapp

Figure 44 – Exemple d'annotation d'entité nommée de type organisation

Des patrons identifiant des noms de fonctions complexes (plus informatifs) tels que « le ministre des affaires étrangères », « le président directeur général » ont aussi été mis en place. Nous avons appliqué des règles heuristiques utilisant les annotations en groupes nominaux complexes disponibles, comme :

« si le groupe nominal complexe contient un nom de fonction, alors cet élément est un nom de fonction complexe ».

Nous nous sommes ensuite focalisée sur les relations établies entre entités nommées, telles que les relations entre une personne et sa fonction : « Marcel Klaus, directeur financier de Swiss ». Les cas de coordination entre deux noms de fonctions (pour une même personne) comme « Le membre de la commission et président de l'association Marc Dupont » sont aussi identifiés par notre extracteur. Dans ces derniers cas, nous utilisons la non-répétition de l'article défini comme preuve pour rattacher les deux noms de fonctions au même nom de personne. Cette relation est intéressante pour identifier des expressions référentielles informatives.

Un dernier type d'entités nommées, étiqueté « `other` » (autre) par le système, a été identifié pour éviter l'attribution de fausses étiquettes à ces entités. Cela s'applique par exemple à « l'affaire Dreyfus », « la loi Falloux ».

Néanmoins, l'extracteur n'est pas en mesure d'identifier les cas d'ellipse partielle (« Michèle et Barack Obama », « le couple Hollande – Trierweiler ») et les surnoms (« Gouvernator »). Ces derniers cas pourraient être identifiés à l'aide d'un ensemble de règles d'équivalence qui prendrait la forme suivante :

- « Michèle et Barack Obama » est équivalent à « Michèle Obama et Barack Obama » (répétition du nom propre) ;
- « Gouvernator » est équivalent à « Arnold Schwarzenegger »[29].

[29] L'activité d'association entre surnom et nom propre de personne peut se révéler complexe. Par exemple, l'ancien président du conseil italien Giulio Andreotti a reçu de nombreux surnoms : « le Divin Giulio », « la première lettre de l'Alphabet », « le bossu », « le renard », « le Moloch »,

En plus de l'identification des groupes nominaux complexes, des entités nommées et des relations entre entités nommées, *RefAnnot* identifie les emplois impersonnels du pronom *il*.

3.3 Annotation du pronom *il* impersonnel

De nombreux travaux sur la résolution des anaphores pronominales en anglais tels que ceux de (Lappin et Leass, 1994), (Kennedy et Boguraev, 1996), (Evans, 2001), (Boyd *et al.*, 2005), (Haghighi et Klein, 2007), (Charniak et Elsner, 2009), (Sobha *et al.*, 2011) se sont d'abord attachés à distinguer les emplois impersonnels du pronom 'it' (*e.g.* « it rains » / « il pleut ») des emplois anaphoriques (*e.g.* « he talks » / « il parle ») avant de s'attaquer à la résolution des anaphores à proprement parler. Pour le français, c'est notamment le système ILIMP de (Danlos, 2005) qui s'intéresse spécifiquement à l'annotation de ces emplois.

Dans cette lignée, *RefAnnot* identifie les occurrences du pronom *il* lorsque ce dernier est un pronom impersonnel[30]. L'intérêt de cette étape de traitement est d'éviter l'appariement des *il* impersonnels avec des antécédents potentiels lors du calcul de la référence. Cependant, à la différence du système ILIMP (Danlos, 2005), qui identifie à la fois les emplois anaphoriques et les emplois impersonnels du pronom *il* en français (ajout d'une étiquette « `ana` » ou « `imp` » après le pronom[31]), notre système cherche uniquement à identifier les emplois non anaphoriques du pronom *il*. En effet, nous partons du principe qu'un candidat non annoté « impersonnel » est potentiellement anaphorique (cela nous permet d'éviter d'éliminer d'emblée les occurrences du « il » qui ne possèderaient pas d'étiquette « personnelle » ou « impersonnelle » à l'issue des phases d'annotations).

Nous[32] avons créé une série de patrons morpho-syntaxiques (387 patrons) pour identifier les emplois impersonnels du pronom *il*. Ces règles symboliques ont été constituées à partir d'un corpus de revues électroniques traitant des nouvelles

« la salamandre », « le Pape noir », « l'éternité », « l'homme des ténèbres », « Belzébuth », « le sphinx », « l'indéchiffrable », « oncle Giulio » (http://it.wikipedia.org/wiki/Giulio_Andreotti#Soprannomi, consulté le 10/02/2012).

[30] Ce pronom impersonnel est aussi appelé « explétif », « semi-explétif » ou « pléonastique » selon les auteurs.

[31] Une troisième étiquette « `amb` » est aussi apposée pour les cas ambigus.

[32] Ce travail a été effectué en collaboration avec Yannick Lutz dans le cadre de son stage de Master 2 recherche « Linguistique, Informatique et Traduction » au sein de l'entreprise RBS, Strasbourg.

technologies de 500 000 *tokens* (période 2008-2009)[33]. Nous nous sommes aussi appuyés sur des listes de verbes météorologiques (« pleuvoir », « grêler », « neiger »), de participes passés à dominante impersonnelle (« il a fallu », « il a suffi ») et d'adjectifs (« il est indéniable », « il est dommage »), constituées à partir de nos divers corpus. Une attention particulière a été apportée aux cas de sujet inversé comme « est-il nécessaire de prendre de telles décisions ». Ainsi, le trait impersonnel « `imp` » est ajouté au *il* impersonnel. Par exemple, dans « il s'agit d'abord d'évaluer l'ampleur et la pertinence de la variation des indices », le patron (patron 337) d'identification du *il* impersonnel est le suivant (voir Figure 45) :

```
- <pattern id="337">
   <one value="il" />
   <one value="s." lemma="se" />
   <one lemma="agir" ana="V.*" />
  </pattern>

  - <action kind="set" tag="feat" value="imp">
      <filter value="il">
      </filter>
    </action>
```

Figure 45 - Exemple de patron d'identification du *il* impersonnel

De la même manière que pour l'exemple du patron d'identification des entités nommées, un filtre est utilisé dans ce patron pour attribuer le trait « `imp` » uniquement au pronom *il* (et non pas à toute la séquence du patron « il s'agit ») (voir Figure 46).

ANA	CHUNK	FEAT	LEMMA	PATTERN	VALUE
Pp3ms		**imp**	il	**337**	Il
Px	Vp#1		se	**337**	s'
Vmis3s	Vp#1		agir	**337**	agit
Spd	Pp#1		de		d'
Ncms	Pp#1 | Np#1		abord		abord

Figure 46 - Exemple d'annotation du *il* impersonnel

A l'issue de ce traitement, les pronoms *il* annotés « `imp` » seront ignorés par notre module de calcul de la référence *CalcRef* (ces *il* impersonnels ne feront donc pas partie des candidats anaphoriques potentiels).

[33] Les sites internet consultés ont été *LeMonde.fr*, *Le nouvel Observateur.com*, *Industrie.com* et *l'Express.fr*.

3.4 Bilan

Comme nous venons de le voir, après avoir été étiqueté et lemmatisé avec TTL, le texte est enrichi en annotations permettant d'identifier des expressions référentielles informatives (les groupes nominaux complexes et les entités nommées) et les emplois impersonnels du pronom *il*.

Les règles d'annotation mises en place dans *RefAnnot* sont en format XML, ce qui les rend facilement convertibles à d'autres formats. Indépendantes de l'architecture Java de *RefAnnot*, ces règles peuvent s'appliquer également à d'autres algorithmes de *pattern-matching*. De plus, grâce à l'architecture modulaire de *RefAnnot*, d'autres types d'entités nommées peuvent être ajoutés (*e.g.* noms de gènes, noms d'événements, etc…).

A partir du texte étiqueté et enrichi en annotations, le calcul de la référence permet de regrouper plusieurs paires antécédents-anaphores dans des chaînes de référence.

4 Calcul de la référence (*CalcRef*)

L'algorithme mis en place pour le calcul de la référence[34] *CalcRef* s'appuie sur des indices de surface (étiquettes morpho-syntaxiques, filtres lexicaux ou de proximité), telle que l'approche adoptée par (Mitkov, 1998, 2000) pour la résolution des anaphores pronominales ou celle de (Bontcheva *et al.*, 2002) pour la résolution de la coréférence. Le calcul des chaînes de référence s'effectue en plusieurs étapes (voir Figure 47).

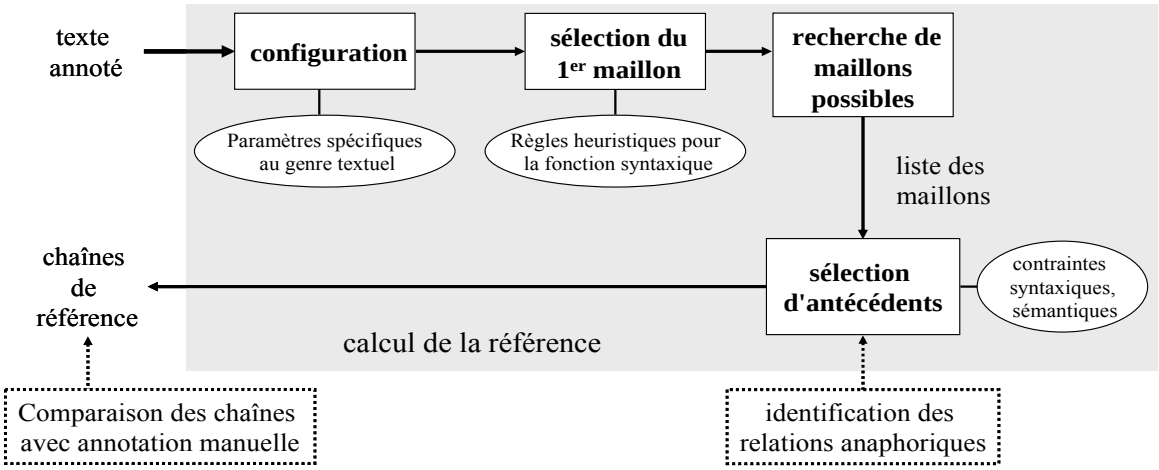


Figure 47 - Algorithme de *CalcRef*

Le module *CalcRef*[35] récupère en entrée le texte enrichi en annotations (par le module *RefAnnot*). Puis *CalcRef* est configuré suivant les paramètres spécifiques au genre textuel du document : longueur moyenne des chaînes de référence, catégorie grammaticale préférée du premier maillon, etc., (voir chapitre 3).

Pour sélectionner les candidats potentiels au poste de premier maillon d'une chaîne de référence, un score est attribué à chaque expression référentielle. Ce score est calculé à partir de :

- l'accessibilité globale du candidat basée sur l'échelle d'accessibilité de (Ariel, 1990),

[34] Ou calcul de saillance.

[35] Le module *CalcRef* a été développé en Java en collaboration avec Amalia Todirascu.

- des paramètres dépendants du genre textuel (distance moyenne entre les maillons d'une chaîne de référence, nombre moyen de maillons d'une chaîne de référence) ;
- des règles heuristiques pour la fonction syntaxique (sujet, objet direct, objet indirect, complément du nom).

A partir de la liste des premiers maillons, *CalcRef* recherche les antécédents possibles. La sélection des antécédents s'effectue grâce à une série de contraintes lexicales, morphosyntaxiques et sémantiques à valider. Les paires antécédents-anaphores valides sont regroupées en chaînes de référence lorsqu'elles partagent le même référent.

Dans les sections suivantes, nous détaillons chacune des étapes du calcul de la référence : le calcul des premiers maillons des chaînes de référence, la recherche de paires antécédents-anaphores valides et leur regroupement, menant à l'identification automatique des chaînes de référence.

4.1 Calcul des premiers maillons des chaînes de référence

La première étape du calcul de la référence s'attache à identifier les premiers maillons potentiels des chaînes de référence. La sélection des candidats susceptibles d'occuper le poste de premier maillon d'une chaîne de référence dépend de trois éléments : une classification des expressions référentielles basée sur l'échelle d'accessibilité de (Ariel, 1990), des règles heuristiques qui prennent en compte des contraintes morpho-syntaxiques et les paramètres des chaînes de référence dépendants du genre textuel.

La classification des expressions est utile pour trouver une première mention possible d'une chaîne de référence et rechercher des paires antécédent-anaphore.

4.1.1 Calcul de l'accessibilité globale

CalcRef sélectionne les candidats initiateurs d'une chaîne de référence dont l' « accessibilité globale » est la plus élevée. L'accessibilité globale est calculée à partir de l'échelle d'accessibilité de (Ariel, 1990) qui classe les expressions référentielles suivant leur degré d'informativité (les descriptions définies longues ou les noms propres complets sont plus informatifs qu'un groupe nominal simple), de rigidité (indication précise du référent) et d'atténuation (la réalisation

phonétique de l'élément) : moins le référent est accessible, plus l'expression référentielle est longue, univoque et accentuée. Ainsi, les groupes nominaux complexes et les entités nommées qui occupent une position de thèmes phrastiques sont utilisés pour introduire une nouvelle entité tandis que les expressions courtes comme les pronoms sont plutôt utilisées pour référer à des entités saillantes car leur expression référentielle figure déjà dans le discours.

Pour chacun des degrés (informativité, rigidité, atténuation), un score de 10 à 110 est attribué à chaque expression référentielle (voir Tableau 25). L'accessibilité globale d'une expression référentielle représente alors la somme de son informativité, sa rigidité et son atténuation. Dans l'exemple suivant[36],

> Moins virulent, [Patrick Devedjian]ı a [lui]ȷ aussi manifesté [son]ĸ soutien à [Michèle Alliot-Marie]ı.

on compte quatre expressions référentielles : « Patrick Devedjian », « Michèle Alliot-Marie », « lui » et « son ». L'accessibilité globale de chacun des noms propres complets (« Patrick Devedjian » et « Michèle Alliot-Marie ») est de 210 (90+90+30), celle des pronoms (« lui » et « son ») est de 150 (30+30+90).

A la hiérarchie d'accessibilité initiale de (Ariel, 1990), nous avons ajouté les groupes nominaux indéfinis, même si la reconnaissance automatique de cette catégorie de candidats génère quelques candidats erronés. En effet, dans les textes descriptifs ou informatifs, les groupes nominaux indéfinis sont souvent les premiers maillons des chaînes de référence, comme par exemple :

(45) Lorsqu'[**un** État membre] estime qu'une compagnie, bien qu'elle soit titulaire d'une attestation de conformité, ne peut exploiter un service de transbordeur roulier sur une ligne régulière à destination ou au départ de ses ports au motif qu'il existe un risque de danger grave pour la sécurité des personnes ou des biens, ou pour l'environnement, l'exploitation du service peut être suspendue jusqu'au moment où le risque a été supprimé. Dans un tel cas, la procédure suivante s'applique : a) [l'État membre] informe immédiatement la Commission et les autres États membres de [sa] décision, en la motivant dûment ; […] (corpus *Acquis Communautaire*)

(46) c'était [**un** garçon tout confit de mystères], répondant peu aux questions qu'on [lui] faisait sur les autres, et éludant celles que l'on faisait sur [lui-même]. (corpus *Les trois Mousquetaires*)

[36] Nous reprendrons cet exemple pour illustrer chaque étape du calcul de la référence.

Expression référentielle	Informativité	Rigidité	Atténuation	Accessibilité Globale
Groupe nominal indéfini	110	110	10	**230**
Np complet avec modifieur	100	100	20	**220**
Np complet	90	90	30	**210**
Description définie longue	80	80	40	**200**
Description définie courte	70	70	50	**190**
Nom de famille	60	60	60	**180**
Prénom	50	50	70	**170**
Démonstratif	40	40	80	**160**
Pronom	30	30	90	**150**
Réfléchi	20	20	100	**140**
Possessif	10	10	110	**130**

Tableau 25 - Accessibilité globale pour chaque expression référentielle

Après le calcul de l'accessibilité globale des expressions référentielles on procède au calcul du rôle syntaxique de chaque expression candidate.

4.1.2 Calcul du rôle syntaxique

Outre l'accessibilité globale, des règles heuristiques identifient la fonction syntaxique du candidat. Ces règles déterminent la position du candidat par rapport au verbe principal (sujet, objet direct, objet indirect, autres compléments). Par exemple, lorsque le candidat est situé avant le verbe principal, il est considéré comme sujet, lorsqu'il est situé après le verbe principal, il est considéré comme objet[37]. Un poids[38] est alors affecté à la fonction (voir Tableau 26) et ajouté au score d'accessibilité globale du candidat.

Position	Score
Sujet	100
Objet direct	50
Objet indirect	30
Autres compléments	20

Tableau 26 – Score pour chaque position syntaxique

En reprenant l'exemple précédent, le sujet « Patrick Devedjian » se voit rajouter 100 à son score d'accessibilité globale initial (son score est donc de 310), l'objet

[37] Notons que les cas d'inversion des interrogatives par exemple sont exclus dans notre approche, mais ces cas particuliers peuvent être pris en compte.

[38] Les scores attribués aux différentes fonctions sont arbitraires (notamment pour ce qui est de la séparation entre les objets indirects et les autres compléments).

direct « son », 50 (score : 200), l'objet indirect « Michèle Alliot-Marie », 30 (score : 240), et le complément de manière « lui », 20 (170).

> Moins virulent, [Patrick Devedjian]ɪ a [lui]ȷ aussi manifesté [son]ĸ soutien à [Michèle Alliot-Marie]ɪ.

4.1.3 Poids global de chaque candidat

Un paramètre lié au genre textuel (préférence d'un type d'expression référentielle particulier pour le premier maillon d'une chaîne de référence) est enfin utilisé pour augmenter le poids (+50) de certains candidats. Ainsi, les candidats ayant le score total (accessibilité globale + fonction syntaxique + paramètre du genre textuel) le plus élevé sont sélectionnés comme premiers maillons potentiels d'une chaîne de référence. Une chaîne de référence est alors ouverte pour chaque candidat retenu.

Dans notre exemple, le genre textuel (en l'occurrence l'analyse politique) nous indique que ce sont les noms propres qui apparaissent fréquemment en position initiale d'une chaîne de référence. De ce fait, les candidats « Patrick Devedjian » et « Michèle Alliot-Marie » bénéficient de 50. « Patrick Devedjian », qui totalise un score de 360, constitue le premier maillon potentiel d'une chaîne de référence.

> Moins virulent, [Patrick Devedjian]ɪ a [lui]ȷ aussi manifesté [son]ĸ soutien à [Michèle Alliot-Marie]ɪ.

Une fois les premiers maillons des chaînes de référence identifiés, la seconde étape du calcul de la référence vise à rechercher les anaphores possibles de ces candidats pour construire les chaînes de référence.

4.2 Recherche de paires valides

Pour identifier les autres maillons des chaînes de référence (rangs 2, 3 et suivants), le système détermine les liens de coréférence s'établissant entre les candidats d'accessibilité haute (pronoms, démonstratifs) et ceux d'accessibilité basse (noms propres complets, groupes nominaux définis). Le système construit toutes les paires antécédent-anaphore possibles présentes dans un intervalle n correspondant à la distance moyenne (en nombre de phrases) entre les maillons (paramètre dépendant du genre textuel).

Puis, pour départager les paires candidates, nous avons adapté l'approche de

(Gegg-Harrison et Byron, 2004). Ces auteurs proposent une méthode utilisant des contraintes (morpho-syntaxiques et sémantiques) à satisfaire pour chaque paire de candidats (en fonction du type d'expression référentielle traitée) afin de résoudre les anaphores pronominales dans plusieurs langues (anglais et coréen). La méthode proposée par (Gegg-Harrison et Byron, 2004) est elle-même inspirée de la théorie de l'optimalité[39] ((Prince et Smolensky, 1993) décrite dans (Beaver, 2004)). Dans cette approche, si l'antécédent et l'anaphore réfèrent à la même entité du discours, c'est qu'ils vérifient une série de contraintes syntaxiques (même fonction syntaxique entre l'antécédent et l'anaphore[40]), morphosyntaxiques (accord en genre et/ou en nombre) et sémantiques (hyponymes/hypéronymes). Le candidat sélectionné est celui qui a validé le maximum de contraintes.

Ainsi, pour une anaphore donnée, nous vérifions les contraintes pour tous les antécédents possibles. Si une seule paire antécédent-anaphore valide le maximum de contraintes, alors elle représente un candidat plausible pour participer à une chaîne de référence. Si plusieurs paires antécédents-anaphores satisfont le même nombre de contraintes, plusieurs chaînes de référence peuvent être générées.

La théorie de l'optimalité limite l'espace de recherche de l'antécédent à la phrase précédente. (Gegg-Harrison et Byron, 2004) proposent un algorithme uniquement pour la résolution pronominale. Nous avons élargi la série des contraintes aux autres catégories d'anaphores (groupes nominaux, pronoms réfléchis).

Nous avons ainsi utilisé plusieurs propriétés linguistiques employées traditionnellement par les systèmes de résolution d'anaphore et appliqué des contraintes fortes (éliminatoires si elles ne sont pas vérifiées) et faibles (non éliminatoires) pour filtrer certaines paires de candidats impossibles ou, au contraire, identifier des candidats valables. Lorsqu'une paire ne valide pas une contrainte forte, la paire est éliminée de la liste de candidats. Sur ce dernier aspect, nous nous démarquons de l'approche de (Beaver, 2004) qui hiérarchise ses six contraintes mais ne définit pas de contraintes éliminatoires. En d'autres termes, dans l'approche de Beaver, même si une paire de candidats ne valide pas la contrainte la plus forte, la paire n'est pas éliminée d'office. Pour toutes les paires ayant validé l'ensemble des contraintes fortes, les contraintes faibles sont alors comptabilisées. Décrivons à présent chacune des contraintes utilisées dans *CalcRef*.

[39] *Cf.* chapitre 2, section 3.

[40] Bien que nous reconnaissons que ce ne soit pas toujours le cas.

4.2.1 Les contraintes fortes

Les contraintes dites fortes doivent obligatoirement être satisfaites :

- **IMB** : la contrainte d'imbrication permet d'éliminer des paires où les deux candidats sont imbriqués. Par exemple, lorsqu'un complément de nom est inclus dans son antécédent (*e.g.* [les répercussions [du changement climatique]$_i$]$_j$*)*, il ne peut pas être coréférent. De même, si les deux éléments de la paire sont des co-arguments d'un verbe, alors la paire n'est pas valide. Par exemple dans « [le véhicule]$_k$ qu'[il]$_l$ a acheté », les éléments k et l sont les arguments du verbe « acheter » (*acheter* [*il*$_l$, *véhicule*$_k$]), donc « véhicule » ne peut pas être un bon antécédent pour l'anaphore « il ».
- **TETELEX** : la contrainte tête lexicale vérifie si les têtes lexicales des groupes nominaux sont identiques (p.e. « le jeune homme » - « cet homme ») ou la reprise partielle du même nom propre (« Barack Obama » - « Obama »).

Pour certaines anaphores, il est utile de définir à l'avance l'ensemble des contraintes fortes à valider et celles qui ne s'appliquent pas (car elles sont non pertinentes dans ce cas). Par exemple, pour les possessifs ou les réfléchis, la contrainte d'imbrication IMB ne s'applique pas.

4.2.2 Les contraintes faibles

Pour chaque paire de candidats ayant satisfait l'ensemble des contraintes fortes, le système compte le nombre de contraintes faibles validées. Lorsque plusieurs paires satisfont le même nombre de contraintes, le système extrait les paires valides à partir d'une grande liste.

A la différence des contraintes fortes, les contraintes faibles peuvent être violées :

- **MORPHO** : la contrainte morphologique vérifie la compatibilité en genre ou en nombre entre les deux candidats (par exemple, entre un pronom personnel et le candidat).
- **PROX** : la contrainte de proximité est validée si l'antécédent et l'anaphore sont des proches voisins (pour les possessifs et les démonstratifs).

- **SYN :** la contrainte syntaxique s'attache à vérifier si l'antécédent et l'anaphore possèdent la même fonction syntaxique. Dans l'exemple suivant :

  > La Commission a présenté au Parlement [le rapport annuel concernant l'emploi de la langue française]$_{\mathbf{m}}$. [Ce rapport]$_{\mathbf{n}}$ a tiré un signal d'alarme […].

  la fonction syntaxique des deux éléments *m* et *n* est différente. Mais, bien que la contrainte syntaxique soit violée ici, les deux éléments sont bien reliés par une relation de coréférence.

- **SEM :** la contrainte sémantique permet de sélectionner un antécédent valide parmi plusieurs antécédents potentiels. Pour ce faire, nous avons appliqué la méthode proposée par (Dagan et Itai, 1991). Nous avons constitué une ressource[41] à partir d'un corpus de 500 000 *tokens* issu de revues scientifiques (informatique). Nous avons extrait les occurrences des verbes principaux et de leurs sujets. Par exemple, si nous recherchons un antécédent pour le pronom *il* dans :

  > [Un virus]$_{o}$ a été trouvé dans mon ordinateur. A cause de ce virus, [l'ordinateur]$_{p}$ tourne lentement. **Il** envoie des messages de publicité.

  Les deux groupes nominaux « un virus » et « l'ordinateur » satisfont le même nombre de contraintes. Pour déterminer lequel des deux est un candidat valide, le système parcourt la liste des paires sujets-verbes disponibles dans la ressource. Le verbe « envoyer » possède plusieurs occurrences pour « virus » en tant que sujet, mais aucune occurrence pour « ordinateur ». L'antécédent préféré de « il » est donc « virus » (car la fonction d'envoi de message est spécifique à une application et non à l'ordinateur lui-même).

- **SEM_NER :** cette contrainte sémantique permet de contrôler s'il existe une relation entre l'antécédent et l'anaphore : un nom propre de personne et un nom de fonction (*e.g.* [*Philippe Dupuy – le président du conseil statutaire*]). Afin d'établir les liens de coréférence entre des noms de personnes et des noms de fonctions (*i.e.* grades, fonctions politiques, titres, métiers), nous nous appuyons sur des heuristiques telles que :

  > « si le candidat A (le nom de personne) est une entité nommée de type « personne » et que la tête nominale du candidat B (le nom

[41] Ce travail a été réalisé en collaboration avec Yannick Lutz, dans le cadre de son stage de Master 2 recherche en 2010.

> de fonction) fait partie d'une liste de noms de fonctions, alors il existe une relation de coréférence entre le candidat A et le candidat B ».

Par exemple, dans :

> « *Barack Obama* a promu lundi l'entrepreneuriat, dans le but de doper la création d'emplois. *Le président des Etats-Unis* a également dit qu'il demanderait au Congrès de rendre permanentes les exemptions d'impôt sur les revenus du capital pour les PME, et de voter d'autres allègement fiscaux pour ce secteur. » (20 minutes, 31/01/2011)

la réalisation de la contrainte **SEM_NER** consiste à vérifier si « Barack Obama » est un nom de personne et si la tête du candidat antécédent « Le président des Etats-Unis » (*i.e.* « président ») est bien un nom de fonction. Vu que ces deux conditions sont vérifiées, la paire candidate (Barack Obama) – (Le président des Etats-Unis) est validée.

4.2.3 Représentation des contraintes

En nous inspirant du modèle de représentation de la validation des contraintes utilisé par (Beaver, 2004), nous proposons de revenir sur l'exemple suivant :

> Moins virulent, [Patrick Devedjian]ı a [lui]ȷ aussi manifesté [son]ĸ soutien à [Michèle Alliot-Marie]ı.

Lors de l'étape du calcul des maillons initiateurs de chaînes de références, le nom propre « Patrick Devedjian » (*i*) est arrivé en tête en totalisant un score de 360 (position sujet, catégorie « préférée » pour le genre textuel). Pour chaque maillon initiateur de chaîne, une liste des candidats anaphoriques potentiels est établie (les candidats doivent avoir une accessibilité globale inférieure ou égale à 190). Dans notre exemple, nous avons le pronom *j* et le possessif *k*.

Pour chaque candidat anaphorique, nous construisons ce que (Beaver, 2004) nomme un « tableau » où sont représentés : à l'horizontale l'initiateur de chaîne (*i* dans notre exemple) et à la verticale les candidats anaphoriques potentiels. Lorsqu'une contrainte est violée pour une paire antécédent-anaphore, elle est représentée par un astérisque '*'. Lorsque la contrainte est validée, c'est l'espace qui est utilisé. Lorsqu'une contrainte ne s'applique pas, elle est schématisée par un tiret '-'. Dans le Tableau 27, est représentée la validation des contraintes pour le candidat *i*.

Les paires qui satisfont le plus de contraintes sont [i, k] et [i, j] (contrainte de proximité). La paire [i, l] s'avère impossible car la contrainte forte TETELEX est violée.

	Id	**MORPHO**	**IMB**	**SYN**	**SEM**	**PROX**	**TETELEX**
i	j			*	-		-
	k	-		-	-		-
	l			*	-		*

Tableau 27 – Validation des contraintes pour le candidat i

4.3 Regroupement des paires

Une fois les relations anaphoriques établies, le système regroupe les anaphores ayant un antécédent commun dans une même chaîne de référence (Ailloud et Klenner, 2009). Pour construire les chaînes de référence, le système applique la propriété de transitivité :

> *si A est un antécédent de B et B un antécédent de C, alors ces trois éléments font partie de la même chaîne de référence.*

En reprenant l'exemple ci-dessus, à partir des paires {« Patrick Devedjian » – « son »} ; {« Patrick Devedjian » – « lui »} ; l'application de la propriété de transitivité permet d'obtenir une chaîne de référence composée de trois maillons : {« Patrick Devedjian », « lui », « son »}.

Le processus se poursuit jusqu'à ce que la longueur de la chaîne de référence en cours soit plus longue que la longueur moyenne d'une chaîne de référence suivant le genre textuel. *CalcRef* relance le processus après avoir sélectionné le candidat suivant au poste de premier maillon (potentiel) de chaîne de référence.

4.4 Bilan

A partir du texte étiqueté et enrichi en annotations, *CalcRef* sélectionne des expressions référentielles au poste de premier maillon potentiel d'une chaîne de référence. Puis, *CalcRef* sélectionne les antécédents possibles qui valident des contraintes fortes (éliminatoires) et faibles (non éliminatoires). Les paires antécédents-anaphores valides sont enfin regroupées en chaînes de référence lorsqu'elles partagent le même référent.

Dans le chapitre suivant, nous évaluons les divers modules de *RefGen* : le module d'annotations (*RefAnnot*) ainsi que le module de calcul de la référence (*CalcRef*).

Chapitre 8

Evaluation de *RefGen*

L'évaluation est une étape cruciale dans le développement d'un outil de TAL. Elle permet de rendre compte de la qualité de l'outil et en dévoile aussi les limites. Depuis la fin des années 80, diverses campagnes d'évaluation se sont

succédées[1]. L'objectif de ces campagnes est de confronter plusieurs systèmes à partir du même jeu de données (corpus d'entraînement et corpus de test) dans un temps limité. Ces campagnes d'évaluation portent sur une ou plusieurs langues et comportent une ou plusieurs tâches en recherche d'information appliquées à des textes oraux et/ou écrits : extraction d'entités nommées, résolution de la coréférence, détection d'opinion, segmentation automatique, résumé, scénario, question/réponse, détection de thèmes, etc.

Les campagnes d'évaluation ont donné lieu à la création de corpus de référence pour plusieurs langues : l'anglais, l'espagnol, l'italien, le catalan, l'allemand, le danois, le chinois, le japonais, le coréen, le basque, le tchèque, l'estonien, le suédois, le français. Néanmoins, pour la tâche de résolution de la coréférence, aucun corpus de référence pour le français n'est actuellement disponible[2]. La preuve en est que, lors de la compétition SemEval 2010 – tâche 1 *« Coreference*

[1] Parmi les nombreuses campagnes d'évaluation internationales, on peut citer :

- les conférences MUC (*Message Understanding Conference*) à l'initiative de l'ARPA, qui ont eu lieu de 1987 (MUC-1) à 1998 (MUC-7), ont proposé des tâches en recherche d'information (extraction d'entités nommées, coréférence, scénarios, etc.) pour les textes écrits en anglais (Grishman et Sundheim, 1996),
- les séries d'évaluation Senseval (de 1998 à 2004), auxquelles ont succédé SemEval (*Semantic Evaluation*) depuis 2007, portent sur la désambiguïsation lexicale et l'évaluation de l'analyse sémantique (entités nommées et coréférence principalement) sur plusieurs langues,
- ACE (*Automatic Content Extraction*) depuis 1999, pour la détection et la caractérisation des entités, des relations et des événements (Doddington *et al.*, 2004),
- TDT (*Topic Detection and Tracking*), de 1998 à 2002, pour la détection de thèmes,
- les campagnes d'évaluation CoNLL (*Conference on Natural Language Learning*) tenues chaque année depuis 1997, portent sur l'apprentissage automatique. Elles se sont attaquées à la modélisation de la coréférence dans la tâche « *Modeling Unrestricted Coreference in OntoNotes* » de l'édition 2011,
- les conférences TREC (*Text REtrieval Conference*) depuis 1992, pour la recherche d'information multilingue, le filtrage d'information, ainsi que sur des tâches telles que la coréférence dans des systèmes de question/réponse, la segmentation de documents vidéo, etc.

Spécifiquement pour le français, on peut citer :

- les campagnes ESTER 1 et 2, pour la reconnaissance d'entités nommées, depuis 2005, auxquelles succède actuellement le projet ETAPE (http://www.afcp-parole.org/etape.html),
- les ateliers DEFT (*DEfi Fouille de Texte*) depuis 2005, avec des tâches allant de la reconnaissance de thèmes à la détection des opinions,
- les campagnes d'évaluation autour des entités nommées du programme sur le traitement automatique de documents multimédias Quæro (http://www.quaero.org), depuis 2008.

[2] Un corpus de référence du français parlé annoté en coréférence, ANCOR (Muzerelle *et al.*, 2013), a été mis à disposition en novembre 2013, voir chapitre 5.

Resolution in Multiple Languages », le français n'était pas représenté. De ce fait, pour évaluer notre système *RefGen*, nous devions d'abord constituer notre propre corpus de référence[3].

L'évaluation de l'étiqueteur TTL ayant été effectuée dans la section 2.3 du chapitre 7, la présente évaluation porte sur les modules *RefAnnot* et *CalcRef*. Ainsi, nous avons évalué *RefGen* de deux manières[4] :

a) manuellement, sur un corpus de rapport publics issus de la *Commission des Communautés Européennes* (7230 mots). Pour cette évaluation manuelle, nous avons utilisé les mesures d'évaluation classiques : le rappel, la précision et la f_mesure (cf. infra 1.1). Nous avons évalué chacun des modules de *RefGen* (les trois modules de *RefAnnot*, les paires antécédent-anaphores et les chaînes de référence du module *CalcRef*). En complément des mesures classiques, nous avons aussi calculé le *Slot Error Rate* (Makhoul *et al.*, 1999) spécifiquement pour les entités nommées, afin d'évaluer le type d'erreur d'identification des entités (erreur sur la délimitation des bornes de l'entité et/ou sur le type de l'entité).

b) automatiquement, sur un corpus composé de genres textuels variés pour tester la robustesse de *RefGen*. Pour cette évaluation, nous avons uniquement évalué la sortie finale de *RefGen* (les chaînes de référence complètes). Nous avons utilisé les 4 métriques d'évaluation actuelles de la coréférence : MUC, B^3, CEAF et BLANC.

Dans cette partie, nous rappellerons les diverses mesures utilisées pour la double évaluation de *RefGen*. Nous présenterons ensuite chacune des deux évaluations et discuterons les résultats obtenus.

[3] Le corpus de référence que nous présentons ci-après est restreint en taille, tant la tâche d'annotation se révèle complexe.

[4] Il nous est paru difficile d'effectuer une évaluation comparative de chacun de nos modules de *RefGen* avec des modules de systèmes équivalents car :

- pour le module d'identification des groupes nominaux complexes, il n'existe pas d'équivalent pour le français à notre connaissance,
- pour le module d'identification des *il* impersonnels, le système Il_{Imp} développé par (Danlos, 2005) n'utilise pas le même système d'étiquetage que notre module,
- pour le module d'identification des entités nommées, la granularité des annotations que nous obtenons n'est pas similaire aux autres outils disponibles pour le français. Par exemple, CasEN (Friburger, 2002 ; Friburger et Maurel, 2004) annote plus finement les entités de type fonction (fonction politique, fonction administrative, etc.) et il ne découpe pas les entités nommées complexes de la même manière.
- Pour *CalcRef*, les systèmes existants du calcul de la référence n'identifient pas les mêmes types d'expressions référentielles (ils se restreignent souvent à la référence pronominale) et utilisent la notion de coréférence, non soumise à un nombre minimal de maillons (de l'ordre de 3 dans notre approche).

1 Mesures utilisées

Pour l'évaluation manuelle de *RefGen*, nous avons utilisé les mesures d'évaluation classiques (rappel, précision, f-mesure). Spécifiquement pour les entités nommées, nous utilisons aussi le *Slot Error Rate* pour prendre en compte les erreurs de délimitation des frontières (bornes) de l'entité nommée et celles de classe (type) d'entité nommée.

Pour l'évaluation automatique des chaînes de référence complètes, nous avons utilisé les métriques actuelles d'évaluation de la coréférence (MUC, B^3, CEAF et BLANC).

1.1 Mesures d'évaluation classiques

Sous l'impulsion de la campagne d'évaluation MUC (*Message Understanding Conference*), trois mesures ont été définies pour évaluer les sorties des systèmes, le rappel, la précision et la f_mesure (van Rijsbergen, 1979) :

- Le **rappel** est une mesure de quantité. C'est le rapport établi entre le nombre d'informations correctement identifiées par le système et le nombre total d'informations contenues dans la référence (*i.e.* le nombre total d'informations à identifier), soit :

$$Rappel = \frac{nombre\ d'informations\ correctement\ identifiées}{nombre\ total\ d'informations\ attendues}$$

- La **précision** est une mesure de qualité. Il s'agit du rapport entre le nombre d'informations correctement identifiées par le système et le nombre d'informations ramenées par le système, soit :

$$Précision = \frac{nombre\ d'informations\ correctement\ identifiées}{nombre\ total\ d'informations\ identifiées}$$

- La **f_mesure** est une mesure de synthèse du rappel et de la précision : c'est la moyenne pondérée de ces deux mesures. Dans la formule ci-dessous, la valeur attribuée au coefficient β permet d'attribuer plus d'importance au rappel ($\beta > 1$) ou bien à la précision ($\beta < 1$) ou d'équilibrer les deux mesures ($\beta = 1$) :

$$f_{mesure} = (1 + \beta^2) \times \frac{Précision \times Rappel}{\beta^2 \times Précision + Rappel}$$

1.2 Le Slot Error Rate (SER)

Le *Slot Error Rate* (SER) est une mesure d'erreur définie par (Makhoul *et al.*, 1999)[5] qui combine et pondère différents types d'erreurs. Elle a notamment été utilisée comme mesure principale dans la campagne d'évaluation ESTER 2 (Galliano *et al.*, 2009). Cette mesure fournit un taux d'erreur qui prend en compte plusieurs types d'erreurs :

- l'insertion (I), lorsque le système a annoté une entité qui ne figurait pas dans la référence[6]
- la suppression (D), lorsque le système n'a pas annoté une entité qui figurait dans la référence
- le bornage (B), lorsque le système n'a pas correctement délimité les frontières de l'entité
- le typage (T), lorsque le système n'a pas attribué la classe adéquate à l'entité
- le bornage et typage (BT), lorsque le système a fait à la fois une erreur de classe et de frontières.

Le SER est alors le rapport du coût associé à chaque type d'erreur (0 si l'annotation est correcte ; 0,5 pour une erreur de typage ou de bornage ; 1 pour une insertion ou une suppression) sur le nombre d'entités contenues dans la référence, soit[7] :

$$Slot\ Error\ Rate = \frac{I + D + 0{,}5 \times (B + T) + BT}{nombre\ total\ d'entités\ de\ la\ référence}$$

A la différence des mesures d'évaluation classiques dont le score est compris entre 0 (0%) et 1 (100%), le SER peut être supérieur à 1. Lorsque le SER est supérieur à 1, cela signifie que le système est encore moins performant que s'il n'avait fourni aucune réponse. Par exemple, si le système n'a reconnu aucune entité, son SER sera de 100%. Si ce même système n'a reconnu aucune entité et qu'il a en plus annoté 2 entités qui ne figuraient pas dans la référence, son SER sera alors de 120% (Makhoul *et al.*, 1999 : 251).

[5] C'est une mesure analogue au *Word Error Rate* (WER) utilisé pour évaluer les performances des systèmes de transcription de la parole.

[6] La référence est le nombre d'entités attendues.

[7] Nous reprenons les poids attribués dans le programme Quæro (Grouin *et al.*, 2011 : 51).

1.3 Les mesures MUC, B³, CEAF et BLANC pour la coréférence

Trois mesures d'évaluation sont couramment utilisées pour évaluer les performances des systèmes de détection de la coréférence : MUC (Vilain *et al.,* 1995), B-Cubed (Bagga et Baldwin, 1998) et CEAF (Luo, 2005)[8]. Depuis les campagnes d'évaluation SemEval 2010, une autre mesure est utilisée : BLANC (Recasens, 2010). Mais, bien que (Popescu-Belis, 2000) l'ait déjà signalé, il n'existe toujours pas de consensus, à l'heure actuelle, sur les métriques d'évaluation de la coréférence : *"there is no agreement at present on a standard measure for coreference resolution evaluation"* (Recasens *et al.*, 2010 : 4). De ce fait, pour l'évaluation automatique de *RefGen*, nous avons choisi d'utiliser les quatre métriques d'évaluation actuelles de la coréférence : MUC, B³, CEAF et BLANC. Ces mesures ont en commun : a) le fait qu'elles procèdent en une comparaison des chaînes de référence identifiées par le système (*S*) avec celles de la référence (*T*) (*e.g.* les chaînes de référence annotées manuellement) et b) elles rendent compte des performances en terme de rappel et de précision. Ces mesures diffèrent cependant dans leur méthode pour calculer ces scores, chaque score produisant un biais différent (Trouilleux *et al.*, 2000 ; Trouilleux, 2001 ; Denis, 2007 ; Denis et Baldridge, 2009 ; Recasens, 2010 ; Recasens et Hovy, 2011).

– La métrique **MUC** (Vilain *et al.*, 1995) est une mesure d'évaluation basée sur le calcul des **liens** communs à *S* et *T*. Le rappel est le rapport établi entre le nombre de liens communs à *S* et *T* et le nombre total de liens dans *T*. La précision est le rapport entre le nombre de liens communs à *S* et *T* et le nombre total de liens dans *S*. Ainsi, le rappel pénalise les suppressions tandis que la précision pénalise les insertions. Nous reprenons ci-après les formules de calcul et de précision de MUC adaptées par (Denis, 2007)[9] :

$$Rappel_{\text{MUC}} = \frac{\sum_{S \in S \cap T \in T \neq \emptyset} |S \cap T| - 1}{\sum_{T \in T} |T| - 1}$$

$$Précision_{\text{MUC}} = \frac{\sum_{S \in S \cap T \in T \neq \emptyset} |S \cap T| - 1}{\sum_{S \in S} |S| - 1}$$

[8] D'autres mesures sont aussi utilisées : ACE (NIST, 2004), l'information mutuelle H (Downey *et al.*, 2005), Pairwise F1 (Manning *et al.*, 2008), Rand (Rand, 1971), etc.

[9] Dans les formules : S est une chaîne de *S*, T est une chaîne de *T*, |S|-1 est le nombre de liens de la chaîne S, |T|-1 est le nombre de liens de la chaîne T, $S \cap T - 1$ est le nombre de liens communs entre S et T.

(Bagga et Baldwin, 1998) et (Luo, 2005) ont mis en évidence deux lacunes majeures dans la mesure MUC. D'une part, cette mesure est indulgente dans le sens où elle ne compte que le nombre de liens manquants et de liens faux. De ce fait, le rattachement erroné d'une mention (*e.g.* un maillon d'une chaîne de référence) à une entité (*e.g.* un référent) compte pour une erreur de rappel et de précision alors que la fusion de deux entités ne compte que pour une erreur de rappel (bien que ce dernier cas soit plus éloigné de la réponse attendue). Ainsi, cette mesure favorise-t-elle les systèmes qui identifient des chaînes longues : par exemple, un système qui ne proposerait qu'une seule chaîne pour tout le document obtiendrait un score supérieur à n'importe quel système déjà publié ((Finkel et Manning, 2008) cité par (Recasens et Hovy, 2011)). D'autre part, la métrique MUC ne prend en compte que les liens de référence (deux mentions au moins), ce qui signifie que les ajouts erronés de singletons (entités mentionnées une seule fois) dans S ne sont pas comptabilisés. Le score baisse uniquement si une mention n'est pas rattachée à la bonne chaîne.

– La métrique **B-Cubed** ou **B³** (Bagga et Baldwin, 1998) calcule le rappel et la précision de chaque **mention** en incluant les singletons. Dans B^3, le rappel d'une mention (m) est le rapport entre le nombre de mentions communes entre la chaîne trouvée par le système (S_m) et la chaîne de la référence (T_m) (*e.g.* $|S_m \cap T_m|$) et le nombre total de mentions de T_m. La précision de m est le rapport entre le nombre de mentions dans $|S_m \cap T_m|$ et le nombre total de mentions (M) dans S_m. Les taux de rappel et de précision pour le document sont obtenus en calculant la moyenne du rappel et de la précision de chaque mention, soit :

$$Rappel_{B3} = \frac{1}{|M|} \sum_{m \in M} \frac{|S_m \cap T_m|}{|T_m|}$$

$$Précision_{B3} = \frac{1}{|M|} \sum_{m \in M} \frac{|S_m \cap T_m|}{|S_m|}$$

B^3 permet de résoudre la lacune rencontrée par MUC pour traiter les singletons. De plus B^3 ne favorise pas les chaînes longues par rapport aux chaînes courtes. Néanmoins, cette mesure ne pénalise pas l'insertion ou la suppression d'une mention dans une chaîne lors de la comparaison des

mentions des chaînes de la référence avec les mentions des chaînes fournies par le système.[10]

– La métrique **CEAF** « Constrained Entity Aligned F-Measure » (Luo, 2005) est basée sur les entités (mentions ou chaînes). Dans cette mesure, les entités ne peuvent être utilisées qu'une seule fois : chaque chaîne du système (S) est associée à une vraie chaîne de la référence (T) au plus[11]. CEAF calcule les meilleures associations possibles entre l'ensemble des chaînes S et T, G(S, T). La meilleure des associations (g^*) est celle qui maximise la similarité totale $\Phi(g)$ pour une association g qui est la somme des paires de similarité $\phi(S_i, T_i)$[12] entre les paires des chaînes alignées S_i et T_i. Ainsi, le rappel est le rapport entre la similarité totale de g* et le nombre total de mentions dans T. La précision est le rapport entre la similarité totale de g* et le nombre total de mentions dans S, soit :

$$Rappel_{\text{CEAF}} = \frac{\Phi(g*)}{\sum_i \phi(T_i, T_i)}$$

$$Précision_{\text{CEAF}} = \frac{\Phi(g*)}{\sum_i \phi(S_i, S_i)}$$

Le fait que CEAF ne permette pas de réutiliser une chaîne peut entraîner la perte de liens de coréférence corrects. Par exemple, si le système fournit une chaîne `{m1, m2, m3, m4, m5}` alors que la référence contient deux chaînes `{m1, m2, m3}` et `{m4, m5}`, avec CEAF, seule une chaîne de la référence va être alignée avec la chaîne du système. La seconde chaîne de la référence, pourtant bien présente dans la chaîne du système (lien `m4 – m5`), ne sera pas testée (Denis et Baldridge, 2009 ; Cai et Strube, 2010).

– La métrique BLANC « BiLateral Assessment of Noun-phrase Coreference » (Recasens, 2010) est une mesure récemment développée pour résoudre les problèmes rencontrés par les autres mesures. C'est une variante du Rand Index (Rand, 1971)[13] adaptée pour la résolution de la

[10] Ces mentions qui ne possèdent pas d'équivalent dans l'une ou l'autre des parties (référence ou système) sont appelées *twinless* par (Stoyanov *et al.*, 2009 : 660) qui ont proposé des améliorations à l'algorithme de B^3.

[11] Cela n'est pas le cas avec les mesures MUC ou B^3 qui autorisent la possibilité que chaque chaîne puisse être utilisée plusieurs fois.

[12] $\phi(S_i, T_i)$ est le nombre de mentions communes aux deux chaînes.

[13] Le Rand index utilise :

- le nombre de paires qui se situent à la fois dans la référence et dans le système
- mais aussi le nombre de paires qui se situent dans des chaînes différentes à la fois dans la référence et dans le système.

coréférence. L'objectif est d'obtenir une granularité d'évaluation fine permettant une meilleure distinction des performances des différents systèmes. Dans son calcul, BLANC prend en compte non seulement les liens de coréférence mais aussi les liens de « non-coréférence » (Recasens et Hovy, 2011 : 495), c'est-à-dire la présence de singletons. Le rappel et la précision de BLANC sont calculés séparément pour les deux types de liens (coréférence ou non-coréférence, voir Tableau 28).

Score	**Coréférence**	**Non-coréférence**
Rappel	$Rappel_c = \frac{rc}{rc + wn}$	$Rappel_n = \frac{rn}{rn + wc}$
Précision	$Précision_c = \frac{rc}{rc + wc}$	$Précision_n = \frac{rn}{rn + wn}$

Tableau 28 - Rappel et précision de la métrique BLANC

Dans les scores, *rc* désigne le nombre de liens de coréférence justes (*i.e.* les liens présents à la fois dans la référence et dans les résultats du système), *wc* désigne le nombre de liens de coréférence faux, *rn* les liens de non-coréférence justes et *wn* le nombre de liens de non-coréférence faux.

L'algorithme de BLANC fait alors la moyenne du score pour la détection des singletons et du score pour la détection des liens de coréférence, ce qui permet d'atténuer l'impact des singletons sur les résultats (ce que les métriques B^3 et CEAF ne parvenaient pas à effectuer).

$$RAPPEL_{BLANC} = \frac{Rappel_c + Rappel_n}{2}$$

$$PRECISION_{BLANC} = \frac{Précision_c + Précision_n}{2}$$

Ainsi, pour être performant sous BLANC, un système doit obtenir un bon score de rappel et de précision à la fois pour l'identification des liens de coréférence et de non-coréférence.

2 Evaluation manuelle

Nous avons procédé à une première évaluation du module *RefGen*. Nous avons ainsi comparé l'étiquetage fourni par TTL ainsi que les diverses annotations de *RefAnnot* (entités nommées, groupes nominaux complexes et *il* impersonnel) et le calcul de la référence *CalcRef* fournis par *RefGen* par rapport à une annotation manuelle. Nous avons aussi testé la pertinence des paramètres liés au genre dans le calcul de la référence. Nous reprenons, en détails, les diverses étapes de cette évaluation.

2.1 Corpus d'évaluation

Le corpus d'évaluation (*i.e.* la référence) est composé de rapports publics issus de la Commission des Communautés Européennes (EUROSFAIRE[14], service d'accès à l'information sur la recherche en Europe) d'avril 2009. L'extrait choisi traite des mesures prises pour limiter les effets du changement climatique et compte 7230 *tokens*.

Nous avons annoté manuellement les expressions référentielles nominales du corpus, soit :

- les entités nommées (noms de personnes, lieux, organisations et fonctions)
- les groupes nominaux définis et indéfinis (simples et complexes)
- les pronoms personnels
- les réfléchis
- les démonstratifs

Les relatifs, les marques d'accord des verbes conjugués ainsi que les participes n'ont pas été annotés, de même que les indices coréférentiels « faibles » (Landragin, 2011) tels que les sujets zéros. En effet, ces derniers s'avèrent particulièrement difficile à identifier automatiquement.

Nous avons ensuite ajouté des indices pour identifier les maillons des différentes chaînes de référence. Dans l'exemple ci-dessous, tous les maillons de la chaîne de référence numéro 17 portent le même indice (_17)[15] :

[14] http://www.eurosfaire.prd.fr/7pc/

[15] Dans l'exemple, seuls sont annotées les expressions référentielles de la chaîne numéro 17.

> Quantité d'informations et d'études sur ce sujet existent déjà, mais force est de constater que ces données ne sont pas partagées entre les différents États membres. L'un des moyens d'améliorer efficacement la gestion des connaissances serait de créer **un centre d'échange d'informations**__17 qui servirait d'outil informatique et de base de données en matière d'incidences du changement climatique, de vulnérabilité et de bonnes pratiques dans le domaine de l'adaptation. **Ce centre d'échange**__17 participerait au système de partage d'informations sur l'environnement, une initiative de collaboration entre la Commission européenne et l'Agence européenne pour l'environnement (AEE) visant à mettre en place, en coopération avec les États membres, un système intégré de partage d'informations sur l'environnement au niveau européen. **Le centre d'échange**__17 devra également **se**__17 fonder sur les données géographiques recueillies dans le cadre de la surveillance globale de l'environnement et de la sécurité (GMES).

2.2 Résultats

2.2.1 Evaluation de TTL

Nous avons procédé à l'évaluation de TTL en comparant ses sorties à une annotation manuelle du corpus d'évaluation. L'annotation manuelle du corpus d'évaluation a suivi la procédure utilisée pour construire le corpus de référence de TTL (étiquetage avec TreeTagger puis Flemm, application de scripts de corrections automatiques et corrections manuelles). TTL obtient des précisions de 98,96% pour l'étiquetage et 99,41% pour la lemmatisation[16].

Les erreurs relevées pour l'étiquetage portent, pour la majeure partie, sur des erreurs dans l'attribution de la catégorie (72% des cas). Cela concerne des sigles et abréviations (par exemple, *AEE* a été étiqueté comme nom commun masculin singulier, *PAC* ou *EIE* comme noms communs féminins singuliers), des noms propres (par exemple, « Arctique » a été étiqueté en tant qu'adjectif), certains adjectifs (par exemple « particuliers » dans « certains particuliers » a été étiqueté comme nom commun) ou adverbes (par exemple, « densément » a été étiqueté comme nom commun). Il s'agit aussi d'erreurs de genre, comme dans « les pays les plus vulnérables », où TTL a attribué de manière erronée le genre féminin à l'adjectif. Les adjectifs sont parfois confondus avec les participes passés (*e.g.* « une adaptation plus poussée »). D'autres erreurs encore proviennent

[16] Nous sommes consciente que les performances obtenues sont à nuancer, compte-tenu de la taille réduite du corpus d'évaluation. Nous souhaitons mener une évaluation sur un corpus plus large, dans les perspectives de ce travail.

d'ambiguïtés non résolues (22,7% des cas), par exemple pour certains articles définis « les » ou « l' », pour lesquels TTL n'a pas tranché sur le genre (*e.g.* « l' » a été étiqueté `(Da-fs, Da-ms)`). Ces dernières erreurs sont aussi présentes pour la lemmatisation.

En effet, les erreurs de lemmatisation touchent aussi certains articles définis. Par exemple, les occurrences de « les » et « l' » possédant une étiquette ambiguë se sont vues attribuer un lemme identique à leur forme. De même, les confusions entre adjectifs et participes passés entraînent des erreurs dans les lemmes (par exemple, dans « une adaptation plus poussée », TTL a attribué le lemme « pousser » à l'adjectif. On relève enfin d'autres erreurs de lemmatisation plus marginales pour les sigles (*e.g.* lemme « avoir » pour *EU*, lemme « ce » pour *CE*) ou le participe « dites », lemmatisé « di » (dans une seule occurrence).

Même si TTL reste encore perfectible, les résultats obtenus sont comparables aux performances d'étiqueteurs état-de-l'art disponibles pour le français, tels que $MeLT_{fr}$ (Denis et Sagot, 2010)[17].

2.2.2 Evaluation de *RefAnnot* et de *CalcRef*

Dans le Tableau 29, nous avons reporté le nombre d'éléments (entités nommées (Ner), groupes nominaux complexes (CNp) et *il* impersonnels (Il_i)) annotés dans la référence ainsi que les mesures de rappel, précision et f_mesure pour chacune des annotations fournies par *RefAnnot* et le calcul de la référence (*CalcRef*).

	RefAnnot			***CalcRef***	
	Ner	**CNp**	**Il_i**	**paires**	**chaînes**
Référence	113	79	43	118	24
Rappel	0,85	0,87	0,91	0,69	0,58
Précision	0,91	0,91	1	0,78	0,70
F-mesure	0,88	0,89	0,95	0,73	0,63

Tableau 29 - Résultats de l'évaluation manuelle des modules de *RefGen*

Pour le calcul de la référence, nous avons séparé l'identification des paires antécédents-anaphores des chaînes de référence à proprement parler. Nous détaillons ci-après les résultats de l'évaluation pour les annotations et le calcul de la référence.

[17] Notons que TTL bénéficie du jeu d'étiquettes MULTEXT, sensiblement plus fin que celui utilisé par $MeLT_{fr.}$

2.2.3 Evaluation des annotations

La performance (f_mesure) pour les annotations est comprise entre 88% (pour les entités nommées) et 95% (pour les *il* impersonnels) (voir Tableau 29). Concernant les entités nommées, la plupart des erreurs ont porté sur les sigles et abréviations (par exemple *AEE, GES, OMS, EU, GISC, CAA, ALE, EIE, PAC, RTE-T*, etc.) qui se sont révélés être un procédé fréquemment utilisé dans le genre textuel du rapport public. Plusieurs types d'erreurs sont à identifier ici :

- soit *RefAnnot* a annoté des abréviations qui n'étaient pas des entités nommées (insertions). Par exemple, les abréviations « GES » (gaz à effet de serre) et « GIZC » (gestion intégrée des zones côtières) se sont vues attribuer l'étiquette `<Org>` alors qu'il s'agissait d'une abréviation de CNp,
- soit il n'existait pas de patron pour typer ces entités nommées (cas d'*AMCC* (Alliance mondiale pour la lutte contre le changement climatique)),
- soit les sigles n'ont pas été identifiés (suppressions) car ils ont été étiquetés par TTL dans une autre catégorie. On trouve par exemple une occurrence d'« EU » (European Union) qui a été étiquetée comme forme conjuguée du verbe « avoir ».

Pour ce qui est des groupes nominaux complexes, les erreurs sont dues soit à une sous-spécification dans nos patrons (le patron était « trop court » à gauche ou à droite), soit elles étaient issues d'erreurs d'étiquetage de TTL. Par exemple, les groupes nominaux coordonnés plus de deux fois ont été mal délimités, parce que nous avions fait le choix de ne pas définir de patron pour éviter d'obtenir trop de bruit (*i.e.* de fausses annotations). Aussi, une ambiguïté maintenue pour l'étiquetage des articles indéfinis pluriels va-t-elle provoquer une « rupture » dans le *chunk* « `Np` ». En conséquence, nos patrons ne pouvaient plus correspondre.

Du côté des *il* impersonnels, nous observons de bonnes performances (95%) : sur 43 emplois impersonnels du pronom *il*, 39 ont bien été reconnus (les 4 cas qui ont échappé au module sont liés à l'absence de patron). Il est à noter que le corpus d'évaluation ne comptait que 4 occurrences de *il* anaphorique. Il serait donc éventuellement prévisible que le système obtienne des résultats moins élevés sur un corpus contenant plus d'occurrences de *il* anaphorique, tant les cas d'ambigüité sont fréquents pour ce phénomène.

2.2.4 Evaluation du calcul de la référence

Le corpus d'évaluation contenait 118 paires antécédent-anaphore et 24 chaînes de référence (suivant notre définition). Nous avons analysé les résultats obtenus par *CalcRef* à la fois pour les paires antécédent-anaphore (pronominales et nominales) retrouvées et le nombre de chaînes de référence complètes retrouvées. Nous avons relevé un nombre important de relations anaphoriques (impliquant surtout des anaphores nominales, démonstratives ou définies) mais peu de chaînes de référence (17% du nombre total de relations anaphoriques). Les performances sont de 73% pour l'identification des paires antécédent-anaphore et de 63% pour les chaînes de référence complètes :

- concernant les paires antécédent-anaphore, les erreurs relevées proviennent d'erreurs d'étiquetage ou du fait que plusieurs antécédents d'un même candidat ont satisfait le même nombre de filtres (filtres pour la fonction syntaxique, le genre, le nombre) comme cela a par exemple été le cas pour les candidats « ils » et « les efforts d'adaptation entrepris par les Etats membres » ou bien pour les candidats « ils » et « des Etats membres ». Leur filtrage nécessiterait des connaissances supplémentaires (*i.e.* l'utilisation d'une ontologie).

- les erreurs d'identification des chaînes de référence sont dues essentiellement aux fausses paires antécédent-anaphore retrouvées mais aussi aux erreurs d'étiquetage de TTL (principalement rencontrées pour les abréviations et sigles). Ce dernier cas pourrait se résoudre en partie par l'ajout de connaissances dans le dictionnaire de l'étiqueteur.

2.2.5 Evaluation des paramètres du genre textuel

Nous avons ensuite procédé à la modification de certains paramètres (Tableau 30) pour tester la pertinence des paramètres spécifiques au genre textuel dans notre calcul de la référence. Les paramètres du genre du corpus d'évaluation (rapports publics) étaient les suivants :

- distance moyenne entre les mentions = 2 phrases,
- longueur moyenne des chaînes de référence = 4 maillons,
- type de premier maillon = groupe nominal défini.

Dans un premier temps, nous avons utilisé les paramètres d'un autre genre textuel, l'analyse politique (distance = 1 phrase, nombre moyen de maillons = 3, type de premier maillon préféré = nom propre) et nous les avons appliqués à notre corpus d'évaluation. Nous obtenons une performance similaire pour les

paires (f_mesure = 0,70) mais une baisse significative pour les chaînes de référence (la f_mesure baisse de 9 points).

Ensuite, nous avons totalement supprimé les paramètres du genre textuel dans notre calcul de la référence en utilisant uniquement une distance intermaillonnaire par défaut de 20 phrases. Ici, nous obtenons une performance comparable pour les paires (f_mesure = 0,71) mais une baisse importante de la f_mesure pour la détection des chaînes de référence (12 points en moins). Le nombre de paires est quasiment le même (quelques paires sont ajoutées entre des candidats éloignés). En revanche, comme certaines chaînes identifiées regroupent des chaînes plus courtes, nous obtenons peu de chaînes valides.

F_mesure	***CalcRef***	
	paires	**chaînes**
avec les paramètres du genre rapport public	0,73	0,63
avec les paramètres du genre analyse politique	0,70	0,54
sans aucun paramètre du genre textuel	0,71	0,51

Tableau 30 – Evaluation de *CalcRef* avec modification des paramètres liés au genre textuel

Ainsi, les deux variations dans la configuration de *CalcRef* ont-elles mené à une baisse des performances de notre système, ce qui nous permet de valider[18] l'importance des spécificités liées au genre textuel pour le calcul de la référence.

2.3 Discussion

Nos premières évaluations en terme de rappel, précision et f_mesure sont strictes (Béchet et *al.*, 2011). Cela sous-entend qu'une annotation automatique est considérée comme correcte si elle correspond exactement à l'annotation humaine. Par exemple, pour l'annotation des entités nommées, l'annotation automatique est correcte si sa délimitation (segmentation) et son typage sont corrects. En comparaison, les mesures utilisées dans les campagnes ESTER2 par exemple attribuent un poids différent aux erreurs strictes, aux erreurs de frontières et aux erreurs de typage.

De même, pour l'évaluation des chaînes de référence, nous avons considéré que la chaîne de référence était correcte uniquement lorsque celle qui était proposée par le système était exactement la même que celle annotée manuellement. Nous

[18] Il est évident que ces résultats seraient à vérifier sur des corpus de plus grande taille.

rejoignons ainsi la procédure adoptée dans la dernière édition de la campagne CoNLL 2011, en l'appliquant aux chaînes de référence :

> « Unlike MUC, or ACE, the OntoNotes data does not explicitly identify the minimum extents of an entity mention, so for the official evaluation, we will consider a mention to be correct only if it matches the exact same span in the annotation key. » (Pradhan *et al.*, 2011 : 19).

Néanmoins, dans les campagnes MUC ou ACE, les correspondances partielles entre les résultats des annotations manuelles et automatiques sont permises. Par exemple, dans les conférences MUC, une mesure d'erreur ERR prend en compte les substitutions, les suppressions et les insertions.

De ce fait, nous avons souhaité calculer le *Slot Error Rate* (Makhoul *et al.*, 1999, voir section 1.2 *supra*) pour prendre en compte les insertions (I), les suppressions (D), les erreurs de bornage (B) et de typage (T). Nous proposons de calculer le SER pour les entités nommées (Ner) de notre corpus d'évaluation manuelle. Pour rappel :

$$SER\_Ner = \frac{I + D + 0{,}5 \times (B + T) + BT}{nombre\ total\ d'entités\ de\ la\ référence}$$

Dans le Tableau 31, nous avons détaillé le nombre d'occurrences de chaque type d'erreur pour l'identification des entités nommées :

Nombre total de Ner	***I***	***D***	***B***	***T***	***BT***	**SER**
113	2	12	3	0	1	**0,14**

Tableau 31 – Calcul du SER pour les entités nommées du corpus d'évaluation manuelle

Les performances pour l'identification des entités nommées (tous types confondus) sont élevées : sur 113 entités nommées, le taux d'erreur est de 14,6%. Nous pouvons remarquer que ce sont les erreurs de suppression qui sont majoritaires (12 cas) car nous avons construit nos règles en privilégiant la qualité de nos annotations. En effet, nous avons préféré ne pas créer de patron d'identification si le contexte n'était pas assez discriminant (*i.e.* lorsque nous n'avions pas ou peu d'indices internes et externes) plutôt que de proposer un typage erroné.

3 Evaluation automatique

Nous avons effectué une évaluation automatique des annotations en chaînes de référence fournies par *RefGen*[19]. Après avoir présenté le corpus d'évaluation et l'outil utilisé pour son annotation manuelle, nous expliquerons les étapes du processus permettant de comparer les annotations manuelles de référence aux annotations automatiques de *RefGen*.

3.1 Annotation du corpus d'évaluation

3.1.1 Corpus d'évaluation

Le corpus d'évaluation est constitué d'extraits issus de plusieurs genres textuels et compte 15192 *tokens* (voir Tableau 32):

GENRE	SOUS-CORPUS	PERIODE	NOMBRE DE TOKENS
presse quotidienne	*L'Est Républicain*[20]	1999 – 2003	3225
résumés de films	*Resumes-films-a*[21]	1940 – 2008	2930
roman	*Les trois Mousquetaires (Dumas)*	1844	3154
lois européennes	*Acquis Communautaire (Steinberger et al.)*	2006	3076
rapports publics	*Commission des Communautés Européennes*	2009	2807
		TOTAL	**15192**

Tableau 32 – Répartition des sous-corpus pour le corpus d'évaluation automatique

Pour annoter manuellement le corpus d'évaluation, nous avons utilisé la plateforme *Glozz*[22] (Widlöcher *et al.*, 2009).

[19] L'évaluation automatique a été effectuée en collaboration avec Eric Vallette D'Osia dans le cadre de son stage de Master Recherche « Linguistique, informatique et traduction » mené à l'Université de Strasbourg en 2010 – 2011.

[20] Le corpus est disponible à : http://www.cnrtl.fr/corpus/estrepublicain/

[21] Les résumés de films sont issus du site : http://joeyy.free.fr/resumes-films-a.htm

[22] Le logiciel *Glozz* et son manuel sont librement téléchargeables à : http://www.glozz.org/

3.1.2 Annotation du corpus avec *Glozz* (Widlöcher *et al.*, 2009)

Glozz est une plateforme d'annotation manuelle et d'exploration de corpus textuels initiée dans le cadre du projet ANR Annodis[23] (Péry-Woodley *et al.*, 2009) puis poursuivi actuellement au sein du laboratoire GREYC (Université de Caen). Cette plateforme développée en Java est indépendante d'un modèle théorique spécifique et elle n'est pas limitée à certains types d'objets linguistiques.

A partir du corpus brut (texte en format UTF-8), il est possible d'annoter et de visualiser des structures simples ou complexes (unités, relations et schémas) en XML *via* l'interface. *Glozz* est ainsi facilement configurable et permet de définir un modèle d'annotation des chaînes de référence.

3.1.2.1 Schéma d'annotations adopté

Dans *Glozz*, nous avons défini notre propre modèle d'annotation[24] où figurent (voir Figure 48) :

- les **unités** à annoter (*i.e.* les expressions référentielles) : pronoms (*il*), possessifs (*son*), réfléchis (*se*), groupes nominaux simples définis (*la vérité*) et indéfinis (*un lien*), groupes nominaux complexes (*les risques du premier tour*), démonstratifs (*ce*), noms propres (*Pierre Denier*);
- les **relations** entre ces unités : les relations anaphoriques (*e.g.* entre deux expressions référentielles, où l'interprétation de l'anaphore dépend de son antécédent) et de coréférence ;
- les **schémas** entre ces relations : les chaînes de référence ((*e.g.* relation établie entre trois expressions référentielles désignant le même référent au moins, dans notre approche).

[23] Le projet *Annodis* vise la création d'un corpus de référence pour l'analyse du discours.

[24] La définition de ce modèle d'annotation s'effectue de manière déclarative suivant un schéma XML simple qui comprend divers éléments (`units`, `relations`, `schema`) pour définir les catégories d'objets à manipuler et leurs valeurs possibles.

Figure 48 – Modèle d'annotation du corpus d'évaluation défini dans *Glozz*

3.1.2.2 Méthode d'annotation

Nous avons choisi d'effectuer l'annotation du corpus d'évaluation en plusieurs « lectures » :

- lors d'une première lecture, nous annotons toutes les expressions référentielles contenues dans le texte (qu'elles soient incluses dans une chaîne de référence ou non) ;
- puis, en seconde lecture, nous relions les diverses expressions référentielles d'une même chaîne de référence deux à deux (relation anaphorique et de coréférence) et nous lions ces couples de relations pour former les chaînes de référence finales (schéma)[25].

Notons que les relations anaphoriques associatives et plurielles ne sont pas annotées vu que nous avons fait le choix de ne pas les traiter dans *RefGen* (*cf.* 3.2.2).

3.1.2.3 Exemple d'annotation

La Figure 49 illustre un extrait d'annotation du corpus d'évaluation issu de *l'Est Républicain* :

> Abdelmalek Benbara, retrouvé mort mercredi dans le coffre de son véhicule, a reçu plusieurs coups de pied ou de poing au visage et trois coups de couteau qui lui ont été portés alors qu'il agonisait. L'un de ces coups lui a brisé le larynx, provoquant son asphyxie. Les deux autres coups ont touché le cœur et le thorax. Toujours selon les constatations du légiste, la mort de M. Benbara est intervenue rapidement. La date de sa mort serait contemporaine à celle de sa disparition. En effet, sa barbe n'a pas eu le temps d'apparaître.

[25] *Glozz* ne permet pas de créer des chaînes de référence en une seule passe.

Dans cet exemple, pour des raisons de lisibilité, seules ont été annotées les expressions référentielles relatives au référent « Abdelmalek Benbara ». Les expressions référentielles apparaissent sous forme de blocs colorés (suivant le code couleur des catégories d'expressions référentielles défini pour notre modèle d'annotation) et peuvent être imbriquées (par exemple, un nom propre peut être inclus dans un groupe nominal complexe : « [Le fils de [Paul]] »). Les relations sont représentées par des lignes en pointillés et les schémas apparaissent via un sur-encadrement en bleu des unités et des relations qui les composent. Les chaînes sont identifiables grâce aux points bleus déportés à gauche.

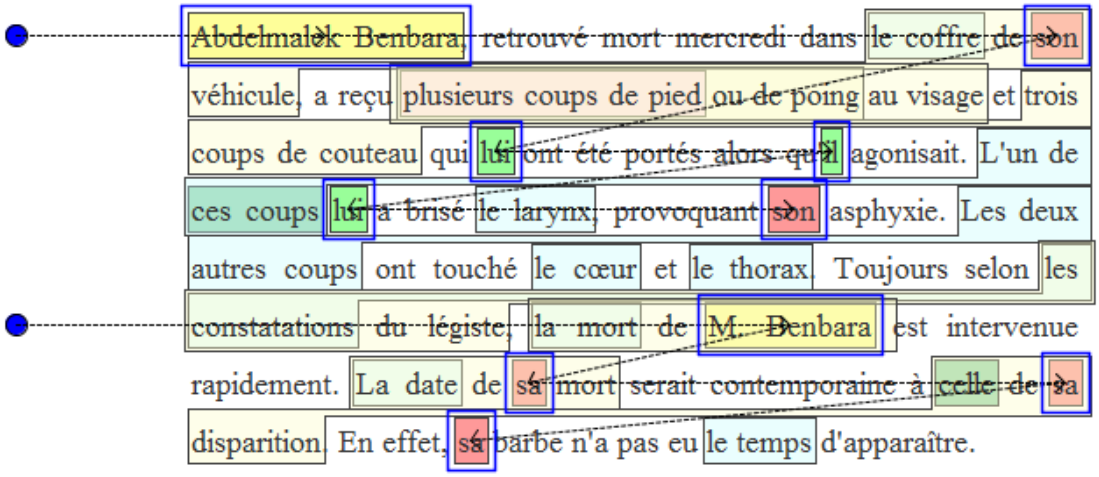


Figure 49 - Exemple d'annotation du corpus d'évaluation avec *Glozz*

Une fois les annotations du corpus d'évaluation effectuées vient l'étape de comparaison (évaluation) automatique des sorties de *Glozz* et de *RefGen*. Pour ce faire, nous avons utilisé *Scorer*[26], le kit d'évaluation de la coréférence issu de la campagne Semeval 2010. *Scorer* permet de calculer automatiquement les quatre métriques MUC, B^3, CEAF et BLANC. Néanmoins, pour utiliser *Scorer*, il nous fallait transformer au préalable les sorties de *Glozz* et de *RefGen* en format compatible (le format CoNLL).

[26] *Scorer* (Perl package for scoring coreference resolution systems using different metrics) version 1.08 est disponible à : http://www.lsi.upc.edu/~esapena/downloads/index.php?id=3

3.2 Transformation des sorties *Glozz* et *RefGen*

Le module *RefGen* fournit directement une sortie en format parenthésé (format Semeval, voir Figure 30, chapitre 6). De même, dans *Glozz*, un export parenthésé est aussi disponible (voir Figure 50).

```
[Abdelmalek Benbara]_16, retrouvé mort mercredi dans le coffre de [son]_16
véhicule, a reçu plusieurs coups de pied ou de poing au visage et trois coups de
couteau qui [lui]_16 ont été portés alors qu'[il]_16 agonisait. L'un de ces coups
[lui]_16 a brisé le larynx, provoquant [son]_16 asphyxie. Les deux autres coups ont
touché le cœur et le thorax. Toujours selon les constatations du légiste, la mort
de [M. Benbara]_17 est intervenue rapidement. La date de [sa]_17 mort serait
contemporaine à celle de [sa]_17 disparition. En effet, [sa]_17 barbe n'a pas eu le
temps d'apparaître.
```

Figure 50 - Exemple de sortie parenthésée dans *Glozz*

Le module Perl *Scorer* permet l'évaluation automatique de la coréférence. Il nécessite en entrée un format inspiré de la tâche CoNLL des éditions 2008 et 2009[27]. Le format CoNLL s'appuie sur la segmentation du texte en *tokens*. Dans cette représentation, chaque mot ou signe de ponctuation est présenté en ligne. Une première colonne identifie chaque *token* par un numéro (ID). Puis, les colonnes suivantes permettent de fournir des informations morphosyntaxiques, les dépendances syntaxiques, le type d'entité nommée, les rôles sémantiques et la coréférence. Par exemple, pour la coréférence dans l'exemple de la Figure 51, le maillon « Abdelmalek Benbara » est identifié par une parenthèse ouvrante et le numéro 16 pour le premier *token* du maillon (*e.g.* « Abdelmalek ») et le numéro suivi de la parenthèse fermante pour le dernier *token* du maillon (*e.g.* « Benbara »). Lorsque les informations ne sont pas disponibles, elles sont remplacées par un _.

ID	token	morpho-syntaxe	dépendances	Ner	rôles sémantiques	Coref
156	Abdelmalek	NNP (NP*	- - - -	*	* (ARG0* (ARG0* *	(16
157	Benbara	NNP *)	- - - -	(PERSON)	* *) *) *	16)

Figure 51 - Exemple de représentation en format CoNLL

Pour pouvoir utiliser *Scorer*, un script a été défini[28]. Ce script permet de transformer l'affichage parenthésé des maillons des chaînes de référence en représentation CoNLL simplifiée. Dans cette représentation CoNLL simplifiée, seules les colonnes `ID`, `token` et `Coref` sont renseignées (les autres colonnes sont remplies par des « _ », vu que nous n'utilisons pas les autres informations ici).

[27] http://ufal.mff.cuni.cz/conll2009-st/scorer.html
[28] Voir (Valette d'Osia, 2011 : 63-64).

Après le passage du script de réécriture, on obtient la sortie suivante (voir Figure 52) :

```
156   Abdelmalek   _   _   _   _   _   _   _   (16
157   Benbara      _   _   _   _   _   _   _   16)
158   ,            _   _   _   _   _   _   _   _
159   retrouvé     _   _   _   _   _   _   _   _
160   mort         _   _   _   _   _   _   _   _
161   mercredi     _   _   _   _   _   _   _   _
162   dans         _   _   _   _   _   _   _   _
163   le           _   _   _   _   _   _   _   _
164   coffre       _   _   _   _   _   _   _   _
165   de           _   _   _   _   _   _   _   _
166   son          _   _   _   _   _   _   _   (16)
167   véhicule     _   _   _   _   _   _   _   _
```

Figure 52 – Exemple de sortie en format CoNLL simplifié

Une fois les sorties *RefGen* et *Glozz* transformées en format CoNLL simplifié, le module *Scorer* a pu être utilisé pour évaluer l'annotation automatique des chaînes de référence.

3.3 Evaluation de l'annotation automatique des chaînes de référence

Nous avons procédé à l'évaluation automatique de trois des cinq sous-corpus constituant le corpus initial : journalistique (presse quotidienne), littéraire (roman) et juridique (lois européennes), suite à des erreurs rencontrées notamment avec l'application de la métrique MUC. Les résultats de l'application des quatre métriques sur les trois corpus sont présentés dans le Tableau 33 :

Corpus	**Identification des mentions**			**MUC**			**B^3**			**CEAF**			**BLANC**		
	R	**P**	**F**	**R**	**P**	**F**	**R**	**P**	**F**	**R**	**P**	**F**	**R**	**P**	**F**
Journalistique	73,1	84,5	78,4	51,0	66,7	57,8	47,1	78,7	58,9	51,2	34,6	44,7	56,9	73,1	59,2
Littéraire	40,3	79,3	53,5	27,3	52,9	36,0	57,9	85,3	69,0	64,9	41,8	50,9	60,0	71,9	63,5
Juridique	22,2	16,0	18,6	100	7,7	14,3	100	68,4	81,3	57,3	85,3	68,8	99,0	53,3	55,7

Tableau 33 - Evaluation automatique de *RefGen* avec les 4 métriques

Les résultats de l'application des métriques montrent des variations significatives suivant le genre textuel : la f_mesure du corpus journalistique est comprise entre

44,7% pour CEAF et 59,2% pour BLANC, celle du corpus littéraire varie entre 36% pour MUC et 69% pour B^3 et la f_mesure du corpus juridique varie de 14,3% pour MUC à 81,3% pour B^3. Entre les divers genres, on observe aussi des différences. Par exemple, pour la métrique MUC, la f_mesure du corpus journalistique est nettement plus élevée que celle du corpus juridique (57,8% contre 14,3%). La tendance est inverse pour la f_mesure de B^3 (58,9% pour le corpus journalistique contre 81,3% pour le corpus juridique). Ces mêmes variations sont observées par exemple dans les résultats officiels obtenus par les participants de la campagne SemEval-2010 (Recasens *et al.*, 2010 : 7), que nous reportons à l'annexe 8. Les tendances obtenues par notre système seraient évidemment à confirmer sur un plus large corpus annoté.

A titre comparatif, sur des données réelles, (Ailloud et Klenner, 2009) relèvent des performances des systèmes actuels de résolution anaphorique (uniquement) allant de 55% à 70% en moyenne. Notre système, qui calcule des relations de coréférence (et pas uniquement anaphoriques), obtient des performances allant de 36,03% pour MUC à 69,73% pour B^3 (soit une moyenne de 55% pour les trois corpus sur les quatre métriques). Notre système obtient donc des performances comparables aux autres systèmes existants, en annotant plus de phénomènes (anaphores et coréférence).

3.4 Discussion

L'évaluation automatique de *RefGen* fournit des résultats variables (et même contradictoires) suivant la mesure utilisée, ce qui rend l'interprétation des performances de notre outil quelque peu difficile. Ce problème est pointé par (De Clercq *et al.*, 2011 : 192) :

> « the different evaluation metrics for coreference research in use today, (MUC, B-cubed, CEAF and BLANC) tend to contradict each other and as a consequence hamper interpretation. This is a well-known problem within the community for which no solution has been found yet. »

De la même manière, (Stoyanov *et al.*, 2010 : 156) estiment que :

> « However, it is still frustratingly difficult to compare results across different coreference resolution systems. Reported coreference resolution scores vary wildly across data sets, evaluation metrics, and system configurations. »

(Popescu-Belis, 2000 : 143) a même été jusqu'à dire « qu'il n'existe pas de mesure unique capable de saisir objectivement la qualité d'un programme de résolution de la référence. ». En effet, aux différents paramètres pris en compte dans les diverses métriques d'évaluation se superposent les différents types de relations entre expressions référentielles identifiées par les algorithmes de résolution de la référence (*i.e.* anaphores pronominales, anaphores nominales, coréférence avec désignation stricte (Popescu-Belis *et al.*, 1998) de l'entité (*e.g. une mesure…cette mesure…elle*), anaphores associatives).

Les organisateurs des campagnes d'évaluation reconnaissent cette difficulté (Recasens *et al.*, 2011). Ainsi, chaque campagne d'évaluation adapte et décide d'un sort différent pour ces métriques afin de départager les participants. Par exemple, pour CoNLL2011, en dépit des quatre métriques disponibles, les organisateurs nécessitaient un unique score pour départager les participants. Ce score a été calculé en utilisant la moyenne non pondérée de MUC, B^3 et CEAF (et BLANC n'a pas été retenu dans ce score final).

Aussi, certains systèmes, conçus pour être performants sur une mesure spécifique, en viennent à obtenir les scores les plus bas lorsqu'on les évalue avec d'autres mesures d'évaluation (Ekbal *et al.*, 2011). Vu qu'il n'existe pas à l'heure actuelle de mesure satisfaisante pour évaluer les systèmes, (Uryupina, 2010 ; Ekbal *et al.*, 2011) ont adapté leur système pour qu'il corresponde à l'une ou l'autre des mesures MUC, CEAF, B^3, BLANC (*i.e.* ils ont modifié leur système pour chaque mesure d'évaluation). Leur système était donc décliné en quatre versions différentes. On peut donc en venir à se demander si les mesures d'évaluation sont mises en place pour évaluer des systèmes développés ou bien si ce sont les systèmes qui doivent être développés pour correspondre aux attentes des métriques…

Nous avons évalué notre système *RefGen* avec les métriques existantes pour la résolution de la coréférence. Toutefois ces métriques utilisent les singletons, les mentions, les liens et la définition de chaîne de référence que nous suivons n'est pas prise en compte en particulier (*i.e.* à partir de 3 expressions coréférentielles). De ce fait, il serait judicieux de définir une métrique adaptée à notre approche (*i.e.* qui prenne en compte le nombre de maillons minimal, les divers types d'expressions référentielles annotés, etc.) ou bien de procéder à une évaluation qualitative des chaînes de référence qui ont été identifiées et des expressions qui n'ont pas été reconnues.

4 Bilan

Dans ce chapitre, nous avons procédé à l'évaluation de TTL et des deux modules de *RefGen* : le module d'annotations (*RefAnnot*) ainsi que le module de calcul de la référence (*CalcRef*). Nous avons évalué *RefGen* en utilisant les mesures d'évaluation classiques (rappel, précision, f_mesure) puis, plus spécifiquement pour les chaînes de référence complètes, nous avons utilisé les métriques actuelles d'évaluation de la coréférence (MUC, B^3, CEAF et BLANC).

Les résultats de l'évaluation de *RefGen* avec les mesures classiques se situent entre 88% et 95% (f_mesure) pour les annotations et entre 63 et 73% pour les chaînes de référence. Nous avons obtenu des erreurs d'identification dans *RefAnnot* et *CalcRef* qui sont dues, pour la plupart, à des erreurs d'annotation dans TTL (mots inconnus ou ambiguïté). Le choix que nous avions fait d'adopter une procédure en chaîne (*i.e.* les annotations de TTL ont servi de point d'entrée à *RefAnnot* puis à *CalcRef*) nous a nécessairement menée à récolter des erreurs issues de chaque étape du processus. Néanmoins, cette méthode nous a permis d'utiliser chaque annotation pour « alléger » le nombre et la complexité de nos règles dans *RefAnnot* par exemple. Aussi, nous avions fait le choix d'annoter uniquement les emplois impersonnels du pronom *il*. Dans l'évaluation de ces emplois, nous aurions pu prendre en compte le fait que nous n'avions pas établi de règles permettant d'identifier les pronoms anaphoriques *il* « sûrs ».

L'évaluation des chaînes de référence avec les métriques d'évaluation de la coréférence a fourni des résultats variables suivant la métrique utilisée et le genre textuel du corpus. Dans l'état actuel de notre travail, nous avons comparé notre système avec un corpus annoté manuellement via *Glozz*. Nous n'avons pas eu à notre disposition des outils de calcul de la référence pour le français. Le manque d'éléments de comparaison (*i.e.* de système d'identification automatique de la coréférence en français qui prennent en compte les mêmes types d'unités (pronoms, noms propres, groupes nominaux) que ceux reconnus par *RefGen*) ne nous permet pas de positionner actuellement notre système. Aussi, serait-il intéressant de comparer notre système à des systèmes par apprentissage, comme BART (Versley *et al.*, 2008) et Corry (Uryupina, 2010). Néanmoins, l'utilisation d'un système par apprentissage nécessiterait la création préalable d'un corpus de référence du français écrit de taille importante annoté en chaînes de référence qui n'existe pas encore à l'heure actuelle (voir chapitre 5).

Conclusion et perspectives

Afin d'optimiser la classification des documents dans les archives internes d'un moteur de recherche industriel, nous avons proposé une nouvelle méthode d'indexation des documents basée sur la détection automatique des thèmes (ATDS-Fr). Dans la lignée des systèmes hybrides de détection de thèmes (Hernandez, 2004), notre méthode utilise un système statistique de segmentation thématique éprouvé, *C99* (Choi *et al.,* 2001), afin de découper les documents en segments textuels thématiquement homogènes. Puis, pour chaque segment, nous identifions des marqueurs linguistiques de cohésion permettant d'attribuer des thèmes à chaque segment thématique.

Comme nous l'avons vu, la notion de *thème* conserve encore à l'heure actuelle des contours flous en linguistique comme en TAL. Pour identifier les thèmes saillants des documents, nous avons montré qu'il était nécessaire d'utiliser la structure textuelle du document (et pas uniquement ses phrases prises isolément), donc qu'il fallait considérer la notion de thème à un niveau global. En suivant (Goutsos, 1997), nous avons défini le thème textuel comme un agrégat des différents thèmes phrastiques, identifié *via* des marqueurs linguistiques de cohésion utilisés seuls ou en combinaison. Nous avons sélectionné deux types de marqueurs fiables pour la détection des thèmes : les cadres de discours (Charolles, 1997), tournés vers l'aval du discours et les chaînes de référence (Schnedecker, 1997), tournées vers l'amont, que nous combinons afin d'obtenir des descripteurs thématiques.

Nous avons ensuite focalisé notre travail sur l'étude, la modélisation et le développement d'un module d'identification des chaînes de référence, *RefGen*, qui constitue le module central du système ATDS-Fr. Les deux études menées sur les chaînes de référence dans des corpus de genres textuels variés ont permis de montrer l'influence du genre sur la composition des chaînes de référence. La typologie des chaînes de référence obtenue à l'issue de ces études a servi à paramétrer notre système d'identification des chaînes de référence *RefGen*. Développé à partir de peu de ressources (étiqueteur et dictionnaire), ce module identifie automatiquement une série d'expressions référentielles (*i.e.* noms propres, pronoms, groupes nominaux simples et complexes, possessifs) qui n'ont pas (ou très peu) été traitées dans leur ensemble dans les systèmes existants (la majorité des systèmes symboliques se restreint à l'identification des anaphores pronominales). A l'heure actuelle, l'algorithme d'identification des chaînes de référence de *RefGen* est fonctionnel : nous obtenons les paires candidates filtrées et l'étape de rapprochement des divers maillons des chaînes est stabilisée.

L'évaluation de *RefGen* avec les métriques d'évaluations de la coréférence actuellement disponibles n'a pas permis de positionner notre système car les paramètres utilisés dans ces mesures ne correspondent pas à la définition des chaînes de référence que nous avons adoptée. Une mesure spécifique serait à mettre en place pour pouvoir évaluer notre travail, mais elle nécessiterait une étude et un développement qui dépassent le cadre de ce travail.

Il paraît évident que la première des perspectives à ce travail serait le développement et l'intégration du système de détection automatique de thèmes ATDS-Fr, qui n'est pour le moment qu'à l'état de modèle. Le développement de notre système consisterait à identifier automatiquement, dans chaque segment thématique fourni par *C99*, les marqueurs lexicaux cadratifs *via* un module comparant les listes d'introducteurs de cadres que nous avons choisi d'identifier (*i.e.* champs thématiques, domaines qualitatifs, espaces de discours) à ceux présents dans le texte et à extraire le complément nominal les succédant. A l'issue de ce traitement, le texte passerait dans *RefGen* pour identifier les chaînes de référence. Puis, un module, testant les diverses configurations possibles entre marqueurs (présence d'un type de marqueur ou des deux, répétition du même premier maillon dans plusieurs chaînes de référence, etc.), fournirait la liste des thèmes et des sous-thèmes du texte.

L'évaluation de TTL et des modules *RefAnnot* et *CalcRef* ayant porté sur des corpus réduits, nous souhaiterions évaluer *RefGen* sur un corpus plus large, afin de confirmer les résultats obtenus.

Aussi, dans un futur proche, nous souhaiterions apporter une série d'améliorations/extensions à *RefGen.* Ces dernières porteraient sur l'ajout de connaissances externes afin de réduire les erreurs d'identification des paires dans *CalcRef*, l'annotation de nouveau types de relations anaphoriques non prises en compte à l'heure actuelle par le système, l'utilisation de classes d'équivalence et de la dimension temporelle pour le suivi d'entité. Nous détaillons ces perspectives dans les sections suivantes.

1. Améliorations et extensions de *RefGen*

Nous proposons ci-dessous diverses améliorations de *RefGen* à court, moyen et long terme.

1.1 Utilisation de connaissances externes

A la suite de (Weissenbacher, 2008) et (Boudreau et Kittredge, 2005), nous pensons que certaines anaphores nominales peuvent difficilement être résolues sans l'utilisation de connaissances externes (listes de synonymes, ontologies). Pour ce faire, nous pourrions par exemple utiliser la base des « voisins » distributionnels de Wikipedia[1]. L'utilisation de telles ressources nous permettra d'identifier, par exemple, les liens de parenté établis entre des noms d'humains (père – fille), des synonymes (« *la modification … ce changement* »), des relations ontologiques du type « *Mr X est mort … Cette disparition …* ». L'utilisation de ces ressources nous permettra aussi d'éliminer des paires antécédents-anaphores non valides. Il sera alors possible de traiter les cas :

- d'hyperonymie, par exemple « *Le disjoncteur* évite les surintensités. *Cette protection* est destinée aux matériels et non pas aux personnes. »,
- d'hyponymie, par exemple « Pour être couvert en cas d'intrusion, vous devez installer *un système de sécurité*. *Cette alarme* devra être conforme aux normes en vigueur. ».

1.2 Traiter des cas d'anaphores plus complexes

Dans les futures extensions du module *RefGen*, nous prévoyons de considérer :

- *les anaphores plurielles* : à la lumière des travaux relatifs à la reprise des entités plurielles tels que (Schnedecker et Bianco, 1995 : 79-108), il nous faudra ajouter des contraintes dans notre calcul de la référence. En effet, la reprise plurielle, qu'elle soit pronominale ou nominale, peut regrouper deux éléments qui n'occupent pas la même position syntaxique. On trouvera par exemple : « *Pierre a connu Marie il y a 10 ans.* ***Ils*** *ont eu trois enfants.* ». Dans cet exemple, *Pierre* occupe la position *sujet* et on pourrait prévoir que cette entité fasse l'objet de reprises dans la suite. Or, le sujet et l'objet sont repris par un pronom pluriel. Cela sous-entend que nous devrons prendre en compte ce phénomène dans notre calcul et offrir la possibilité de regrouper deux antécédents possibles en un seul : « si la phrase n+1 contient un pronom pluriel, alors c'est une reprise potentielle d'un antécédent « dispersé » (Corblin, 1985) formé d'un sujet et d'un objet de type « nom de personne » de la phrase n ».

  Dans les cas plus classiques, où les deux entités qui font l'objet d'une

[1] http://redac.univ-tlse2.fr/voisinsdewikipedia/

reprise occupent la même position syntaxique, par exemple : « ***Pierre et Marie*** *se sont mariés en 1974.* ***Ils*** *ont eu 3 enfants. »,* il nous faudra identifier le référent « pluriel » issu de la coordination des deux entités de type « nom de personne » en position sujet « Pierre » et « Marie » comme premier maillon potentiel d'une chaîne de référence. Nous pourrons nous appuyer sur les travaux de (Schnedecker et Bianco, 2000) portant sur la préposition « avec ».

D'autres cas de reprise où le groupe nominal possède le sème pluriel, comme dans : « *Pierre et Marie…le couple »* pourront aussi bénéficier de ce même traitement, à la différence près qu'il nous faudra établir un liste des groupes nominaux représentant des entités plurielles, telles que « couple, bande, groupe, paire, … ».

Demeureront encore des cas d'anaphores à antécédent « flou » définis par (Landragin, 2007) pour lesquels le traitement automatique reste encore très incertain, par exemple, dans : « *Les manifestants… Pierre …… »* où l'on assiste à une extraction d'un élément issu d'un groupe. Ici, « Pierre » fait partie intégrante du groupe de manifestants, mais ce n'est pas le groupe entier qui est repris. Doit-on effectuer une double ouverture de chaîne de référence (une pour « les manifestants », l'autre pour « Pierre ») et rattacher la deuxième chaîne à la première (ce qui signifierait que l'on considèrerait des chaînes de référence à deux niveaux, la deuxième étant inclue dans la première) ? Devrait-on plutôt envisager une modification possible du contenu du référent de la chaîne de référence au fil du texte ?

– *les anaphores associatives* (relation méronymique ou partie-tout, fonctionnelle, locative, actancielle, (Kleiber, 2001 : 263-367)), comme « Les policiers inspectèrent la voiture. Les roues étaient pleines de boue. » ((Fradin, 1984) cité par (Kleiber et *al.*, 1994)). Dans *CalcRef*, lors du test des paires antécédent-anaphore possibles, le système doit valider la contrainte TETELEX consistant à vérifier que la tête lexicale des deux groupes nominaux de la paire est identique (voir chapitre 7 section 4.2.2). Dans le cas d'anaphores associatives, ce n'est pas le cas. Avec l'ajout de connaissances et d'une exception à cette contrainte (*e.g.* « si les deux groupes nominaux de la paire font partie du lexique de parenté par exemple, alors la règle TETELEX ne s'applique pas ») il sera alors possible de prendre en compte les anaphores associatives,
– les *anaphores génériques*, tel que « Ma *Clio* est encore en panne. *Ces citadines* ne sont plus aussi fiables qu'avant. ».

Pour caractériser ces cas, nous utiliserons la typologie des anaphores non pronominales définis par (Le Pesant, 2008) et les indices permettant de les identifier (distance, nature de la tête nominale, présence ou absence de modifieur, etc., *cf.* (Le Pesant, 2008 : 194-195)).

- D'autres cas, tels que les antonomases (le fait d'utiliser un nom commun pour désigner un nom propre, ou inversement ; comme dans « Au fond, Zucca, c'est *une sorte de Jacques Tourneur français* » (Leroy, 2001 : 192)), seront aussi à considérer dans le futur. Nous pourrons nous appuyer sur les scripts Perl développés par (Leroy, 2001) dans des corpus issus de deux genres textuels : les portraits journalistiques et les critiques de films (issus du quotidien *Libération*) pour identifier de manière automatique les cas d'antonomase du nom propre en français.

1.3 Utilisation de classes d'équivalence

Afin d'assurer le suivi des différentes dénominations d'une personne au cours du temps, il serait intéressant de mettre en place des classes d'équivalence. En effet, l'identification de tels phénomènes permettrait de résoudre l'ambiguïté entre les différentes facettes d'un même référent (Croft et Cruse, 2004 ; Hearst, 2006) dans un texte (*i.e.* les différentes fonctions politiques, métiers, statuts) et permettrait d'éviter la création automatique de plusieurs chaînes de référence (alors qu'il s'agit toujours de la même entité) en rattachant ces chaînes par une relation de coréférence d'« identité numérique » (Wiggins, 2001). Par exemple :

- Carla Bruni a été mannequin, chanteuse, première dame de France, ex-première dame de France et elle a aussi été surnommée « Carlita »,
- Arnold Schwarzenegger a été culturiste, acteur, gouverneur et il a reçu des surnoms différents suivant son statut : « Schwarzy » lorsqu'il est acteur et « Gouvernator » lorsqu'il est devenu gouverneur.

Des techniques de *clustering*, telles que (Santamaría *et al.*, 2010 ; Bernardini *et al.*, 2009) ont déjà tenté de désambiguïser ce type de phénomènes et pourraient constituer une piste intéressante à explorer.

1.4 Utilisation de la dimension temporelle

Aussi, dans un cadre de veille, le suivi d'entité permettrait d'identifier le changement de statut d'une personne, afin de mettre à jour une base de données. Par exemple, (Adam, 2007) relève que, suite à l'avant-veille du premier tour des élections présidentielles de 2007, soit le 20 avril, le statut de plusieurs

personnalités politiques a changé. C'est ainsi que F. Bayrou n'est plus désigné dans le corpus[2] par « le candidat UDF à la présidentielle », mais par « le candidat malheureux au premier tour ». De son côté, N. Sarkozy devient « le Président de la République » ce qui fait perdre ce statut à J. Chirac. Aussi, à partir du 26 mars, l'autrice constate que « ministre de l'Intérieur » cesse de désigner N. Sarkozy pour désigner F. Baroin. Ainsi, l'utilisation des informations temporelles présentes dans les documents permettrait de mettre à jour les statuts des référents au cours du temps afin d'éviter la présence d'informations obsolètes. Ces informations temporelles pourraient être identifiées *via* des règles symboliques intégrées au module *RefAnnot*.

2. Constitution d'un corpus de référence annoté en chaînes de référence

Comme nous l'avons vu aux chapitres 5 et 8, à la différence de nombreuses langues (anglais, allemand, espagnol, néerlandais), le français ne possède toujours pas de corpus de référence (d'un million de formes au moins) annoté en chaînes de référence[3]. Or, la constitution d'un tel corpus annoté, et son libre accès à la communauté scientifique, répondrait à un besoin réel. Ainsi, l'acquisition d'une telle ressource rendrait possible l'utilisation des techniques d'apprentissage automatique[4] pour l'entraînement d'un système d'apprentissage automatique, tels que celui de (Denis, 2007). De ce fait, nous pourrions (enfin) nous aligner sur les autres langues qui possèdent déjà une telle ressource et ainsi participer à des campagnes d'évaluation comme SemEval.

Aussi, afin de réduire sensiblement le temps d'annotation manuelle du corpus de référence, nous pourrions utiliser notre module *RefGen* comme outil de pré-annotation des expressions référentielles (avec *RefAnnot*) et des relations de coréférence (avec *CalcRef*). Dans la continuité du projet Peps *MC4*, un projet ANR nommé DEMOCRAT (*DEscription et MOdélisation des Chaînes de*

[2] Le corpus utilisé contient des articles de *Le Monde*, *Libération* et *Le Figaro* de 2006-2007.

[3] Un premier corpus oral de 418 000 mots annoté en coréférence (ANCOR, Muzerelle *et al.*, 2013), a été rendu disponible en novembre 2013 (http://tln.li.univ-tours.fr/Tln_Corpus_Ancor.html).

[4] Nous pourrions par exemple utiliser des Champs Markoviens Conditionnels (CRF, Conditional Random Fields). Ces modèles graphiques discriminants (Lafferty *et al.*, 2001 ; Sutton et McCallum, 2006) ont souvent obtenu les meilleures performances pour des tâches de TAL telles que la reconnaissance d'entités nommées (McCallum et Li, 2003), l'extraction d'information, l'étiquetage en parties du discours (Altun *et al.*, 2003). Les CRF permettent de combiner des approches symboliques et l'apprentissage statistique pour apprendre à annoter des données en se basant sur des exemples.

Référence : Outils pour l'Annotation de corpus et le Traitement automatique), piloté par Frédéric Landragin (CNRS, Lattice), en partenariat avec le LiLPa (Strasbourg) et ICAR (Lyon) a été déposé en ce sens et fournira, entre autres, un corpus de 300 000 mots annotés en relations de coréférence. Nous espérons que ce projet pourra être mené à bien.

3. Utilisation des chaînes de référence comme marqueurs de segmentation thématique

(Sitbon, 2004) propose, dans ses perspectives, de faire appel à des marqueurs linguistiques afin de déterminer des frontières candidates avant d'effectuer le calcul statistique des frontières thématiques. Suivant cette idée, il serait intéressant d'effectuer des tests psycholinguistiques afin de déterminer si les chaînes de référence constitueraient des indices forts de la segmentation thématique d'un document. En effet, comme nous avons pu le présenter dans les chapitres 2 et 3, des auteurs tels que (Charolles, 1995 ; Schnedecker, 1997, 2005) ont montré que certaines expressions référentielles présentes dans une chaîne de référence (*i.e.* les noms propres, les démonstratifs) marquent des ruptures thématiques dans la chaîne et, ce faisant, l'obligent à redémarrer. Par exemple :

> **La France**, en forme longue la République française, est une république constitutionnelle unitaire ayant un régime parlementaire à tendance présidentielle, dont la majeure partie du territoire et de la population est située en Europe occidentale, mais qui comprend également plusieurs régions et territoires répartis à travers le monde. Elle a pour capitale Paris, pour langue officielle le français et pour monnaie l'euro. Sa devise est « Liberté, Égalité, Fraternité », et son drapeau est constitué de trois bandes verticales respectivement bleue, blanche et rouge. Son hymne est La Marseillaise. Son principe est gouvernement du peuple, par le peuple et pour le peuple.
>
> **La France** est un pays ancien, formé au Haut Moyen Âge. Du début du XVIIe siècle à la première moitié du XXe siècle, elle possède un vaste empire colonial. À partir des années 1950, elle est l'un des acteurs de la construction de l'Union européenne. Elle est une puissance nucléaire, et l'un des cinq membres permanents du Conseil de sécurité des Nations unies. **La France** joue un rôle important dans l'histoire mondiale par l'influence de sa culture, de sa langue et de ses valeurs démocratiques, laïques et républicaines.

> **La France** occupe, en 2012, le cinquième rang mondial pour le produit intérieur brut. Son économie, de type capitaliste avec une intervention étatique assez forte, fait d'elle un des leaders mondiaux dans les secteurs de l'agroalimentaire, de l'aéronautique, de l'automobile, des produits de luxe, du tourisme et du nucléaire.
>
> Peuplée de 65,8 millions d'habitants au 1er janvier 2013, **la France** est un pays développé, avec un indice de développement humain très élevé. (http://fr.wikipedia.org/wiki/France).

Les chaînes de référence joueraient un rôle (dont l'importance est à quantifier) dans la structuration des textes, qui permettraient de proposer un découpage thématique original. De ce fait, les chaînes de référence pourraient être utilisées pour pré-segmenter thématiquement les documents. Cette technique permettrait d'obtenir des segments thématiques plus précis que ceux obtenus actuellement par les méthodes statistiques. Les discussions que nous avons pu entretenir à ce propos avec Y. Bestgen nous invitent vivement à aller en ce sens.

4. Extension des types de maillons des chaînes de référence

Nous avons pu montrer dans cette thèse que la composition des chaînes de référence étaient contraintes notamment par leur genre d'occurrence. L'étude de textes non-narratifs que nous poursuivons actuellement sur un corpus de textes juridiques a fait émerger la présence d'autres facteurs de cohésion tels que la répétition lexicale pour l'identification des référents non humains dans les chaînes de référence. Les chaînes de référence, initialement composées exclusivement d'expressions référentielles nominales (noms propres, groupes nominaux, etc.) pourrait en venir à évoluer pour inclure des verbes (conjugués ou à l'infinitif) permettant de référer à des entités abstraites dans le discours (actions, événements, faits). En effet, la prise en compte et l'identification automatique des coréférences événementielles (Davidson, 1967 ; Danlos, 2006 ; Bittar, 2006) permettrait d'inclure dans des chaînes de référence des textes non narratifs certains éléments, jusque-là ignorés, comme, par exemple, les mentions événementielles « retirer » ou « retirent » dans la chaîne {*retirer* le permis de séjour, ce retrait, *retirent* le permis de séjour, ce retrait, *retirent* le permis de séjour, ce retrait}, qui constituerait le thème central de l'extrait suivant :

> L'article 6, paragraphe 1, premier tiret, de la décision n° 1/80 […] lu en combinaison, notamment, avec le principe de sécurité juridique interdit-il aux autorités nationales compétentes de **retirer le permis de séjour** d'un travailleur turc, qui ne s'est rendu coupable d'aucun comportement

> frauduleux, avec effet rétroactif à la date à laquelle le motif auquel le droit national subordonnait l'octroi du permis de séjour a cessé d'exister, **ce retrait** intervenant après l'expiration du délai d'un an visé à l'article 6, paragraphe 1, premier tiret, susvisé ?»
>
> Par sa question, la juridiction de renvoi demande, en substance, si l'article 6, paragraphe 1, premier tiret, de la décision n° 1/80 doit être interprété en ce sens qu'il s'oppose à ce que les autorités nationales compétentes **retirent le permis de séjour** d'un travailleur turc avec effet rétroactif à la date à laquelle le motif auquel le droit national subordonnait l'octroi de son permis a cessé d'exister, lorsque ledit travailleur ne s'est rendu coupable d'aucun comportement frauduleux et que **ce retrait** a lieu après l'expiration de la période d'un an d'emploi régulier prévue audit article 6, paragraphe 1, premier tiret.
>
> Eu égard à ce qui précède, il y a lieu de répondre à la question posée que l'article 6, paragraphe 1, premier tiret, de la décision n° 1/80 doit être interprété en ce sens qu'il s'oppose à ce que les autorités nationales compétentes **retirent le permis de séjour** d'un travailleur turc avec effet rétroactif à la date à laquelle le motif auquel le droit national subordonnait l'octroi de son permis a cessé d'exister, lorsque ledit travailleur ne s'est rendu coupable d'aucun comportement frauduleux et que **ce retrait** a lieu après l'expiration de la période d'un an d'emploi régulier prévue audit article 6, paragraphe 1, premier tiret. (*corpus Unal*)

Toutes ces améliorations potentielles et ces questions en cours de réflexion illustrent l'étendue du travail qu'il reste encore à accomplir. Elles montrent la richesse de la référence textuelle qui offre encore de belles perspectives de recherche.

Annexes

1. Annexe 1 : exemples de corpus d'apprentissage

Nous présentons ci-dessous quelques exemples de phrases annotées issues de corpus d'apprentissage proposés lors des diverses campagnes d'annotations (ACE, SemEval, CoNLL). La plupart du temps, on retrouve les annotations suivantes :

- annotations morphosyntaxiques,
- *chunks*,
- propositions,
- entités nommées (*i.e.* lieux, organisations, personnes, etc.),
- relations de coréférence.

1.1 Campagne d'évaluation CoNLL 2011

```
0       The      DT (TOP_(S_(NP_* -        -       -       -       *
        *(ARG0****       (11

1       U.S.     NNP      *)       -        -       -       -       (GPE)*   *)
        *        *        *        11)

2       ,        ,        *        -        -       -       -       *
        *        *        *        *        *       -

3       claiming VBG      (S_(VP_* claim    01      2       -       *(V*)
        (ARGM-ADV*       *        *        *       -
```

1.2 Campagne MUC 6

```
[...]

<s>
        By/IN proposing/VBG
        <COREF ID="13" TYPE="IDENT" REF="6" MIN="date">
                a/DT meeting/NN date/NN
        </COREF>
        ,/,
        <COREF ID="14" TYPE="IDENT" REF="0">
                <ORGANIZATION>
                        Eastern/NNP
                </ORGANIZATION>
        </COREF>
```

```
        moved/VBD one/CD step/NN closer/JJR toward/IN reopening/VBG
       current/JJ high-cost/JJ contract/NN agreements/NNS with/IN
        <COREF ID="15" TYPE="IDENT" REF="8" MIN="unions">
                <COREF ID="16" TYPE="IDENT" REF="14">
                        its/PRP$
                </COREF>
                unions/NNS
        </COREF>
        ./.
</s>

[...]
```

1.3 Campagne ACE 2004

```
[...]

<entity ID="20001115_AFP_ARB.0212.eng-E1" TYPE="ORG" SUBTYPE="Educational"
CLASS="SPC">
        <entity_mention ID="1-47" TYPE="NAM" LDCTYPE="NAM">
                <extent>
                        <charseq START="475" END="506">
                                the Globalization Studies Center
                        </charseq>
                </extent>
                <head>
                        <charseq START="479" END="506">
                             Globalization Studies Center
                        </charseq>
                </head>
        </entity_mention>

 [...]
```

2. Annexe 2 : Jeu d'étiquettes utilisées par TTL pour le français[1]

ATTRIBUT	ETIQUETTE	DESCRIPTION	EXEMPLE
chunk	Np #n	groupe nominal	*le livre*
	Pp#n	groupe prépositionnel	*le livre de Luc*
	Ap#n	groupe adjectival	*le beau livre*
	Vp#n	groupe verbal	*mange*
ana	Af-ms	adjectif (masculin singulier)	*ancien*
	Ai-ms	adjectif indéfini	*aucun*
	Cc	conjonction de coordination	*et*
	Cs	conjonction de subordination	*que*
	Da-ms	déterminant (masculin singulier)	*le*
	Dd-fs	déterminant démonstratif féminin singulier	*cette*
	Dg-ms	déterminant contracté	*du*
	Mc	numéral	*1625*
	Pi	pronom indéfini	*chacun*
	Pp3	pronom personnel	*en*
	Pp3ms	pronom personnel 3ème personne du pluriel	*il*
	Pr	pronom relatif	*qui*
	Px	Pronom réfléchi	*se*
	Sp	préposition	*dans*
	Spa	préposition « à »	*à*
	Spd	préposition « de »	*de*
	R	adverbe	*toujours*
	Vasp3p	verbe au passé simple de l'indicatif 3ème personne du pluriel	*fussent*
	Vmip3s	verbe principal au présent de l'indicatif 3ème personne du singulier	*mange*
	Vmmp2s	Verbe principal à l'impératif 2ème personne du singulier	*mange*
	Vmn	verbe à l'infinitif	*manger*
	Vmps-s	participe passé	*mangé*
	Vmpp	participe présent	*mangeant*

[1] La liste exhaustive des étiquettes est disponible à : http://aune.lpl.univ-aix.fr/projects/multext/LEX/LEX.LangSpec.fr.html (consulté le 08/10/12)

3. Annexe 3 : Jeu d'étiquettes utilisées par TreeTagger pour le français

ETIQUETTE	DESCRIPTION
ABR	Abréviation
ADJ	Adjectif
ADV	Adverbe
DET : ART	Article
DET : POS	Pronom possessif (ma, ta, …)
INT	Interjection
KON	Conjonction
NAM	Nom propre
NOM	Nom commun
NUM	Numéral
PRO	Pronom
PRO : DEM	Pronom démonstratif
PRO : IND	Pronom indéfini
PRO : PER	Pronom personnel
PRO : POS	Pronom possessif (mien, tien, …)
PRO : REL	Pronom relatif
PRP	Préposition
PRP : det	Préposition + article (au, du, aux, des)
PUN	Ponctuation
PUN : cit	Ponctuation de citation
SENT	Balise de phrase
SYM	Symbole
VER:cond	Verbe au conditionnel
VER:futu	Verbe au futur
VER:impe	Verbe à l'impératif
VER:impf	Verbe à l'imparfait
VER:infi	Verbe à infinitif
VER:pper	Verbe au participe passé
VER:ppre	Verbe au participe présent
VER:pres	Verbe au présent
VER:simp	Verbe au passé simple
VER:subi	Verbe à l'imparfait du subjonctif
VER:subp	Verbe au présent du subjonctif

4. Annexe 4 : Erreurs récurrentes produites par *Flemm* (Namer, 2000)

Le tableau suivant présente les principales erreurs systématiques produites par l'étiqueteur *Flemm* :

NIVEAU	NATURE	ERREUR
Nom	genre	Information manquante
	nombre	Pas de nombre si le nom a la même forme pour le pluriel et le singulier
Adjectif	type	Pas de catégorie pour les adjectifs qualificatifs
	participe	Étiquetage comme verbe participe passé
Pronom	fonction syntaxique	Annotation erronée
Déterminant	genre	Manque pour les agrégats (*e.g.* « des »)
	catégorie	Information parfois indisponible
Verbe	ambiguïté	Plusieurs temps et modes
	catégorie	principal ou auxiliaire

5. Annexe 5 : Exemple de script Perl pour les corrections automatiques du corpus d'apprentissage de TTL

```
########## Script repair_det.pl ########################################
#                                                                      #
# récupère le genre du déterminant et l'ajoute au nom qui le suit      #
#                                                                      #
# - créé par : Amalia Todirascu                                        #
# - modifié par : Laurence Longo                                       #
#                                                                      #
########################################################################

open(K, 'fr1_acq_det.txt');
open(L, ">fr2_temp.txt");
open(H, 'fr2_temp.txt');
open(G, ">fr2_temp2.txt");
open(N, 'fr2_temp2.txt');
open(M, ">fr1_acq_det_out.txt");

while ($ligne2=<K>) {
      chomp($ligne2);
      @tab2 = split ( /\s+/, $ligne2);
      push ( @mot2, $tab2[0]);
      push ( @etiquette2, $tab2[1]);
      push ( @lemme2, $tab2[2] );
      }
$i=$#mot2+1;
for ($i =$#mot2 ; $i >=0 ; $i--) {
      print L "$mot2[$i] $etiquette2[$i] $lemme2[$i]\n";
      };
close K;
close L;

while ($ligne=<H>) {
      chomp($ligne);
      @tab = split (/\s+/, $ligne);
      push (@mot, $tab[0] );
      push (@etiquette, $tab[1]);
      push ( @lemme, $tab[2] );
      };
close H;

for ($i=0 ; $i <= $#etiquette ; $i++) {
      if (($etiquette[$i]=~/NOM\:Nc\-/) &&
      ($etiquette[$i+1]=~/DET.*\:D[ads][0123456789\-][fm].*/)){
      # modif pour les D[ads] pour traiter des cas comme "sa demande"
      $ind=index($etiquette[$i+1], ":D");
      $gen=substr($etiquette[$i+1], $ind+4, 1);
      $nb=substr($etiquette[$i+1], $ind+5, 1);
            if($etiquette[$i]=~/NOM\:Nc\-\-/){
                  $etiquette[$i]=~s/NOM\:Nc\-\-/NOM\:Nc$gen$nb/;
                  }
            else {
                  $etiquette[$i]=~s/NOM\:Nc\-/NOM\:Nc$gen/;
```

```
                  };
            print (G "$mot[$i] $etiquette[$i] $lemme[$i]\n");
      }
      else {
            if(($etiquette[$i]=~/NOM\:Nc\-/) &&
            ($etiquette[$i+1]=~/PRP.*\:Sp\+D[ads][0123456789\-][fm].*/)){
            $ind=index($etiquette[$i+1], ":Sp+D");
            $gen=substr($etiquette[$i+1], $ind+7, 1);
            $nb=substr($etiquette[$i+1], $ind+8, 1);
                  if($etiquette[$i]=~/NOM\:Nc\-\-/){
                        $etiquette[$i]=~s/NOM\:Nc\-\-/NOM\:Nc$gen$nb/;
                        }
                  else{
                        $etiquette[$i]=~s/NOM\:Nc\-/NOM\:Nc$gen/;
                        };
                  print (G "$mot[$i] $etiquette[$i] $lemme[$i]\n");
            }
            else{
                  print (G "$mot[$i] $etiquette[$i] $lemme[$i]\n");
                  }
      }
};
close G;

while ($ligne1=<N>) {
      ($ligne1);
      @tab1 = split (/\s+/, $ligne1);
      push ( @mot1, $tab1[0] );
      push ( @etiquette1, $tab1[1] );
      push ( @lemme1, $tab1[2] );
      }
$i=$#mot1+1;
for ($i =$#mot1 ; $i >=0 ; $i--) {
      print (M "$mot1[$i] $etiquette1[$i] $lemme1[$i]\n");
      };
close M;
close N;


######################################################################
```

6. Annexe 6 : Résultats de la campagne d'évaluation SemEval-2010 tâche 1

	Mention detection			CEAF			MUC			B^3			BLANC		
	R	P	F_1	R	P	F_1	R	P	F_1	R	P	F_1	R	P	Blanc
Catalan															
closed×gold															
RelaxCor	100	100	100	70.5	70.5	**70.5**	29.3	77.3	42.5	68.6	95.8	**79.9**	56.0	81.8	59.7
SUCRE	100	100	100	68.7	68.7	68.7	54.1	58.4	**56.2**	76.6	77.4	77.0	72.4	60.2	**63.6**
TANL-1	100	96.8	98.4	66.0	63.9	64.9	17.2	57.7	26.5	64.4	93.3	76.2	52.8	79.8	54.4
UBIU	75.1	96.3	84.4	46.6	59.6	52.3	8.8	17.1	11.7	47.8	76.3	58.8	51.6	57.9	52.2
closed×regular															
SUCRE	75.9	64.5	69.7	51.3	43.6	47.2	44.1	32.3	**37.3**	59.6	44.7	51.1	53.9	55.2	**54.2**
TANL-1	83.3	82.0	82.7	57.5	56.6	**57.1**	15.2	46.9	22.9	55.8	76.6	**64.6**	51.3	76.2	51.0
UBIU	51.4	70.9	59.6	33.2	45.7	38.4	6.5	12.6	8.6	32.4	55.7	40.9	50.2	53.7	47.8
open×gold															
open×regular															
Dutch															
closed×gold															
SUCRE	100	100	100	58.8	58.8	**58.8**	65.7	74.4	**69.8**	65.0	69.2	**67.0**	69.5	62.9	**65.3**
closed×regular															
SUCRE	78.0	29.0	42.3	29.4	10.9	15.9	62.0	19.5	**29.7**	59.1	6.5	11.7	46.9	46.9	**46.9**
UBIU	41.5	29.9	34.7	20.5	14.6	**17.0**	6.7	11.0	8.3	13.3	23.4	**17.0**	50.0	52.4	32.3
open×gold															
open×regular															
English															
closed×gold															
RelaxCor	100	100	100	75.6	75.6	**75.6**	21.9	72.4	33.7	74.8	97.0	**84.5**	57.0	83.4	61.3
SUCRE	100	100	100	74.3	74.3	74.3	68.1	54.9	**60.8**	86.7	78.5	82.4	77.3	67.0	**70.8**
TANL-1	99.8	81.7	89.8	75.0	61.4	67.6	23.7	24.4	24.0	74.6	72.1	73.4	51.8	68.8	52.1
UBIU	92.5	99.5	95.9	63.4	68.2	65.7	17.2	25.5	20.5	67.8	83.5	74.8	52.6	60.8	54.0
closed×regular															
SUCRE	78.4	83.0	80.7	61.0	64.5	**62.7**	57.7	48.1	**52.5**	68.3	65.9	**67.1**	58.9	65.7	**61.2**
TANL-1	79.6	68.9	73.9	61.7	53.4	57.3	23.8	25.5	24.6	62.1	60.5	61.3	50.9	68.0	49.3
UBIU	66.7	83.6	74.2	48.2	60.4	53.6	11.6	18.4	14.2	50.9	69.2	58.7	50.9	56.3	51.0
open×gold															
Corry-B	100	100	100	77.5	77.5	77.5	56.1	57.5	56.8	82.6	85.7	84.1	69.3	75.3	**71.8**
Corry-C	100	100	100	77.7	77.7	**77.7**	57.4	58.3	57.9	83.1	84.7	83.9	71.3	71.6	71.5
Corry-M	100	100	100	73.8	73.8	73.8	62.5	56.2	**59.2**	85.5	78.6	81.9	76.2	58.8	62.7
RelaxCor	100	100	100	75.8	75.8	75.8	22.6	70.5	34.2	75.2	96.7	**84.6**	58.0	83.8	62.7
open×regular															
BART	76.1	69.8	72.8	70.1	64.3	67.1	62.8	52.4	57.1	74.9	67.7	71.1	55.3	73.2	57.7
Corry-B	79.8	76.4	78.1	70.4	67.4	68.9	55.0	54.2	54.6	73.7	74.1	**73.9**	57.1	75.7	**60.6**
Corry-C	79.8	76.4	78.1	70.9	67.9	**69.4**	54.7	55.5	55.1	73.8	73.1	73.5	57.4	63.8	59.4
Corry-M	79.8	76.4	78.1	66.3	63.5	64.8	61.5	53.4	**57.2**	76.8	66.5	71.3	58.5	56.2	57.1
German															
closed×gold															
SUCRE	100	100	100	72.9	72.9	72.9	74.4	48.1	**58.4**	90.4	73.6	81.1	78.2	61.8	**66.4**
TANL-1	100	100	100	77.7	77.7	**77.7**	16.4	60.6	25.9	77.2	96.7	**85.9**	54.4	75.1	57.4
UBIU	92.6	95.5	94.0	67.4	68.9	68.2	22.1	21.7	21.9	73.7	77.9	75.7	60.0	77.2	64.5
closed×regular															
SUCRE	79.3	77.5	78.4	60.6	59.2	**59.9**	49.3	35.0	**40.9**	69.1	60.1	**64.3**	52.7	59.3	**53.6**
TANL-1	60.9	57.7	59.2	50.9	48.2	49.5	10.2	31.5	15.4	47.2	54.9	50.7	50.2	63.0	44.7
UBIU	50.6	66.8	57.6	39.4	51.9	44.8	9.5	11.4	10.4	41.2	53.7	46.6	50.2	54.4	48.0
open×gold															
BART	94.3	93.7	94.0	67.1	66.7	**66.9**	70.5	40.1	**51.1**	85.3	64.4	**73.4**	65.5	61.0	**62.8**
open×regular															
BART	82.5	82.3	82.4	61.4	61.2	**61.3**	61.4	36.1	**45.5**	75.3	58.3	**65.7**	55.9	60.3	**57.3**
Italian															
closed×gold															
SUCRE	98.4	98.4	98.4	66.0	66.0	**66.0**	48.1	42.3	**45.0**	76.7	76.9	**76.8**	54.8	63.5	**56.9**
closed×regular															
SUCRE	84.6	98.1	90.8	57.1	66.2	**61.3**	50.1	50.7	**50.4**	63.6	79.2	**70.6**	55.2	68.3	**57.7**
UBIU	46.8	35.9	40.6	37.9	29.0	32.9	2.9	4.6	3.6	38.4	31.9	34.8	50.0	46.6	37.2
open×gold															
open×regular															
BART	42.8	80.7	55.9	35.0	66.1	45.8	35.3	54.0	**42.7**	34.6	70.6	46.4	57.1	68.1	**59.6**
TANL-1	90.5	73.8	81.3	62.2	50.7	**55.9**	37.2	28.3	32.1	66.8	56.5	**61.2**	50.7	69.3	48.5
Spanish															
closed×gold															
RelaxCor	100	100	100	66.6	66.6	66.6	14.8	73.8	24.7	65.3	97.5	**78.2**	53.4	81.8	55.6
SUCRE	100	100	100	69.8	69.8	**69.8**	52.7	58.3	**55.3**	75.8	79.0	77.4	67.3	62.5	**64.5**
TANL-1	100	96.8	98.4	66.9	64.7	65.8	16.6	56.5	25.7	65.2	93.4	76.8	52.5	79.0	54.1
UBIU	73.8	96.4	83.6	45.7	59.6	51.7	9.6	18.8	12.7	46.8	77.1	58.3	52.9	63.9	54.3
closed×regular															
SUCRE	74.9	66.3	70.3	56.3	49.9	52.9	35.8	36.8	**36.3**	56.6	54.6	55.6	52.1	61.2	**51.4**
TANL-1	82.2	84.1	83.1	58.6	60.0	**59.3**	14.0	48.4	21.7	56.6	79.0	**66.0**	51.4	74.7	**51.4**
UBIU	51.1	72.7	60.0	33.6	47.6	39.4	7.6	14.4	10.0	32.8	57.1	41.6	50.4	54.6	48.4
open×gold															
open×regular															

Tableau 34 - Résultats officiels de SemEval-2010 (Recasens *et al.*, 2010 : 7)

7. Annexe 7 : Le Nouveau chapitre de la thèse (NCT)[2]

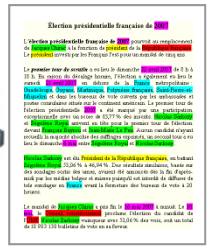
ABG
L'intelli'agence

Valorisation des compétences des docteurs, NCT®

Laurence LONGO

Ecole doctorale des Humanités
Université de Strasbourg
Nom du "mentor" : Bernard JUND

Un outil de détection automatique de thèmes pour améliorer les résultats des moteurs de recherche


Election présidentielle française de 2007

L'élection présidentielle française de 2007 pourvoit au remplacement de Jacques Chirac à la fonction de président de la République française. Le président sortant est Jacques Chirac, élu en 1995 puis réélu en 2002 pour un mandat de cinq ans.

Le premier tour du scrutin a eu lieu le dimanche 22 avril 2007 de 8 h à 18 h. En raison du décalage horaire, l'élection a également eu lieu le samedi 21 avril 2007 en métropole, à Mayotte, à Saint-Pierre-et-Miquelon et dans les bureaux de vote ouverts par les ambassades et postes consulaires situés sur le continent américain. Le premier tour de l'élection présidentielle 2007 a été marqué par une participation exceptionnelle avec un score de 83,77 % des inscrits. Ségolène Royal et Nicolas Sarkozy arrivent en tête pour le premier tour de l'élection, devant François Bayrou et Jean-Marie Le Pen. Aucun candidat n'ayant recueilli la majorité absolue des suffrages exprimés, un second tour a eu lieu le dimanche 6 mai entre Nicolas Sarkozy et Ségolène Royal.

Nicolas Sarkozy est élu Président de la République française, en battant Ségolène Royal 53,06 % à 46,94 %. Des résultats similaires, basés sur des sondages sortie des urnes, avaient été annoncés dès la fin d'après-midi par les médias belges et suisses pourvu est interdit de diffuser de tels sondages en France avant la fermeture des bureaux de vote à 20 heures.

Le mandat de Jacques Chirac a pris fin le 16 mai 2007 à minuit. Le 16 mai, Nicolas Sarkozy prêtait serment. Le Conseil constitutionnel proclame l'élection du candidat de l'UMP, Nicolas Sarkozy, vainqueur avec 53,06 % des voix, soit un total de 18 983 138 bulletins de vote en sa faveur.

Date de présentation orale du « NCT » : 14 juin 2011

[2] Cette formation doctorale s'est déroulée lors de la 4ème année de thèse.

1. Cadre général et enjeux de la thèse

1.1 Présentation générale du projet

Pour faire face à la constante augmentation du nombre de documents disponibles sur Internet, la plupart des systèmes de recherche d'information font appel aux techniques de Traitement Automatique des Langues (TAL) qui exploitent les informations syntaxiques ou sémantiques, dans le but d'améliorer la qualité des résultats fournis par les moteurs de recherche.

Force est de constater que, parmi la multitude des résultats renvoyés à l'issue d'une recherche, rares sont ceux qui comportent les informations attendues. En parallèle à cela, certains documents pertinents ne sont malheureusement pas retrouvés par les moteurs de recherche. Ce manque de pertinence est dû à la méthode d'indexation par mots-clés utilisée par les moteurs de recherche, qui ne tient pas compte des propriétés linguistiques des textes (syntaxe, contenu, genre textuel, etc.).

Ainsi, notre objectif est de proposer une méthode innovante d'indexation de documents par thèmes. Ces thèmes sont identifiés automatiquement dans chaque document. Dans notre approche, les thèmes textuels constituent les sujets d'un texte, ou d'un fragment de document. Pour identifier les thèmes, nous utilisons les propriétés globales du texte : la cohésion et la cohérence, mais aussi les propriétés spécifiques liés au genre textuel.

L'outil de détection de thèmes mis en place permet d'effectuer, en parallèle à la recherche plein texte classique des moteurs de recherche, une recherche par thèmes. Nous appliquons une méthode hybride qui combine techniques statistiques et marqueurs linguistiques pour identifier les thèmes des documents. Parmi ces marqueurs linguistiques, les chaînes de référence font l'objet d'une attention particulière. Les chaînes de référence sont des marqueurs linguistiques permettant d'identifier des ruptures ou des continuations thématiques dans le discours. Construites au fil du texte par la présence de diverses expressions référentielles (noms propres, pronoms, groupes nominaux) se rapportant à la même entité, les chaînes de référence sont des indices forts pour la détection des thèmes. Par exemple, la présence des expressions « Barack Obama …il…il…ce président » dans une même portion de texte va constituer une chaîne de référence permettant d'indiquer que « Barack Obama » constitue le personnage central du texte. Ainsi, « Barack Obama » est alors le thème dans cette portion du discours. L'identification automatique des chaînes de référence s'effectue suivant des paramètres dépendants du genre textuel, définis à l'issue d'une étude en corpus (recueil de textes) menée sur cinq genres différents (analyses politiques, éditoriaux, roman, rapports publics, normes européennes).

Dans sa version actuelle, l'outil de détection automatique de thèmes, développé en Java, récupère en entrée un texte brut. Le texte est ensuite étiqueté (rajout d'informations sur la catégorie grammaticale, le genre, le nombre, …) et annoté au niveau des entités nommées (identification et catégorisation des noms propres de personne, d'organisation et de fonction). Le texte ainsi enrichi en annotations passe dans le module d'identification des chaînes de référence. Les chaînes de référence obtenues sont ensuite comparées suivant des critères de cohésion lexicale (identité des chaînes lexicales) afin de proposer une liste de thèmes décrivant les documents.

1.2 La thèse dans son contexte

Ma thèse s'est déroulée dans le cadre d'une convention CIFRE (Convention Industrielle de Formation par la Recherche) de trois ans entre le laboratoire LiLPa (Linguistique, Langues et Parole) de l'Université de Strasbourg et la société RBS (Ready Business System) basée à

Entzheim. Cette collaboration a permis d'offrir des compétences complémentaires : théoriques et méthodologiques pour le LiLPa et plutôt techniques (intégration et mise en œuvre des travaux dans un milieu industriel) pour RBS.

Dans ce contexte, j'ai été amenée à partager mon temps de travail entre mon laboratoire de recherche et la section Recherche et Développement de l'entreprise.

L'équipe LiLPa (Linguistique, Langues et Parole) est une équipe d'accueil de l'Université de Strasbourg qui regroupe des chercheurs en linguistique, socio-linguistique, didactique des langues, linguistique de corpus, phonétique. Etant co-dirigée par Mme Todirascu (Maître de conférence en informatique) et Mr Kleiber (professeur en linguistique et sémanticien de renommée internationale), j'appartiens aux composantes Fonctionnements Discursifs pour la partie Traitement Automatique des Langues et Scolia (Sciences Cognitives, Linguistique et intelligence artificielle) pour la partie Linguistique. La composante Fonctionnements Discursifs effectue des recherches en linguistique et didactique des langues vivantes étrangères. L'axe de recherche « Lexicologie et TAL » crée, entre autres, des ressources linguistiques en format électronique (lexiques, grammaires, corpus annotés), ainsi que des outils TAL pour l'acquisition des connaissances. La composante SCOLIA a pour objet central l'étude du sens sous toutes ses facettes. La méthode adoptée est avant tout celle des indices ou manifestations linguistiques présentes en corpus.

Prestataire de service en ingénierie informatique, RBS compte plus de 160 collaborateurs. Cette entreprise est un intégrateur de logiciels et d'équipements mais elle développe aussi ses propres logiciels, dans les domaines de la gestion de contenus, des applications collaboratives et de la mobilité. L'entreprise est soutenue par Oséo, elle a obtenu le label Entreprise Innovante et bénéficie du statut de « PME de croissance » (entreprise Gazelle). RBS se déploie à présent à l'international, notamment dans les pays de l'Est et du Golfe persique.

1.3 Réseau scientifique dans le cadre de la thèse

Les diverses composantes de mon laboratoire sont impliquées dans de nombreux projets nationaux, européens et internationaux. Plus spécifiquement, l'axe « lexicologie et TAL » travaille en étroite collaboration avec l'Académie Roumaine, l'Université de Stuttgart (projet sur les genres) et est membre de l'infrastructure de recherche européenne CLARIN (projet sur la relation lexicale hiérarchique CAP).

Depuis cette année, le laboratoire LilPa est membre de l'ILF (Institut de Linguistique Française). Cet institut a été créé pour assurer la visibilité internationale de la recherche en linguistique du français et faciliter les projets multi-laboratoires.

Dans le cadre de ma thèse, j'ai pu participer au groupe de travail sur la coréférence (les modes de reprise lexicale dans le texte) au laboratoire LATTICE à Paris. Les réunions mensuelles ont été l'occasion de rencontrer de nombreux spécialistes en linguistique et en traitement automatique des langues. Les divers échanges m'ont permis de faire avancer mes réflexions sur la référence en général (le choix des marqueurs pertinents pour détecter les thèmes), sur les problèmes d'annotation des marqueurs linguistiques en particulier (délimitation des entités nommées notamment). Les questions soulevées lors de ces réunions rentraient totalement en adéquation avec mes propres interrogations et m'ont aidé à construire peu à peu mon projet d'après thèse.

1.4 Ma place dans ce contexte

Depuis mon entrée en Master Sciences du Langage mention Technologies du Langage effectué à Aix en Provence, j'ai pu apprécier le milieu de la recherche. J'ai d'abord travaillé lors de mon stage de Master 2 Recherche sur du discours oral spontané où j'ai confronté les spécificités de l'oral (telles que les reformulations, les répétitions, les chevauchements de parole) à des analyseurs syntaxiques prévus pour traiter du texte écrit. J'ai voulu poursuivre cette étude de l'oral en me focalisant sur les chevauchements de parole dans le cadre d'un projet de thèse, mais, n'ayant pas obtenu de financement, le projet a été abandonné. J'ai ensuite effectué un master 2 professionnel qui m'a permis d'intégrer le centre de recherche Xerox à Grenoble. Pendant 6 mois, j'ai participé à une campagne d'évaluation d'analyseurs syntaxiques (SemEval), ce qui m'a permis de passer de l'étude de l'oral à l'étude de l'écrit et plus spécifiquement à la syntaxe. L'objectif de ce projet était d'évaluer l'analyseur syntaxique de Xerox (XIP), face à une tache d'identification des emplois métonymiques des entités nommées (déterminer si derrière l'emploi de l'entité nommée « France » dans « la France a signé le traité » par exemple, se cache une institution, un groupe d'humains…). Ce deuxième stage a conforté mon désir de poursuivre en thèse, avec pour objectif, cette fois de me tourner vers la sémantique et pourquoi pas le milieu industriel. C'est pourquoi, j'ai candidaté à 3 appels d'offre diffusés sur une liste spécialisée (liste Ln). J'ai été sélectionnée parmi 20 candidats pour une thèse au Listic de Chambéry (laboratoire d'Informatique Système, Traitement de l'Information et de la Connaissance) sur l'acquisition des connaissances à partir de textes et parmi 10 candidats pour une thèse au LiLPa à Strasbourg, sur l'amélioration des moteurs de recherche. Le projet de thèse sur les moteurs de recherche correspondait totalement à mes attentes. Les grandes lignes du projet avaient déjà été définies entre le laboratoire et l'entreprise pour la demande de convention CIFRE. Je suis plutôt intervenue au cours du projet, notamment en encadrant trois stages (deux en entreprise et un dans le laboratoire) pour lesquels le sujet se rapprochait ou faisait partie intégrante de mon projet de recherche.

2. Déroulement, gestion et coût du projet de recherche

2.1 Préparation et cadrage du projet de recherche

Pour mener à bien ce projet et de manière à répondre aux objectifs académiques et industriels, le projet a été rigoureusement délimité. Côté académique, la détection automatique de thèmes n'avait pas fait l'objet de travaux antérieurs dans le laboratoire. Néanmoins, des études linguistiques fines sur les anaphores et les chaînes de référence (un des marqueurs utilisés pour détecter les thèmes) avaient été effectuées par deux professeurs du laboratoire, ou plus largement sur les relations de (co)référence, ce qui nous a permis d'avoir des appuis théoriques solides. De plus, ma co-directrice de thèse avait déjà développé des applications de TAL, notamment pour l'étude des collocations (une association fréquente de deux ou plusieurs mots se suivant dans une phrase, comme : « entraîner des conséquences »).

Côté industriel, l'entreprise existait depuis 10 ans et était en croissance constante. L'entreprise était à la recherche d'une nouvelle solution pour améliorer son moteur de recherche interne. L'entreprise mettait à disposition du matériel et était assez souple pour le déroulement du projet, car, composée essentiellement d'informaticiens, elle ne détenait pas la connaissance linguistique nécessaire.

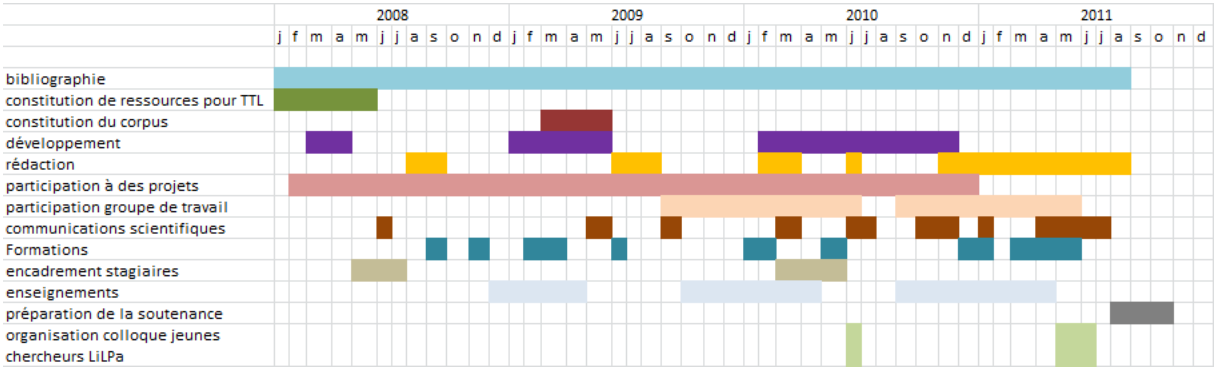
Parmi les facteurs de risque potentiels dans le cadre d'une CIFRE, le premier aurait été l'abandon du projet, soit pour des raisons économiques, soit parce que le projet n'aurait pas pu être mené à terme dans les temps impartis. Néanmoins, sur ce point, les objectifs et limites étaient bien définis dès le départ et nous nous sommes efforcés de suivre à la lettre le calendrier. Un autre

facteur important aurait été la difficulté de dialogue entre les linguistes et les informaticiens, concernant le transfert du besoin de l'entreprise vers l'Université.

Une partie du projet a été effectuée en collaboration avec le laboratoire RACAI de l'Académie Roumaine à Bucarest qui avait déjà collaboré dans des projets avec ma co-directrice de thèse. Nous avons participé à la mise en place des ressources linguistiques d'apprentissage nécessaires pour l'adaptation au français d'un étiqueteur (outil de traitement automatique des langues permettant de préciser des informations telles que la catégorie grammaticale de chacun des mots d'un corpus, la forme infinitive des verbes...) conçu pour l'anglais et le roumain. Un facteur de risque potentiel dans cette collaboration aurait été que le laboratoire roumain se rétracte ou qu'il nous fournisse un outil non exploitable.

2.2 Conduite du projet de recherche

Les grandes étapes de la conduite de mon projet de recherche peuvent être représentées par le diagramme suivant :



Ainsi, la première étape de la thèse a été de rechercher les articles et ouvrages traitant des thèmes, afin de délimiter les problèmes posés par cette notion (flou terminologique) et de donner la définition de cette notion dans notre approche. Ce travail a été repris plusieurs fois au cours de la thèse et élargi aux domaines du traitement automatique des langues (pour les entités nommées et le calcul de la référence), de la psychologie (pour la notion de saillance référentielle). En parallèle à cette première recherche bibliographique, nous avons passé 4 mois à constituer les ressources linguistiques nécessaires à l'adaptation au français de l'étiqueteur TTL. Ce travail a consisté à étiqueter un corpus de grande taille (1 million de termes) avec divers outils puis de corriger les erreurs d'annotations automatiques (soit avec des règles automatiques créées en Perl pour les erreurs récurrentes, soit manuellement). En même temps a débuté un projet de recherche mené par ma co-directrice sur la relation hiérarchique « CAP ». Ce projet proposait d'étudier les relations hiérarchiques entre les noms propres (de personne et de fonction) dans plusieurs langues (français, anglais, allemand, roumain). Ce projet s'intégrait totalement dans mon projet de recherche car, pour détecter les thèmes dans les documents, j'ai étudié les entités nommées (noms de personne, de lieu, d'organisation, de fonction) et défini des patrons de formation pour les identifier de manière automatique.

Dès la fin de ma première année (décembre 2008), j'ai eu la possibilité d'être vacataire d'enseignement en informatique au département LSHA (Lettres et Sciences Humaines Appliquées). J'ai pu enseigner des cours de bureautique (Microsoft Office Word, Access, Excel), de création de sites Web ainsi qu'un cours sur les ressources électroniques, auprès d'un public varié d'étudiants de licence en Lettres, Sciences du Langage, LEA, Langues Vivantes. En plus de

ces enseignements (204 heures de travaux dirigés sur 3 ans), j'ai pu encadrer deux stagiaires en master linguistique et informatique, qui ont participé à mon projet de thèse.

Une phase importante qui a occupé une bonne partie de la deuxième et troisième année a concerné le développement des différentes briques de notre système : écriture des règles d'identification des marqueurs, mise en place de l'algorithme pour identifier les chaînes de référence, ainsi que leur évaluation.

Une autre phase très formatrice a consisté à soumettre des communications à diverses conférences nationales et internationales, en axant et en adaptant notre propos suivant les thématiques des conférences (étude de corpus, développement d'outil, …). Les derniers mois de ma thèse sont essentiellement voués à la rédaction de mon mémoire de thèse.

Mon projet de recherche s'est toujours déroulé en étroite collaboration avec ma co-directrice de thèse. Nous avons eu des réunions régulières, ce qui m'a permis, au besoin, de réorienter mes recherches rapidement. Des réunions de travail ont été organisées, plus ponctuellement, avec mon directeur de thèse, pour obtenir son avis éclairé sur nos choix méthodologiques. Côté entreprise, des réunions mensuelles avec mon responsable technique ont permis un suivi régulier de l'état d'avancement des travaux et des questions en cours. Une présentation annuelle de l'avancée des travaux de thèse a eu lieu en présence du Président Directeur Général de l'entreprise ainsi que des responsables des pôles Recherche et Développement, Intranet et Commercial ; pour évaluer les retombées de ce projet et estimer sa future mise en commercialisation. J'ai également échangé avec les ingénieurs de la section Recherche et Développement, notamment lors de l'implémentation du système de détection automatique des thèmes. Ces échanges m'ont permis d'adapter mon discours face à des non linguistes et à évaluer la faisabilité de mon projet et sa portabilité vers le langage Java.

Nous avons éprouvé des difficultés, durant la première année, avec l'étiqueteur car le laboratoire roumain, impliqué dans de nombreux autres projets, a tardé à nous livrer une version stable de l'étiqueteur, qui est le premier élément de notre chaîne de traitements linguistiques. Nous avons quand même travaillé avec une version bêta de cet étiqueteur et nous avons pu l'évaluer et lister d'autres erreurs récurrentes rencontrées. Durant la seconde année, c'était l'accès en ligne (*via* un service web) à cet étiqueteur qui nous a fait défaut, car nous étions limités sur la taille des corpus. Nous avons alors mis en place des scripts en langage Perl, pour découper nos corpus en plus petits bouts, les envoyer les uns après les autres à l'outil en ligne, puis regrouper tous ces fragments étiquetés en un seul gros fichier.

Une autre difficulté à laquelle nous avons dû faire face était l'apprentissage du langage Java pour le développement de notre système. En effet, au cours de mon cursus en traitement automatique des langues, j'ai appris le langage Perl et le C++. Or, l'entreprise souhaitait absolument que le système soit développé en Java. J'ai suivi quelques cours de Master 2 enseignés par ma co-directrice de thèse et j'ai ensuite effectué de l'autoformation pour me mettre au niveau rapidement.

Pour valoriser mes travaux de recherche, j'ai eu l'occasion de présenter ma thèse lors de plusieurs conférences nationales et internationales ; ainsi que lors de séminaires et réunions (dans mon laboratoire de recherche ou dans le groupe de travail sur la coréférence au laboratoire LATTICE à Paris).

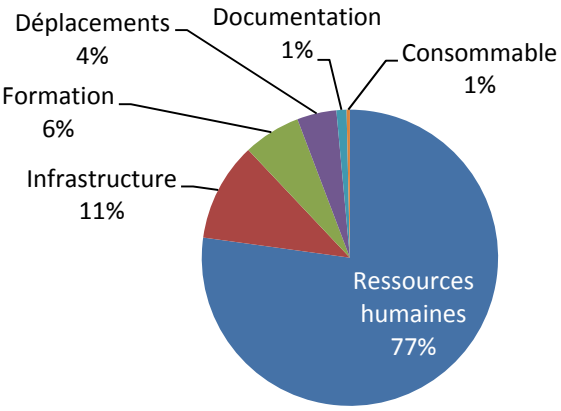
2.3 Evaluation et prise en charge du coût du projet

2.3.1 Répartition du coût du projet

Mon projet s'est déroulé sur 4 ans. Il est estimé à 167 000€ (voir le détail page suivante). Le principal poste de dépenses concerne des ressources humaines (77%). Outre le salaire de mes co-directeurs, de mon directeur technique et du mien, nous avons bénéficié d'un ingénieur développement (de RBS) pendant 25 jours ainsi que de deux stagiaires (4 mois chacun).

Viennent ensuite les frais d'infrastructure (11%) et de formation (14 formations en 3 ans, dont les Doctoriales en 2008, une formation sur les réseaux sociaux, sur la propriété industrielle et une école d'été sur les méthodes et outils d'exploration de corpus).

Aucun investissement lourd n'a été engagé pour ce projet, les moyens techniques utilisés étaient déjà disponibles dans le laboratoire ou dans l'entreprise (et amortis).



Répartition du coût du projet

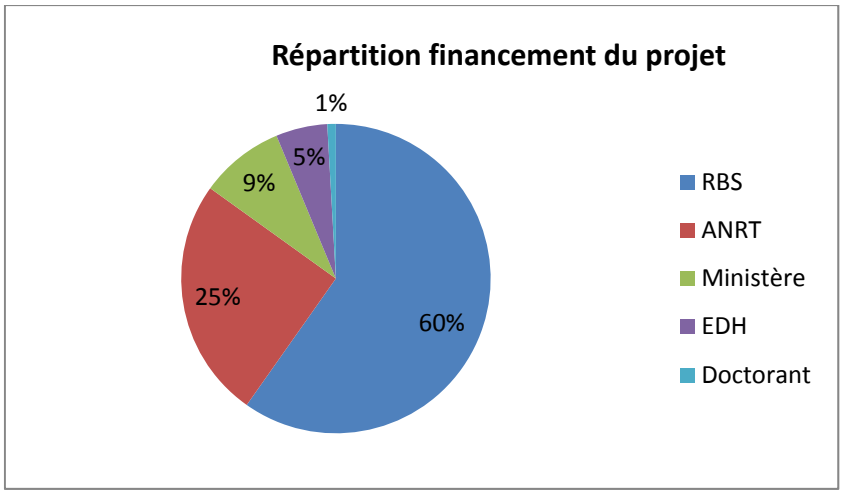
Montants en euros TTC

	Nature de la dépense	Détails		Coûts totaux (euros TTC)				Crédit
				Nombre d'unités	Coût unitaire moyen	Quote-part utilisation	Total	
1	**Ressources Humaines**							
1.1	Doctorant	salaire brut : 2200	charges patronales : 500	36	97200	100%	97200	ANRT + RBS
1.2	Encadrant 1	salaire brut : 6000	Charges : 1800	36	280800	1%	2808	Ministère
1.3	Prime Encadrement	637		4	2548	100%	2548	Ministère
1.4	Encadrant 2	salaire brut : 4000	Charges : 1200	36	187200	5%	9360	Ministère
1.5	Prime Encadrement							
1.6	Autre personnel (hors sous-traitance)	2 stagiaires salaire brut : 500	Charges : 200	8	700	30%	1680	RBS
1.7	Sous-traitance	1 ingénieur : 2600	charges : 700	4	3100	10%	12400	RBS
1.7	Sous-traitance	1 responsable produit : 3200	charges : 900	36	4100	2%	2952	RBS
	Sous-total Ressources Humaines						128948	
2	**Consommables**							
2.1	Fournitures de bureau			3	200	100%	600	RBS
	Sous-total Consommables						600	
3	**Infrastructures**							
	Sous-total Infrastructures			36	500	100%	18000	RBS
5	**Déplacements**							
5.1	Missions en France	Transport	Hébergement + autres frais				3000	RBS
5.3	Congrès en France	Transport	Hébergement + autres frais				1200	RBS
5.4	Congrès à l'étranger	Transport	Hébergement + autres frais				3000	RBS
	Sous-total Déplacements						7200	
6	**Formation**							
6.1	Formations			15	600		9000	EDH
6.2	Autres frais (Inscription à l'Université, SS)	inscription		4	370		1480	doctorant
	Sous-total Formation						10480	
7	**Documentation et communication**							
7.1	Affranchissements, Internet, téléphone			36	30		1080	RBS
7.2	Publicité, communication, impressions	Direct : 2 posters	Sous-traitance agence	2	30		60	RBS
7.3	Documentation (périodiques, livres, bases de données, bibliothèque, etc.)	abonnement revues, livres		2	350		700	RBS
	Sous-total Documentation et communication						1840	
10	**TOTAL**						**167068**	

2.3.2 Répartition du financement

Concernant la répartition du financement, dans le cadre de la convention CIFRE, l'ANRT a versé une subvention totale d'environ 42 000€ (sur 3 ans) à l'entreprise.

En plus du versement de mon salaire, l'entreprise RBS a versé au laboratoire de recherche académique une subvention de 18 000€ pour 3 ans (pour couvrir les frais de déplacement à des conférences, l'achat de livres, etc.). RBS a aussi rémunéré les deux stagiaires et l'ingénieur développement qui a construit la coque du module d'annotations de notre système.



3 Compétences mises en œuvre durant la thèse

3.1 Compétences scientifiques et techniques

Mon projet de recherche étant situé à la croisée de la linguistique et de l'informatique, j'ai donc fait appel à des méthodes issues de la linguistique de corpus, de la sémantique lexicale, du traitement automatique des langues (TAL) et de l'algorithmique. Pour développer l'outil de détection de thèmes, j'ai utilisé le langage Perl mais aussi le Java ainsi que divers outils de TAL tels que les concordanciers, les étiqueteurs, et des outils de statistiques textuelles. Pour la rédaction scientifique ainsi que les réunions j'ai utilisé des outils de travail collaboratif (agendas partagés, gestionnaires de versions). Enfin, pour éditer certains articles scientifiques, j'ai appris (en autoformation) à utiliser l'outil LaTeX, encore peu utilisé dans le domaine des Sciences Humaines.

Avant tout développement (étude de corpus, création de règles de formation, programmation) une recherche bibliographique doit être réalisée. Celle-ci nous permet d'acquérir des connaissances générales du domaine (p.e. résolution des anaphores) ou plus approfondies sur des aspects particuliers (p.e. étude de marqueurs linguistiques de cohésion). Par la suite, je me suis tenue à jour des différentes avancées dans mon domaine, à la fois par la mise en place d'un système de **veille scientifique** et par la participation à des conférences nationales et internationales.

D'un point de vue méthodologique, la **rigueur** a été une des qualités majeures dont je me suis efforcée de faire preuve tout au long de la thèse. J'ai toujours été très attentive aux remarques et conseils de mes encadrants, ce qui m'a permis de progresser dans l'écriture des articles scientifiques par exemple. J'ai aussi tenu les délais pour la soumission d'articles dans des conférences, les états d'avancement de ma thèse ou les présentations mensuelles dans l'entreprise. Dans le même ordre d'idée, j'ai développé une capacité à **mener de front plusieurs**

travaux (participation à des projets universitaire, encadrement, rédaction) en établissant toujours un planning. Cela m'a permis de **gérer** mon projet de thèse pour respecter au mieux le calendrier que nous avions défini au départ, en aménageant aussi quelques plages pour parer à d'éventuels aléas (indisponibilité de l'étiqueteur en ligne, ordinateur défaillant).

Aussi, grâce à l'enseignement dispensé pendant 3 ans, j'ai pu **approfondir** mes connaissances techniques en informatique, dans les domaines de la programmation web et dans la conception et l'exploitation des systèmes de bases de données et de bases documentaires (ressources électroniques).

3.2 Compétences « humaines »

Concernant l'encadrement, j'ai été **tutrice** industrielle de deux stagiaires en Master 2 recherche linguistique, informatique et traduction (stages effectués chez RBS) et j'ai suivi depuis son master 1 un étudiant de la même formation (qui a repris ses études en vue de gérer un groupe d'annotateurs chez VecSys) qui s'est chargé de l'annotation manuelle du corpus d'évaluation de notre outil. J'ai appris à **gérer** au quotidien les questions posées par les stagiaires, à propos de l'organisation de leurs recherches et de l'exploitation de leurs résultats. J'ai aussi développé mon sens de l'**écoute** et l'**adaptation** de mon discours suivant le niveau de connaissances (ou le caractère !) de chacun. Ces mêmes compétences ont été utilisées tout au long de ma collaboration avec l'entreprise RBS. J'ai dû user de persuasion pour **convaincre** mon chef de produit du bien-fondé de la recherche bibliographique, même si l'on pouvait penser que cette étape était non productive. J'ai appris à **déléguer** certains points de mon travail, pour tenir les délais fixés (mise en place de règles pour annoter les emplois impersonnels du pronom « il » par un stagiaire) ou lorsque mes connaissances s'avéraient insuffisantes (développement par l'ingénieur RBS de la coque Java de l'outil d'annotation automatique de marqueurs linguistiques) ; toujours en gardant le **contrôle** sur l'avancée et la bonne marche des différentes étapes. Cela m'a permis entre autres de m'assurer de la bonne compréhension du problème posé pour recadrer au besoin.

Avec aussi les présentations mensuelles (avec support Powerpoint) effectuées chez RBS, j'ai pu développer une certaine **aisance en expression orale** qui m'a énormément servi lors de mes interventions dans des conférences et plus récemment, depuis que je suis représentante des doctorants au conseil de l'école doctorale. En effet, j'ai pu **négocier** le réexamen d'un dossier de candidature pour une bourse doctorale d'un futur doctorant qui avait été écarté d'office par manque de connaissances des membres du jury. Je suis aussi membre actif de l'association DoXtra (doctorants en sciences humaines et sociales) qui permet, entre autres, d'**organiser** des conférences et des formations destinées aux doctorants, ce qui est encore un moyen d'échanges et de travail en groupe.

Une thèse demande en grande partie de la **persévérance**. Formatrice, la thèse permet de transformer nos échecs en expériences pour reconstruire notre façon de voir les choses.

Outre les compétences techniques spécifiques développées durant la thèse, certaines compétences sont transférables autant au milieu académique, qu'industriel. Cela concerne, entre autres, les compétences managériales, de conseil, de contrôle, la capacité à planifier son travail et être rigoureux pour mener à terme un projet.

4. Résultats et impacts de la thèse

Né d'un besoin formulé par l'entreprise RBS, les retombées directes de ce projet de recherche consistent, pour l'entreprise, à fournir à ses clients une solution d'indexation de documents déjà présente chez d'autres concurrents français ou étrangers, mais à moindre coût. Les apports de ce programme de recherche et de développement ont amené à une amélioration substantielle du moteur de recherche développé par RBS pour ses projets d'Intranet, Extranet et Sites Internet. Aux outils statistiques déjà déployés dans les solutions RBS, cette approche basée sur le sens même du contenu indexé fournit plus de services : détection automatique des personnes et sociétés citées dans le document, détermination de thématiques, proposition de documents semblables.

Côté universitaire, les outils et ressources développés à l'issue de la thèse sont à présent utilisés dans d'autres projets de recherche et ont donné naissance à une thèse et un mémoire de Master 2. Ainsi, L'étiqueteur TTL dans sa version française est à présent utilisé dans d'autres projets de recherche du laboratoire LiLPa, tels que le projet CAP. TTL est disponible gratuitement en ligne pour la communauté scientifique ; ceci dans une optique de partage de la connaissance. Le module d'identification des chaînes de référence *RefGen* va être utilisé dans un projet de recherche mené en partenariat avec le laboratoire Lattice. Dans ce cadre, *RefGen* (et ses modules d'annotation des entités nommées, des groupes nominaux complexes, …) sera utilisé comme outil de pré-annotation de corpus (pour faciliter cette lourde tâche d'annotation manuelle), afin de mettre en place un corpus de référence, annoté en chaînes de référence pour le français, toujours accessible gratuitement à la communauté scientifique (il n'en existe pas encore à l'heure actuelle pour le français). Enfin, *RefGen* va être optimisé dans le cadre d'une thèse sur la répétition lexicale, par le rajout, entre autres, d'ontologies (liste de mots reliés par des relations de synonymie ; par exemple « homme » et « humain » ; d'hyperonymie, comme « chat » et « animal », etc.) et en traitant d'autres genres textuels que ceux déjà étudiés.

Aussi, le travail de thèse a été plusieurs fois valorisé lors de communications à des séminaires, à des colloques nationaux (notamment au colloque « Traitement Automatique des Langues Naturelles » (TALN) qui est la référence du domaine) et internationaux. Certaines communications ont fait l'objet de publication dans des actes (7), des revues (3) et des chapitres d'ouvrages (2).

D'un point de vue personnel, la thèse et plus particulièrement la CIFRE m'a permis d'apprécier à la fois la vie en laboratoire et la vie en entreprise. J'ai pu prendre conscience des enjeux propres à chaque secteur et ainsi mieux cerner le monde industriel et académique. J'ai notamment apprécié la liberté qui nous a été accordée pour définir la méthode à utiliser pour détecter les thèmes (et la prise de risque encourue). D'autre part, la thèse m'a appris à oser aller vers les autres pour demander conseil, pour faire partie d'une communauté de chercheurs et ne pas rester isolée. Elle m'a montré l'importance du travail en groupe et des échanges qui ont enrichi mes recherches et m'ont fait prendre conscience de mon envie de partager mes connaissances.

5. Pistes professionnelles

La première piste professionnelle vers laquelle je tends est universitaire ; même si je suis bien consciente de l'actuelle difficulté pour obtenir un poste d'enseignant-chercheur à l'Université. Cela me permettrait de rester dans le domaine du TAL que j'affectionne particulièrement et de poursuivre mes recherches sur la résolution de la référence. C'est un domaine que j'ai découvert

et apprécié durant la thèse et que je souhaite vivement continuer à explorer. De plus, en étant maître de conférences, je pourrais faire partie intégrante du projet de constitution de corpus annoté en chaînes de co-référence (mené avec le laboratoire Lattice à Paris) et continuer ainsi à entretenir des relations étroites avec les membres du groupe de travail sur la coréférence que j'ai rejoint depuis 2 ans.

Bien évidemment, je n'exclue pas le secteur privé, qui m'a accompagné pendant 3 ans. Dans le cadre de ma thèse, j'ai plutôt amélioré l'existant (en optimisant les résultats du moteur de recherche interne de l'entreprise). A présent, je souhaiterai intervenir dans un projet d'innovation (développer un nouveau produit) et animer une équipe. Le secteur du TAL ne m'est pas inconnu car j'ai pu acquérir une connaissance du marché dans ce secteur lors de mes recherches de stages de Master. Ainsi, les postes de chef de projet en linguistique informatique, d'ingénieur assurance qualité linguiste ou d'ingénieur support linguiste – service clients me permettraient de modéliser des produits et solutions innovantes issues du TAL, tout en gardant un contact étroit avec les clients.

Les entreprises sont de plus en plus à la recherche de solutions innovantes incluant de la sémantique pour retrouver de manière plus pertinente leurs documents ou effectuer de la veille stratégique (flux d'actualités, réseaux sociaux, etc.). Des sociétés comme *Viavoo* fournissent déjà des solutions permettant d'analyser des « remontées » clients via les sites communautaires, les emails, etc. Leurs solutions utilisent des plates-formes sémantiques et font appel à des ontologies. D'autres sociétés comme *Synomia* proposent une navigation assistée, un moteur de recherche « intelligent » personnalisé pour guider les utilisateurs. D'autres sociétés encore, telles que *Syllabs*, permettent d'extraire des informations (caractéristiques d'un produit, opinions), les structurer et les enrichir d'informations sémantiques.

Notons bien que ces deux perspectives de carrière (académique et industrielle) ne s'excluent pas, bien au contraire. Si je deviens enseignant-chercheur, je souhaite conserver des liens avec le monde de l'entreprise et intervenir ponctuellement en tant que consultante. Si j'intègre une entreprise privée, je souhaiterai poursuivre mes collaborations avec l'Université pour des projets dans le domaine du TAL, ou bien intervenir dans le cadre de formations.

Au terme de plusieurs années menées dans l'un ou l'autre des secteurs et grâce à l'expérience que j'aurai acquise en enseignement et en management, une dernière piste professionnelle, qui s'écarte de mon domaine de recherche, me mènerait vers les métiers du conseil ou du coaching. Je souhaiterai intégrer un cabinet de conseil en management de la personne, pour accompagner les personnes dans leur recherche d'emploi ou dans leur réorientation professionnelle ; les écouter, les aider à valoriser leurs compétences, à prendre confiance en eux et à s'organiser dans leurs démarches. Aussi, le coaching de vie m'attire car il intervient autant dans la vie professionnelle que personnelle, en analysant le discours des personnes pour leur fournir des méthodes et des outils adaptés, les orienter et les soutenir dans leurs projets.

Table des figures

Liste des tableaux

Index des auteurs

Publications

Publications liées à la thèse

Communications à des manifestations internationales à comité de lecture

- Schnedecker, C., **Longo, L.** (2012). Impact des genres sur la composition des chaînes de référence : le cas des faits divers, *Actes du Congrès Mondial de Linguistique Française (CMLF)*, 4 – 7 juillet 2012, Lyon, France, pp. 1957-1972.
- **Longo, L**., Todiraşcu, A. (2010). RefGen: a Tool for Reference Chains Identification, *Actes de CLA'10 (Computational Linguistics-Applications)*, IMCSIT (International Multiconference on computer Science and Information Technology), pp. 447-454, 18-20 octobre 2010, Wisla, Pologne (indexé par DBLP-Trier).
- **Longo, L**., Todiraşcu, A. (2010). RefGen: Identifying Reference Chains to Detect Topics, *Actes des 4th International Workshop on Distributed Agent-Based Retrieval Tools (DART 10)*, 18 juin 2010, Genève, Suisse (indexé par DBLP-Trier).

Communications à des manifestations nationales à comité de lecture

- **Longo, L**., Todirascu, A. (2012). Chaînes de référence et genre textuel pour la détection automatique de thèmes, *Actes de la journée ATALA*, Paris, 11 mai 2012.
- **Longo, L**., Todiraşcu, A. (2011). RefGen, outil d'identification automatique des chaînes de référence en français, session démonstrations industrielles, *Actes de la conférence TALN 2011*, Montpellier, 27 juin-1er juillet 2011.
- **Longo, L**., Todirascu, A. (2011). *RefGen* : un système d'identification automatique de chaînes de référence, *Actes de la journée ATALA*, Paris, 20 juin 2011.
- **Longo, L**. (2011). Un corpus de genres textuels variés pour l'identification automatique des chaînes de référence, *Actes du* ***VIIe Colloque jeunes chercheurs*** *Praxiling*, Montpellier, 9-10 juin 2011.
- **Longo, L.,** Todiraşcu, A. (2010). La saillance référentielle pour la détection des thèmes, *actes du colloque Saillance 2*, Strasbourg, 19-20 novembre 2010.
- **Longo, L.,** Todiraşcu, A. (2010). *RefGen* : un module d'identification des chaînes de référence dépendant du genre textuel, *actes de TALN10,* session poster*,* 19-23 juillet 2010, Montréal, Canada.

Articles dans des revues nationales à comité de lecture

- **Longo L.**, Todiraşcu, A. (2013). Une étude de corpus pour la détection automatique de thèmes. *Revue électronique Texte et corpus*, 4, 143-155, ISSN : 1958-5306.
- **Longo, L.** (2010). Un corpus pour optimiser l'identification automatique des chaînes de référence, in Azzopardi S. (coord). Corpus, Données, Modèles. *Cahiers de Praxématique*, 54-55, PULM, Montpellier, 249-262.
- **Longo, L.**, Todiraşcu, A. (2009). Une étude de corpus pour la détection automatique des thèmes, actes des 6èmes journées de linguistique de corpus, 10-12 septembre 2009, Lorient.
- **Longo, L.** (2009). Un outil de détection automatique de thèmes, actes du Forum Jeunes Chercheurs / INFORSID, pp. 467-468, 26-28 mai 2009, Toulouse (indexé par DBLP-Trier).

Articles dans des revues internationales à comité de lecture

- **Longo, L.**, Todiraşcu, A. (2010). Genre-based Reference Chains Identification for French, *Investigationes Linguisticae*, Volume XXI, 57-75, ISSN : 0378-4169.
- Todiraşcu, A., Ion, R., Navlea, M., **Longo, L**. (2011). French Text Preprocessing with TTL, in *Proceedings of the Romanian Academy – Series A (Mathematics, Physics, Technical Sciences, Information Science)*, Ed. Romanian Academy, Publishing House of the Romanian Academy, Volume 12, number 2/2011, 151-158, ISSN : 1454-9069.

Chapitres d'ouvrages

- **Longo, L.**, Todiraşcu, A. (2011). RefGen: Identifying Reference Chains to Detect Topics, in *Studies in Computational Intelligence*, volume 361, chapitre 3, 27-40, Ed.: Kacprzyk, J., "Advances in Intelligent and Soft Computing", Springer Verlag, ISSN : 1860-949X.
- **Longo, L.**, Todiraşcu, A. (2012). Une approche automatique pour l'identification de référents saillants en discours, Saillance 2, Presses Universitaires de Franche-Comté.

Hors thèse

Article en soumission

- **Longo L.**, Todiraşcu, A. Etude des chaînes de référence dans des textes non-narratifs, *Langages*, (soumis).

Communications à des manifestations internationales à comité de lecture

– **Longo, L.**, Todirascu, A. (2013). Vers une modélisation des chaînes de référence dans des textes non-narratifs, *Actes du colloque International Corpus et Outils en Linguistique, Langues et Parole : Statuts, Usages et Mésusages*, 3 – 5 juillet 2013, Strasbourg, France.

Articles dans des revues nationales à comité de lecture

– Todiraşcu, A., Grass, T., Navlea, M., **Longo, L**. (accepté). La relation de hiérarchie « chef » : une approche translingue français-anglais-allemand, *Revue Meta*.

Bibliographie

Adam, C. (2007). Traitement automatique de la coréférence pour une application de veille terminologique, Mémoire de Master 1, Université de Toulouse 2 Le Mirail, 92p.

Adam, C. (2012). Voisinage lexical pour l'analyse du discours, Thèse de doctorat, Université de Toulouse 2 Le Mirail, 290p.

Adam, C., Morlane-Hondère, F. (2009). Détection de la cohésion lexicale par voisinage distributionnel : application à la segmentation thématique, *actes de Récital*, Senlis, France.

Adam, J.-M. (1977). Ordre du texte, ordre du discours, *Pratiques*, 13, 103-111.

Adam, J.-M. (1990). Éléments de linguistique textuelle. Bruxelles-Liège : Mardaga.

Adam, J.-M. (2004). Linguistique textuelle. Des genres de discours aux textes. Paris, Nathan, coll. « Fac-Linguistique », 208p.

Adam, J.-M. (2011). Les textes : types et prototypes. Récit, description, argumentation, explication et dialogue, 3e édition, coll. « Fac-Linguistique », Armand Colin, 223p.

Aery, M., Ramamurthy, N., Alp Aslandogan, Y. (2003). Topic identification of textual data. *Technical report CSE-2003-25*. Department of computer science and engineering, University of Texas at Arlington.

Afantenos, S., Asher, N., Benamara, F., Bras, M., Fabre, C. Ho-Dac, M., Le Draoulec, A., Muller, P., Pery-Woodley, M.P. Prevot, L., Rebeyrolles, J., Tanguy, L., Vergez-Couret, M., Vieu, L. (2012). An empirical resource for discovering cognitive principles of discourse organisation: the ANNODIS corpus. In eds. Calzolari, N., Choukri, K., Declerck, T., Uğur Doğan, M., Maegaard, B., Mariani J., Odijk, J., Piperidis, S., *Proceedings of the Eight International Conference on Language Resources and Evaluation*, 23-25 may, Istanbul, Turkey, European Language Resources Association (ELRA), 2727-2734.

Agnès, Y. (2009). Pratiquer et transmettre les genres journalistiques, *in* Les genres journalistiques. Savoirs et savoir-faire, *in* Utard J.-M., Ringoot R., *Les genres journalistiques. Savoirs et savoir-faire*, Paris, L'Harmattan, 25-35.

Ailloud, E., Klenner, M. (2009). Vers des contraintes plus linguistiques en résolution de coréférences, *Actes de TALN*, Senlis, France.

Allan, J., Carbonell, J., Doddington, G., Yamron, J., Yang, Y. (1998). Topic Detection and Tracking Pilot Study. Final Report, *Proceedings of the DARPA Broadcast News Transcription and Understanding*.

Allan, J. (2002). Introduction to topic detection and tracking. In: J. Allan (Ed.), Topic Detection and Tracking: Event-Based Information Organization, 1-16. Boston, MA: Kluwer Academic Publishing.

Alshawi, H. (1987). Memory and Context for Language Interpretation. Cambridge, Cambridge University Press.

Altun, Y., Johnson, M. & Hofmann, T. (2003). Investigating loss functions and optimization methods for discriminative learning of label sequences, *Actes de EMNLP*.

Amsili, P., Denis, P., Roussarie, L. (2005). Anaphores abstraites en français : représentation formelle, *revue TAL*, *46* (1), 15-39.

Anscombre, J.-C. (2001). À propos de mécanismes sémantiques de formation de certains noms d'agent en français et en espagnol, *Langages*, *143*, 28-48.

Anscombre, J.-C. (2003). L'agent ne fait pas le bonheur : agentivité et aspectualité dans certains noms d'agent en espagnol et en français, *Thélème, Revista Complutense de Estudios Franceses*, *11*, 11-27.

Aone, C., Bennett, S. W. (1995). Evaluating automated and manual acquisition of anaphora resolution strategies, *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, Cambridge, Mass., 122-129.

Ariel, M. (1990). Accessing Noun-Phrase Antecedents. Londres : Routledge.

Ariel, M. (1996). Referring expressions and the +/- coreference distinction, In T. Fretheim & J.K. Gundel (eds.), 13-25.

Ariel, M. (2001). « Accessibility theory: An overview », *in* Sanders T, Schilperoord J., Spooren W., (dir.), *Text Representation*, Amsterdam, Benjamins, 29-87.

Asher, N. (1993). Reference to Abstract Objects in Discourse. Kluwer, Dortrecht, 455p.

Asher, N., Denis, P., Reese, B., (2006). Names and pops and discourse structure, *Proceedings of the workshop on constraints in discourse*, Maynooth, National University of Ireland, 11-18.

Ašić, T. (2008). Espace, temps, prépositions, *Langue et culture, 41*, Librairie Droz, 319p.

Auclair, G. (1970). Le mana quotidien : structures et fonctions de la chronique des faits divers. Paris : Anthropos.

Azé, J., Heitz, T., Mela, A., Mezaour, A. D., Peinl, P., Roche, M. (2006). Présentation de DEFT'06 (DEfi Fouille de Textes), *Actes de DEFT'06*, *1*, 3-12.

Azzam, S., Humphreys, K., Gaisauskas, R. (1998). Evaluating a focus-based approach to anaphora resolution, *Actes de COLING-ACL'98*, Montréal.

Bagga, A., Baldwin, B. (1998). Algorithms for scoring coreference chains, *Proceedings of the LREC 1998 Workshop on Linguistic Coreference*, Granada, Spain, 563-566.

Bakhtine, M. M. (1975). Esthétique et théorie du roman. Paris, Gallimard, (édition originale 1978).

Baldwin, B. (1997). CogNIAC: high precision coreference with limited knowledge and linguistic resources, *Proceedings of ACL/EACL workshop on Operational factors in practical, robust anaphora resolution (ACL97)*, Madrid, Espagne, 38-45.

Barbu, C., Mitkov, R. (2001). Evaluation tool for rule-based anaphora resolution methods, *Proceedings of the 39th Annual Meeting on ACL*, 34-41.

Barthes, R. (1964). Structure du fait divers. Essais critiques. Paris : Seuil, 188-197.

Barthes, R. (1981). Essais critiques. Paris : Seuil, (édition originale 1964).

Bartning, I. (1996). Eléments pour une typologie des SN complexes en de en français. *Langue française*, *109*, 29-43.

Barzilay, R., Elhalad, M. (1997). Using lexical chains for text summarization, *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and the 8th European Chapter Meeting of the Association for Computational Linguistics*, Workshop on Intelligent Scalable Text Summarization, 10-17, Madrid.

Baumer, E. (2011a). Etude contrastive français / anglais des anaphores lexicales dans la presse et dans la fiction littéraire, *CORELA - RJC Cotexte, contexte, situation / Numéros thématiques*. [En ligne]. URL : http://corela.edel.univ-poitiers.fr/index.php?id=2233 Consulté le 09/04/2012.

Baumer, E. (2011b). Noms propres et anaphores lexicales en anglais et en français : étude comparée des chaînes de référence. Thèse de doctorat, Paris Diderot, 378p.

Bean, D. L., Riloff, E. (2004). Unsupervised Learning of Contextual Role Knowledge for Coreference Resolution, *Proceedings of HLT-NAACL*, 297-304.

Beaver, D. (2000). The optimization of discourse. Stanford University.

Beaver, D. (2002). The optimization of discourse anaphora. Stanford University.

Beaver, D. (2004). The optimization of discourse anaphora, *Linguistics and Philosophy*, *27*(1), 3-56.

Béchet, F., Sagot, B., Stern, R. (2011). Coopération de methodes statistiques et symboliques pour l'adaptation non-supervisée d'un système d'étiquetage en entités nommées, *Actes de TALN*, Montpellier.

Beeferman, D., Berger, A., Lafferty, J. D. (1999). Statistical models for text segmentation, *Machine Learning*, *34* (1–3), 177–210.

Bengtson, E., Roth, D. (2008). Understanding the value of features for coreference resolution, *Proceedings of EMNLP 2008*, 294-303, Honolulu, Hawaii.

Benveniste, E. (1974). Problèmes de linguistique générale, t.2, chap. 8, 113-125.

Berge, C. (1989). Hypergraphs: The Theory of Finite Sets. Amsterdam, Netherlands: North-Holland, 254p.

Bergler, S., Witte, R., Khalife, M., Li, Z., Rudzicz, F. (2003). Using knowledge-poor coreference resolution for text summarization, *Proceedings of the HLT Workshop on Automatic Summarization, DUC 2003*, Edmonton, Canada.

Bernardini, A., Carpineto, C., D'Amico, M. (2009). Full-subtopic retrieval with keyphrase-based search results clustering, *Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*, 1, Washington, DC, USA: IEEE Computer Society, 206-213.

Berrendonner, A. (1994). « Anaphores confuses et objets indiscrets » in Schnedecker, C. *et al.* (L'anaphore associative. Aspects linguistiques, psycholinguistiques et automatiques : éds.), Paris : Klincksieck, 209-230.

Berthoud, A.-C. (1996). Paroles à propos. Approche énonciative et interactive du topic, Paris : Ophrys.

Bessonnat, D. (1988). Le découpage en paragraphes et ses fonctions, *Pratiques*, *57*, 81-105.

Bestgen, Y. (2004). Analyse sémantique latente et segmentation automatique des textes. *In* G. Purnelle, C. Fairon, & A. Dister (Eds.), *Actes des septième Journées internationales d'Analyse statistique des Données Textuelles*, 171-181, Louvain-la-Neuve.

Bestgen, Y. (2005). Amélioration de la segmentation automatique des textes grâce aux connaissances acquises par l'analyse sémantique latente, *Actes de TALN*, Dourdan, 6-10 juin, 203-212.

Bestgen, Y. (2006). Improving Text Segmentation Using Latent Semantic Analysis: A Reanalysis of Choi, Wiemer-Hastings, and Moore (2001), *Computational Linguistics, 32 (1)*, 5-12.

Bestgen, Y. (2012). Évaluation automatique de textes et cohésion lexicale, *Discours* [En ligne], 11 | 2012, mis en ligne le 23 décembre 2012, consulté le 06 mars 2013. URL : http://discours.revues.org/8724 ; DOI : 10.4000/discours.8724

Bestgen, Y., Piérard, S. (2006). Comment évaluer les algorithmes de segmentation automatique ? Essai de construction d'un matériel de référence, *Actes de TALN* : Verbum ex machina, Louvain-La-Neuve, Presse universitaire de Louvain, 407-414.

Biber, D. (1989). A typology of English texts. *Linguistics*, *27*, 3-43.

Biber, D. (1994). « Representativeness in corpus design », Linguistica Computazionale, vol. IX-X, 377-408. *Current Issues in Computational Linguistics*: in honor of Don Walker.

Biber, D. (1995). *Dimensions of register variation: a cross-linguistic comparison*. Cambridge: Cambridge University Press.

Biber, D., Conrad, S. (2009). Register, genre and style. Cambridge: Cambridge University Press, 344p.

Bick, E. (2001). The VISL System: Research and applicative aspects of IT-based learning, *Actes de NoDaLiDa 2001 (Uppsala)*.

Bilhaut, F. (2006). Analyse automatique de structures thématiques discursives, application à la recherche d'information. Thèse de doctorat, Université de Caen, 304p.

Bilhaut, F. (2007). Analyse thématique automatique fondée sur la notion d'univers de discours, *Discours*, *1*.

Bilhaut, F., Enjalbert, P. (2005). Discourse Thematic Organisation Reveals Domain Knowledge Structure, *Proceedings of the 2nd Indian International Conference on Artificial Intelligence (IICAI'05)*, Pune, India, 2815-2831.

Bilhaut, F., Ho-Dac, M., Borillo, A., Charnois, T., Enjalbert, P., Le Draoulec, A., Mathet, Y., Miguet, H., Péry-woodley, M.-P., Sarda, L. (2003). Indexation discursive pour la navigation intradocumentaire : cadres temporels et spatiaux dans l'information géographique, *Actes de TALN*, Batz-sur-Mer, 315-320.

Bittar, A. (2006). Un algorithme pour la résolution d'anaphores événementielles, Mémoire de Master, Université Paris 7, 69p.

Blanchard, A. (2009). La cartographie des brevets dans l'industrie et la recherche : outils et pratiques, *Atelier Veille numérique, regards pluridisciplinaires, 9èmes*

Journées francophones Extraction et gestion des connaissances, Francis Chateauraynaud, Jean-Gabriel Ganascia, Julien Velcin (Eds.), Strasbourg, 3-10.

Blanche-Benveniste, C. (1990). Le français parlé - Etudes grammaticales. Paris, édition CNRS.

Bontcheva, K., Dimitrov, M., Maynard, D., Tablan, V., Cunningham, H. (2002). Shallow Methods for Named Entity Coreference Resolution. Chaînes de références et résolveurs d'anaphores, *Actes de l'atelier TALN 2002*, Nancy, 173-181.

Bosredon, B., Galmiche, M. (1992). Le thème. Présentation, *L'information grammaticale*, *54*.

Bouchekif, A., Damnati, G., Charlet, D. (2013). Segmentation thématique : processus itératif de pondération intra-contenu, *Actes de TALN*, 17-21 juin, Les Sables d'Olonne, 739-746.

Boudreau, S., Kittredge, R. (2005). Résolution des anaphores et détermination des chaînes de coréférences, *Traitement Automatique des Langues (TAL)*, *46*, 41-70.

Boudreau, S., Kittredge, R. (2006). Résolution d'anaphores et identification des chaînes de coréférence : une approche minimaliste, *Actes des 8èmes Journées internationales d'Analyse statistique des Données Textuelles (JADT)*, 201-210.

Boufaden, N., Lapalme, G., Bengio, Y. (2002). Découpage thématique des conversations : un outil d'aide à l'extraction, *Actes de TALN*, 24-27 juin, Nancy, 119-129.

Boufaden, N., Lapalme, G., Bengio, Y. (2010). Segmentation en thèmes de conversations téléphoniques : traitement en amont pour l'extraction d'information, *Actes de TALN2010*, session poster, Montréal, Canada, 377-382.

Bourigault, D., Fabre, C., Frérot, C., Jacques, M.-P., Ozdowska, S. (2005). Syntex, analyseur syntaxique de corpus, *Actes des 12èmes journées sur le Traitement Automatique des Langues Naturelles*, Dourdan, France.

Boyd, A., Gegg-Harrison, W., Byron, D. (2005). Identifying non-referrential it: A machine learning approach incorporating linguistically motivated patterns (expanded version). In *Special Issue of the Journal TAL: Models and algorithms for anaphora resolution*, Busquet, J., Hardt, D., *46* (1), 71-90.

Brants, T. (2000). TnT – A Statistical Part-of-Speech Tagger, *Proceedings of the 6th Applied Natural Language Processing Conference*, Seattle, USA, 224-231.

Brants, T., Chen, F., Tsochantaridis, I. (2002). Topic-based document segmentation with probabilistic latent semantic analysis, *Proceedings of Conference on Information and Knowledge Management*, 211–218.

Brill, E. (1994). Some Advances in Transformation-Based Part-of-Speech Tagging, *Proceedings of AAAI*, 1, 722-727.

Broscheit, S., Poesio, M., Ponzetto, S.-P., Rodriguez, K. J., Romano, L., Uryupina, O., Versley, Y., Zanoli, R. (2010). Bart: A multilingual anaphora resolution system, *Proceedings of the Semantic Evaluation Workshop (SemEval-2)*, 104-107.

Brown, G., Yule, G. (1983). *Discourse Analysis.* Cambridge University Press.

Brun, A., Smaili, K., Haton, J.-P. (2002). WSIM : une méthode de détection de thèmes fondée sur la similarité entre mots, *Actes de TALN*, Nancy, 145-154.

Brunetti, L., Avanzii, M., Gendrotk, C. (2012). Entre syntaxe, prosodie et discours : les topiques sujet en français parlé, *Actes du Congrès Mondial de Linguistique Française*, Lyon, 2041-2054.

Brunn, M., Chali, Y., Pinchak, C. J. (2001). Text Summarization Using Lexical Chains, *Proceedings of the Document Understanding Conference*, New Orleans, 135-140, Published by NIST.

Burgess, C., Livesay, K., & Lund, K. (1998). Explorations in context space: Words, sentences, discourse. *Discourse Processes*, *25*, 211-257.

Cadiot, P., Fradin, B. (1988). Présentation. Une crise en thème ? *Langue française*, *78* (1), 3-8.

Cai, J. Strube, M. (2010). Evaluation metrics for end-to-end coreference resolution systems, *Proceedings of SIGDIAL*, University of Tokyo, Japan, 28-36.

Callan, J. P., Croft, W. B. and Harding, S. M. (1992). The INQUERY retrieval system, *Proceedings of the International Conference on Database and Expert Systems Application*, Berlin and New York: Springer-Verlag, 78-83.

Cardie, C., Wagstaff, K. (1999). Noun phrase coreference as clustering, *Proceedings of the 1999 joint SIGDAT Conference on Empirical Methods in Natural Language Processing And Very Large Corpora*, 82-89.

Caron, F. (2006). Inférence bayésienne pour la détermination et la sélection de modèles stochastiques, Thèse de doctorat, Université de Sciences et Technologies de Lille, 210p.

Carretero, M. (1998). A proposal of a topic structure model for expository texts, *Estudios Ingleses de la Universidad Complutense, 6*, Madrid, 223-237.

Carter-Thomas, S. (2000). La cohérence textuelle – pour une nouvelle pédagogie de l'écrit, Paris, L'Harmattan, coll. Langue et parole.

Carter-Thomas, S. (2009). Texte et contexte pour une approche fonctionnelle et empirique, mémoire d'habilitation à diriger les recherches, Lattice, Paris.

Carthy, J., Smeaton, A. F. (2000). The design of a topic tracking system, *Proceedings of the 22nd Annual Colloquium on Information Retrieval Research*, Cambridge, England, The Information Retrieval Specialist Group of the British Computer Society.

Ceauşu, A. (2006). Maximum Entropy Tiered Tagging, *Proceedings of the Eleventh ESSLLI Student Session*, Malaga, Spain, 173-179.

Chafe, W. (1970). Meaning and the structure of language. Chicago: University of Chicago Press.

Chafe, W. (1976). Givenness, contrastiveness, definiteness, subjects, topics and point of view, *in* Li, C. N. (ed.), *Subject and Topic*, New York; Academic Press, 25-55.

Chafe, W. (1980). The Pear Stories. Ablex Publishing Corporation, Norwood, NJ.

Chafe, W. (1987). Cognitive constraints on information flow. Coherence and Grounding in Discourse, ed. by Russel Tomlin. Amsterdam: John Benjamins.

Chafe, W. (1994). Discourse, consciousness, and time: The flow and displacement of conscious experience in speaking and writing. Chicago: University of Chicago Press, 340p.

Chafe, W. (2001). The Analysis of Discourse Flow, The Handbook of Discourse Analysis, Blackwell.

Chafe, W. (2008). Aspects of discourse analysis, *Brno Studies in English*, *34*, 23-37.

Chali, Y. (2001). Topic detection using lexical chains, *in* L. Monostori, J. Váncza, and M. Ali (Eds.): IEA/AIE 2001, LNAI 2070, 552-558.

Chambers, N., Jurafsky, D. (2008). Unsupervised Learning of Narrative Event Chains, *Proceedings of ACL-08: HLT*.

Chambers, N., Jurafsky, D. (2011). Template-Based Information Extraction Without the Templates, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 1, Stroudsburg, PA, USA, 976-986.

Charniak, E., Elsner, M. (2009). EM Works for Pronoun Anaphora Resolution, *Actes de EACL 2009*, Athènes, Grèce.

Charolles, M. (1987). Contraintes pesant sur la configuration des chaînes de référence comportant un nom propre, *Cahiers du centre de recherches sémiologiques de Neuchâtel*, *53*, 29-55.

Charolles, M. (1988). Les plans d'organisation textuelle : périodes, chaînes, portées et séquences, *Pratiques*, *57*, 3-15.

Charolles, M. (1995a). Anaphore associative. Problèmes de délimitation, *dans* Schnedecker, C. *et al.*, L'anaphore associative, aspects linguistiques, psycholinguistiques et automatiques, Paris, Klincksieck, 67-92.

Charolles, M. (1995b). Cohésion, cohérence et pertinence du discours, *Travaux de Linguistique*, *29*, 125-151.

Charolles, M. (1997). L'encadrement du discours : univers, champs, domaines et espaces, *Cahier de Recherche Linguistique*, *6*, 1-73.

Charolles, M. (1999). Associative Anaphora and its Interpretation, *Journal of Pragmatics*, *31* (3), 311-326.

Charolles, M. (2003). De la topicalité des adverbiaux détachés en tête de phrase, *dans* M. Charolles & S. Prévost (eds). Adverbiaux et topiques, Louvain la Neuve, *Travaux de Linguistique*, *47*, 11-51.

Charolles, M. (2009). Les cadres de discours comme marques d'organisation des discours. In Venier, F (Ed.) *Tra Pragmatica e Linguistica Testuale*, edizioni dell'Orso, Alessandria, 401-420.

Charolles, M., Péry-Woodley, M.-P. (2005). Introduction, dans Charolles, M., Péry-Woodley, M.-P., eds., Les adverbiaux cadratifs, *Langue Française*, *148*, 3-8.

Charolles, M., Prévost, S. (2003) (Eds.). Adverbiaux et topiques, *Travaux de Linguistique*, *47*.

Charolles, M., Vigier, D. (2005). Les adverbiaux en position préverbale : portée cadrative et organisation des discours, *Langue Française*, *148*, 9-30.

Charton, E., Gagnon, M., Jean-Louis, L. (2013). Influence des annotations sémantiques sur un système de détection de coréférence à base de perceptron multi-couches, *Actes de TALN*, 17-21 juin, Les Sables d'Olonne, 612-619.

Chastain, C. (1975). « Reference and Context », *In* Gunderson K. (dir.), Language, Mind and Knowledge, Minneapolis: University of Minnesota Press.

Chauché, J. (1984). Un outil multidimensionnel de l'analyse du discours, *Proceedings of Coling'84*, 11-15.

Chaumartin, F. (2007). Résolution d'anaphores dans une encyclopédie en langue anglaise : conception, implémentation et évaluation des performances, *Actes de la Journée La résolution des anaphores en Traitement Automatique des*

Langues, sous l'égide de l'ATALA (Association pour le Traitement Automatique des Langues), 16 juin 2007, Paris.

Choi, F. Y. Y. (2000). Advances in domain independent linear text segmentation, *Proceedings of the 1st Meeting of the North American Chapter of the ACL*, Seattle, USA, 26-33.

Choi, F. Y. Y., Wiemer-Hastings P., Moore J. (2001). Latent semantic analysis for text segmentation, *Actes de NAACL'01*, 109-117.

Chomsky, N. (1981). Lectures on Government and Biding. Dordrecht: Foris, 371p.

Chomsky, N. (1982). Some concepts and consequences of the theory of Government and Binding. Chicago: MIT Press.

Chomsky, N. (1986). Barriers, Linguistic Inquiry Monograph 13, MIT Press, 102p.

Church, K-W. (1993). Char align: A program for aligning parallel texts at the character level, *Proceedings of the 31st ACL*.

Clark, H. H., Clark, E. V. (1977). Psychology and Language: An Introduction to Psycholinguistics. New York: Harcourt, Brace & Jovanovich, 608p.

Clark, H., Haviland, S. (1977). Comprehension and the given-new contract. *In* Freedie, R. (Ed.), Discourse Production and Comprehension. Hillsdale, N. J.: Lawrence Erlbaum Associates, 1-40.

Clark, H. H., Marshall, C. R. (1981). Definite reference and mutual knowledge. *In* Joshi, A.K., Webber, B. & Sag, I. (Eds.), *Elements of discourse understanding.* Cambridge: Cambridge University Press, 10-63.

Clark, J., Gonzàlez-Brenes, J. (2008). Coreference resolution: Current trends and future directions. *Technical report*, The Language Techonologies Institute, CMU, 19p.

Claveau, V., Lefèvre, F. (2011a). Segmentation thématique : apport de la vectorisation, *Actes de la conférence CORIA - Conférence en recherche d'information et applications*, France.

Claveau, V., Lefèvre, S. (2011b). Topic segmentation of TV-streams by mathematical morphology and vectorization, *Proceedings of the 12th International Conference of the International Speech Communication Association*, Interspeech, 1105–1108.

Combettes, B. (1977). Ordre des éléments dans la phrase et linguistique du texte, *Pratiques*, *13*, 81-101.

Combettes, B. (1978). Thématisation et progression thématique dans les récits d'enfants, *Langue Française*, *38*, 74-86.

Combettes, B. (1988). Pour une grammaire textuelle : la progression thématique. De Boeck, 2ème édition, De Boeck/Duculot : Paris-Gembloux.

Combettes, B. (1992). Hiérarchie des référentes et connaissance partagée : Les degrés dans l'opposition connu/nouveau, *L'information grammaticale, 54*, 11-14.

Combettes, B. (1996). Facteurs textuels et facteurs sémantiques dans la problématique de l'ordre des mots : le cas des constructions détachées. In: *Langue française*, *111* (1), L'ordre des mots, 83-96.

Condamines, A. (2005). Anaphore nominale infidèle et hyperonymie : le rôle du genre textuel, *Revue de Sémantique et Pragmatique*, *18*, 33-52.

Connolly, D., Burger, J. D., Day, D. S. (1994). A machine learning approach to anaphoric reference, *Proceedings of International Conference on New Methods in Language Processing*, 255-261.

Corblin, F. (1985). Les chaînes de référence : analyse linguistique et traitement automatique, *Intellectica*, *5* (1), 123-143.

Corblin, F. (1995a). Les chaînes de référence dans le discours. Presses Universitaires de Rennes.

Corblin, F. (1995b). Les formes de reprise dans le discours : Anaphores et chaînes de référence. Presses Universitaires de Rennes.

Corblin, F. (2005). Les chaînes de la conversation et les autres, in J.-M. Gouvard (ed.) De la langue au style, Lyon : Presses universitaires de Lyon, 233-254.

Cornish, F. (1986). *Anaphoric relations in English and French: a discourse perspective*. Londres: Croom Helm.

Cornish, F. (1989). review article : Discourse, structure and anaphora: Written and conversational (B. Fox), *Lingua 79*, (2-3), 229-243.

Cornish, F. (1995). « Référence anaphorique, référence déictique, et contexte prédicatif et énonciatif », dans *Numéro spécial de Sémiotiques*, *8*, Anaphores : marqueurs et interprétations, 31-55.

Cornish, F. (1998). Les chaînes topicales : leur rôle dans la gestion et la structuration du discours, *Cahiers de Grammaire*, *23*, 19-40.

Cornish, F. (2000). L'accessibilité cognitive des référents, le Centrage d'attention et la structuration du discours : une vue d'ensemble, *Verbum*, *22* (1), 7-30.

Couto, J. (2006). Une plate-forme informatique de Navigation Textuelle : modélisation, architecture, réalisation et applications de Navi Texte, Thèse de Doctorat, Université Paris-Sorbonne - Paris IV.

Couto, J., Ferret, O., Grau, B., Hernandez, N., Jackiewicz, A., Minel, J., Porhiel, S. (2004). RÉGAL, un système pour la visualisation sélective de documents, *Revue d'Intelligence Artificielle*, 481-514.

Croft, W., Cruse, D.A. (2004). Cognitive Linguistics, Cambridge University Press.

Cruse, D. A. (1986). Lexical semantics. Cambridge: Cambridge University Press.

Cunningham, H. (2002). GATE, a General Architecture for Text Engineering, *Computers and the Humanities*, *36*, 223–254.

Dagan, I., Itai, A. (1991). "A statistical filter for resolving pronoun references," Feldman Y. A., Bruckstein A., eds., *Artificial Intelligence and Computer Vision*, Elsevier Science Publishers B.V, 125-135.

Dal, G., Namer, F. (2005). L'exception infirme-t-elle la notion de règle ? ou le lexique construit et la théorie de l'optimalité ?, *Faits de Langues, 25*, 123-130.

Daneš, F. (1968). Some Thoughts on the Semantic Structure of the Sentence, *Lingua*, *21*, 55-59.

Daneš, F. (1974). Functional sentence perspective and the organisation of the text, *In* Daneš F. (dir.), Paper in Functional Sentence Perspective, Prague Academia.

Danlos, L. (2005). ILIMP : outil pour repérer les occurrences du pronom impersonnel *il*, *Actes de TALN'05,* Dourdan, France, 123-132.

Danlos, L. (2006). Verbes causatifs, Discours causaux et Coréférence événementielle, *Revue Lynx*, *54*, 233-246.

Daumé III, H., Marcu, D. (2005). A large-scale exploration of effective global features for a joint entity detection and tracking model, *Proceedings of the Human Language Technology Conference and the 2005 Conference on Empirical Methods in Natural Language Processing*, Vancouver, B.C., Canada, 6-8 October, 97-104.

Davi, A., Haughton, D., Nasr, N., Shah, G. Skaletsky, M., Spack, R. (2005). A review of two text mining packages: SAS TextMining and WordStat. *The American Statistician*, *59* (1), 89-103.

Davidson, D. (1967). La forme logique des phrases d'action. Actes et événements, 149-198.

De Clercq, O., Hoste, V., Hendrickx, I. (2011). Cross-Domain Dutch Coreference Resolution, *Proceedings of the 8th International Conference on Recent Advances in Natural Language Processing*, Hissar, Bulgarie, 186-193.

Deerwester, S., Dumais, S., Furnas, G. W., Thomas K. Landauer, T. K., Harshman, R. (1990). Indexing by Latent Semantic Analysis, *Journal of the Society for Information Science*, *41* (6), 391-407.

Deleu, C. (2005). Le monde selon le nouveau Détective : quand le fait divers renonce au réel, *Les Cahiers du journalisme, 14*, 76-93.

Demol, A. (2010). Les pronoms anaphoriques il et celui-ci, De boeck Duculot, Champs linguistiques, 393p.

De Mulder, W. (1997). Les demonstratifs : des indices de changement de contexte, *in* N. Flaux, D. Van de Velde, W. de Mulder (eds) Entre général et particulier : les Déterminants, Artois Press Université, 137-200.

Denis, P. (2007). New learning models for reference resolution. Thèse de doctorat, Université du Texas, Austin.

Denis, P., Baldridge, J. (2008). Specialized models and ranking for coreference resolution, *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 660-669.

Denis, P., Baldridge, J. (2009). Global joint models for coreference resolution and named entity classification. *Procesamiento del Lenguaje Natural, 42,* 87-96. ISSN: 1135-5948.

Denis, P., Sagot, B. (2010). Exploitation d'une ressource lexicale pour la construction d'un étiqueteur morpho-syntaxique état-de-l'art du français, *Actes de TALN*, Montréal.

Desclés, J.-P., Cartier, E., Jackiewicz, A., Minel, J.-L. (1997). Textual processing and contextual exploration method, *Context*, Universidade Federal do Rio de Janeiro, 189-197.

Di Eugenio, B. (1996). The discourse functions of Italian subjects: a Centering approach, *actes de COLING*, Copenhague, 352-357.

Dik, S. C. (1978). Functional Grammar. Amsterdam: Elsevier.

Dik, S. C. (1997). The Theory of Functional Grammar. Part 1: The Structure of the Clause. 2nd revision, Berlin: Mouton de Gruyter.

Dik, S. C., Hoffmann, M. E., de Long, J. R., Djiang, S. I., Stroomer, H., Devries, L. (1981). On the Typology of Focus Phenomena, *in* Hoekstra, T., van der Hulst, H., Moortgat, M. (eds), *Perspectives on Functional Grammar*, Dordtrecht: Foris Publications, 41-74.

Dimitrov, M., Bontcheva, K., Cunningham, H., et Maynard, D. (2005). "A Light-weight Approach to Coreference Resolution for Named Entities in Text" *in* A. Branco, T. Cenery & R. A. Mitkov (eds) Anaphora Processing: Linguistic, Cognitive and Computational Modelling, John Benjamins.

Doddington, G., Mitchell, A., Przybocki, M., Ramshaw, L., Strassel, S., Weischedel, R. (2004). The automatic content extraction (ACE) program–tasks, data, and

evaluation, *Proceedings of the LREC Conference*, Canary Islands, Spain, July, 837-840.

Dooley, R. A. (2007). Explorations in discourse topicality, *SIL Electronic Working Papers*.

Downey, D., Etzioni, O., Soderland, S. (2005). A probabilistic model of redundancy in information extraction, *Proceedings of IJCAI*, 1034-1041.

Downing, A. (2000). Nominalization and Topic management in leads and headlines. Discourse and Community. Doing Functional Linguistics. Ed. E. Ventola, Tübingen: Gunter Narr Verlag, 355-378.

Downing, A., Neff, J., Carretero, M., Martínez-Caro, E., Pérez de Ayala, S., Marín, J., Simón, J. (1998). Structuring and signalling topic management. *In* S. Embleton (ed.) *LACUS Forum XXIV*. The Linguistic Association of Canada and the United States, 267-278.

Du, L., Buntine, W., Johnson, M. (2013). Topic Segmentation with a Structured Topic Model, *Proceedings of NAACL-HLT*, Atlanta, Georgia, 190-200.

Dubied, A. (2000). Invasion péritextuelle et contaminations médiatiques. Le fait divers, une catégorie complexe ancrée dans le champ journalistique, *Semen, 13*, 51-66.

Dubied, A. (2004). *Les dits et les scènes du fait divers*. Genève : Droz.

Dubied, A., Lits, M. (1999). Le fait divers. Paris : PUF.

Dupont, M. (2002). Une approche cognitive pour le calcul des chaînes de référence, *Actes de TALN*, Nancy.

Dupont, M. (2003). Une approche cognitive du calcul de la référence, thèse de Doctorat, Université de Caen, 314p.

Ehrmann, M. (2008). Les Entités Nommées, de la linguistique au TAL : statut théorique et méthodes de désambiguïsation, thèse de doctorat, Université Paris 7 (Paris Diderot), 283p.

Ehrmann, M., Jacquet, G. (2006). Vers une double annotation des entités nommées, *TAL*, *47* (3), 63-88.

Eisenstein, J. (2009). Hierachical text segmentation from multi-scale lexical cohesion, *Humain Language Technologies: The 2009 annual conference of the Nothern American Chapter of the ACL, Association for computational Linguistics*, 353-361.

Eisenstein, J., Barzilay, R. (2008). Bayesian Unsupervised Topic Segmentation, *Proceedings of EMNLP*, Honolulu, 334-343.

Ekbal, A., Saha, S., Uryupina, O., Poesio, M. (2011). Multiobjective Simulated Annealing Based Approach for Feature Selection in Anaphora Resolution, *8th Discourse Anaphora and Anaphor Resolution Colloquium, DAARC,* Faro, Portugal. Revised Selected Papers, Iris Hendrickx, Sobha Lalitha Devi, António Branco and Ruslan Mitkov (Eds.), *7099*, 47-58.

Elalouf, M.-L. (2006). Thème et thématisation – Présentation, *Linx* [En ligne], *55*, 7-11.

Ellouze, N. (2010). Approche de recherche intelligente fondée sur le modèle des Topic Maps, Application au domaine de la construction durable, Thèse de Doctorat, CNAM, 261p.

Enkvist, **N.** (1978). « Linearity and Text Strategy », *The Nordic Languages and Modern Linguistics*, *3*, 159-172.

Evans, R. (2001). Applying Machine Learning toward an Automatic Classification of it, *Literary and Linguistic Computing*, *16* (1), 45-57.

Fauconnier, G. (1984). *Espaces mentaux*. Paris, Editions de Minuit.

Fellbaum, C. (1998) (ed.). WordNet: An electronic lexical database, Cambridge, MA, The MIT Press, 423p.

Ferret, O. (2006a). Approches endogène et exogène pour améliorer la segmentation thématique de documents, *Traitement Automatique des Langues*, *47* (2), 111-135.

Ferret, O. (2006b). Découvrir les thèmes d'un document pour en améliorer la segmentation thématique, *actes de la 9ème conférence CIDE*, Fribourg, 97-111.

Ferret, O. (2007). Finding document topics for improving topic segmentation, *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, Prague, Czech Republic, 480-487.

Ferret, O. (2009). Utiliser des sens de mots pour la segmentation thématique ?, *Actes de TALN*, session posters, Senlis.

Ferret, O., Grau, B., Masson, N. (1998). Thematic segmentation of texts : two methods for two kinds of texts, *Proceedings of ACL-COLING*, Montréal,, 1, 392-396.

Ferret, O., Grau, B., Minel, J.-L., Porhiel, S. (2001). Repérage de structures thématiques dans des textes, *Actes de TALN*, 163-172.

Filippova, K., Strube, M. (2006). Using linguistically motivated features for paragraph segmentation, *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing,* Sydney, Australia, 267-274.

Finkel, J. R., Manning, C. D. (2008). Enforcing transitivity in coreference resolution, *Proceedings of ACL-HLT*, Columbus, OH, 45-48.

Firbas, J. (1964). On Defining the Theme in Functional Sentence Analysis, *Travaux Linguistiques de Prague*, *1*, 267-280.

Firbas, J. (1966). Non-thematic subjects in contemporary English, *Travaux linguistiques de Prague, 2*, 239-256.

Firbas, J. (1972). On the interplay of prosodic and non-prosodic means of Functional Sentence Perspective. *In* V. Fried (ed.), The Prague School of Linguistics and Language Teaching. London: Oxford University Press.

Firbas, J. (1992). Functional Sentence Perspective, Written and Spoken Communication, Cambridge: Cambridge University Press.

Flament, D. (2006). L'entrée thème/rhème du glossaire de Comenius, *Linx*, *55*, 61-71.

Floor, S. (2004). From information structure, topic and focus, to theme in Biblical Hebrew narrative, Dissertation # 9165 xiv, 360p.

Foucault, N., Rosset, S., Adda, G. (2013). Segmentation textuelle de pages web et sélection de documents pertinents en Questions-Réponses, *Actes de TALN*, Les Sables d'Olonne, 479-492.

Fox, B. (1987), Discourse structure and anaphora. Cambridge: Cambridge University Press.

Fradin, B. (1984). Anaphorisation et stéréotypes nominaux, *Lingua*, *64*, 325-369.

Fragnon, J. (2007). Le fait divers dans la PQR : fenêtre ou miroir sur la violence ? *Les Cahiers du journalisme, 17*, 254-269.

Friburger, N. (2002). Reconnaissance automatique des noms propres : application à la classification automatique de textes journalistiques. Thèse de Doctorat, Université François-Rabelais Tours, France.

Friburger, N., Maurel, D. (2004). Finite-state transducer cascade to extract named entities in texts, *Theoretical Computer Science*, *313*, 94-104.

Galliano, S., Gravier, G., Chaubard, L. (2009). The ESTER 2 evaluation campaign for the rich transcription of French radio broadcasts, *Actes d'Interspeech*, 2583-2586.

Galmiche, M. (1992). Au carrefour des malentendus : le thème, *L'Information grammaticale, 54*, 3-10.

Gardent, C., Manuélian, H. (2005). Création d'un corpus annoté pour le traitement des descriptions définies, *Traitement Automatique des Langues, TAL*, *46* (1), 115-139.

Garrod, S., Sandford, T. (1983). Topic Dependent Effects in Language Processing, *in* The Process of Language Understanding, *ed.* by G. B. Flores d'Arcais, R. Jarvella (Wiley, London).

Ge, N., Hale, J., Charniak, E. (1998). A statistical approach to anaphora resolution, *Proceedings of the Sixth Workshop on Very Large Corpora*, Montreal, Canada, 161-170.

Gegg-Harrison, W., Byron, D. (2004). PYCOT: An Optimality Theory-based Pronoun Resolution Toolkit, *Actes de LREC*, Lisbonne.

Georgescul, M., Clark, A., Armstrong, S. (2006). Word distributions for thematic segmentation in a support vector machine approach, *Proceedings of the 10th Conference on Computational Natural Language Learning, CoNLL*, 101-108.

Gernsbacher, M.A. (1989). Mechanisms that improve referential access, *Cognition*, 32, 99-156.

Givón, T. (1976). Topic, pronoun and grammatical agreement, in Li, C. (ed.), Subject and Topic. New York: Academic Press, 151-188.

Givón, T. (1979). On understanding grammar, New York: Academic Press.

Givón, T. (1983). Topic continuity in discourse: an introduction. In: Talmy Givón (ed.) *Topic continuity in discourse: a quantitative cross-language study.* Amsterdam/Philadelphia: John Benjamins, 5-41.

Givón, T. (1990). Syntax: A Functional Typological Introduction, 2, Amsterdam/Philadelphia: John Benjamins.

Goutsos, D. (1997). Modeling Discourse Topic: sequential relations and strategies in expository text, *Advances in Discourse Processes*, *59*, Norwood: Ablex Publishing Corporation.

Green, S. J. (1996). Using lexical chains to build hypertext links in newspaper articles, *Internet-Based Information Systems, Papers from the AAAI Workshop*, Portland, Oregon, USA, 56-64.

Grishman, R., Sundheim, B. (1996). Message Understanding Conference - 6: A Brief History, *Proceedings of the 16th International Conference on Computational Linguistics* (COLING), Copenhague, 466-471.

Grobet, A. (2002). L'identification des topiques dans les dialogues. De Boeck Université.

Grobet, A. Montemayor-Borsinger, A. (2012). Double éclairage sur l'organisation thématique de discours oraux 'publics', *Actes du CMLF*, Lyon, 545-559.

Grosz, B. J., Joshi, A.K, Weinstein, S. (1995). Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, *21* (2), 203-225.

Grosz, B. J., Sidner, C. L. (1986). Attention, intentions, and the structure of discourse, *Computational Linguistics*, *12* (3), 175-204.

Grouin, C., Galibert, O., Rosset, S., Quintard, L., Zweigenbaum, P. (2011). Mesures d'évaluation pour entités nommées structurées, *Actes de EvalECD'2011 (Évaluation des méthodes d'Extraction de Connaissances dans les Données)*, Brest, 49-62.

Guillot, C. (2006). Anaphores résomptives démonstratives et relations partie/tout en discours. *In* G. Kleiber, C. Schnedecker, A. Theissen (éd.), La relation partie-tout, Louvain-Paris, Peeters (Bibliothèque de l'Information Grammaticale), 289-302.

Guinaudeau, C. (2011). Structuration automatique de flux télévisuels. Thèse de Doctorat, Institut National des Sciences Appliquées de Rennes, 154p.

Gülich, E., Raible, W. (1977). *Linguistische Textmodelle. Grundlagen und Möglichkeiten*, München.

Gundel, J., Hedberg, N. et Zacharski, R. (1993). « Cognitive status and the form of referring expressions », *Language*, *69*, 274-307.

Habert, B., Nazarenko, A. et Salem, A. (1997). Les Linguistiques de corpus. Paris : Armand Colin.

Haghighi, A., Klein, D. (2007). Unsupervised coreference resolution in a nonparametric Bayesian model, *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, 848–855.

Haghighi, A., Klein, D. (2010). Coreference resolution in a modular, entity-centered model, *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (HLT '10), Association for Computational Linguistics, Stroudsburg, PA, USA, 385-393.

Halliday, M. A. K. (1967a). Intonation and Grammar in British English. The Hague: Mouton.

Halliday, M. A. K. (1967b). Notes on transitivity and theme in English, Part II, *Journal of Linguistics*, *3*, 199-244.

Halliday, M. A. K. (1985). An Introduction to Functional Grammar. Arnold.

Halliday, M. A. K., Hasan, R. (1976). Cohesion in English. Longman.

Harman, D. (1993). Overview of the first trec conference, *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval, ACM*, New York, USA, 36-47.

Hartrumpf, S. (2001). Coreference Resolution with Syntactico-Semantic Rules and Corpus Statistics, *Proceedings of the Fifth Computational Natural Language Learning Workshop (CoNLL-2001), 7*, Association for Computational Linguistics, Stroudsburg, PA, USA, 137-144.

Hearst, M. A. (1994). Multi-paragraph segmentation of expository texts, *Proceedings of the 32th Annual Meeting of the Association for Computational Linguistics.*

Hearst, M. A. (1997). TextTiling: Segmenting Text into Multi-paragraphe Subtopic Passages, *Computational Linguistics*, *23* (1), 33-64.

Hearst, M. A. (2006). Clustering versus faceted categories for information exploration, *Communications of the ACM*, 49, 59–61.

Hearst, M. A., Plaunt, C. (1993). Subtopic structuring for full-length document access, *Proceedings of the ACM SIGIR-93 International Conference On Research and Development in Information Retrieval*, 59–68.

Helfman, J. I. (1994). Similarity patterns in language, *IEEE Visual Languages*, 173-175.

Hendrickx, I., Hoste, V. (2009). Coreference Resolution on Blogs and Commented News, *Proceedings of the 7th Discourse Anaphora and Anaphor Resolution Colloquium on Anaphora Processing and Applications* (DAARC), Sobha Lalitha Devi, Antonio Branco, and Ruslan Mitkov (Eds.). Springer-Verlag, Berlin, Heidelberg, 43-53.

Hernandez, N. (2004). Description et Détection Automatique de Structures de Texte, Thèse de doctorat, Université Paris-Sud XI.

Hernandez, N., Boudin, F. (2013). Construction d'un large corpus écrit libre annoté morpho-syntaxiquement en français, *Actes de TALN*, 17-21 juin, Les Sables d'Olonne, 160-173.

Hernandez, N., Grau, B. (2002). Analyse thématique du discours : segmentation, structuration, description et représentation, *Actes du 5ème colloque international sur le document électronique (CIDE)*, Hammamet, Tunisie, 277-288.

Hernandez, N., Grau, B. (2003a). Extraction et typage de termes significatifs pour la description de textes, *Proceedings of the International society for knowledge organization (ISKO)*, Grenoble, 3 et 4 juillet, 61-71.

Hernandez, N., Grau, B. (2003b). What is this text about ?, *Proceedings of ACM SIGDOC*, San Francisco, USA.

Herslund, M. (1996). Partitivité et possession inaliénable. La relation d'appartenance. *Faits de Langue, 7*, 33-42.

Hinds, J. (1977). Paragraph structure and pronominalization, *Papers in Linguistics 10*, 77–99.

Hirst, H., St-Onge, D. (1998). Lexical Chains as Representations of Context for the Detection and Correction of Malapropisms. In: Fellbaum, C. (ed.). WordNet: An electronic lexical database*, Cambridge, MA,* The MIT Press, 305-332.

Hobbs, J. (1978). Resolving Pronoun References, *Lingua*, *44*, 311-338.

Ho-Dac, L.-M. (1999). Au carrefour des cadres de discours et des registres, mémoire de maîtrise, Université de Toulouse Le Mirail.

Ho-Dac, L.-M. (2003). Cadres de discours et chaînes de référence, étude en corpus, séminaire ERSS.

Ho-Dac, L.-M. (2007). La position initiale dans l'organisation du discours : une exploration en corpus. Thèse de Doctorat, Université de Toulouse le Mirail, 362p.

Ho-Dac, L-M., Péry-Woodley, M.-P. (2009). A data-driven study of temporal adverbials as discourse segmentation markers. *Discours* (revue en ligne), *4*.

Hoey, M. (1991). Patterns of lexis in text. Oxford University Press: Oxford.

Hoey, M. (2005). Lexical Priming: A New Theory of Words and Language. Routledge, 202p.

Hofmann, T. (1999). Probabilistic latent semantic indexing, Proceedings of SIGIR, Berkeley, CA, 35-44.

Hofmann, T.R. (1989). Paragraphs, & anaphora, *Journal of Pragmatics*, *13*, 239-250.

Hollingsworth, W., Teufel, S. (2005). Human annotation of lexical chains: Coverage and agreement measures, *Proceedings of the Workshop ELECTRA: Methodologies and Evaluation of Lexical Cohesion Techniques in Real-world Applications, In Association with SIGIR'05*, Salvador, Brazil.

Hoste, V. (2005). Optimization Issues in Machine Learningof Coreference Resolution, thèse de Doctorat, Antwerp University, 261p.

Huet, S., Gravier, G., Sébillot, P. (2008). Un modèle multi-sources pour la segmentation en sujets de journaux radiophoniques, *Actes de TALN'08*, Avignon.

Hurault-Plantet, M., Jardino, M., Berthelin, J.-B. (2006). Ajustement des frontières de segments thématiques détectés automatiquement, *Actes de DEFT'06*, 21 et 22 septembre, Fribourg, Suisse.

Ide, N., Véronis, J. (1994). MULTEXT (Multilingual Tools and Corpora), *Actes de la 14ème conference IAACL*, Kyoto.

Iida, R., Inui, K., Takamura, H., Matsumoto, Y. (2003). Incorporating contextual cues in trainable models for coreference resolution, *Proceedings of the EACL Workshop on The Computational Treatment of Anaphora*, 23-30.

Illouz, G., Habert, B., Folch, H., Fleury, S., Heiden, S., Lafon, P., Prévost, S. (2000). TyPTex: Generic features for Text Profiler, *Content-Based Multimedia Information Access*, *2*, 1526-1540.

Intuition http://www.sinequa.com/html-fr/fr-edition.oem.html

Ion, R. (2007). TTL: A portable framework for tokenization, tagging and lemmatization of large corpora, Bucharest: Romanian Academy.

Jackiewicz, A. (2002). Repérage et délimitation des cadres organisationnels pour la segmentation automatique des textes, *actes de CIFT'02*, Hammamet, Tunisie, 95-107.

Jacquemin, C. (1997). Variation terminologique : reconnaissance et aquisition automatiques de termes et de leurs variantes en corpus. Mémoire d'habilitation à diriger des recherches en informatique fondamentale, Université de Nantes, Nantes.

Jayarjan, D., Deodhare, D., Ravindran, B. (2008). Lexical chains as document features, *Proceedings of The third International Joint Conference on Natural Language Processing*, Hyderabad, India, 111-117.

Jenkins, C. (2002). Les procédés référentiels dans les portraits journalistiques, *Actes du 15ème congrès Skandinaviske romanistkongress*, Oslo.

Johnsen, L. A. (2010). Anaphore et recatégorisation des objets-de-discours en français parlé, in Florea L. *et al.* (éds.), Directions actuelles en linguistique du texte. *Actes du colloque international Le texte : modèles, méthodes, perspectives*, Cluj-Napoca, Casa Cărţii de Ştiinţă, 2, 23-31.

Jones, L. K. (1977). Theme in English Expository Discourse. Lake Bluff, IL: Jupiter Press.

Kan, M.-Y., Klavans, J. L., Mckeown, K. R. (1998). Linear segmentation and segment significance, *Proceedings of the 6th International Workshop of Very Large Corpora (WVLC-6)*, 197-205.

Karttunen, L. (1976). Discourse referents, *in* J. D. McCawley (Dir.) *Syntax and Semantics. Notes from the Linguistic Underground*, Academic Press, New York, 363-385.

Keenan, E., Schieffelin, B. (1976). Topic as a discourse notion: A study of topic in the conversations of children and adults, *in* Subject and topic, ed. by C. Li. New York: Academic Press.

Kennedy, C., Boguraev, B. (1996). Anaphora for everyone: pronominal anaphora resolution without a parser, *Proceedings of the 16th conference on Computational linguistics - Volume 1* (COLING '96), Stroudsburg, PA, USA, 113-118.

Khalis, Z. (2006). La segmentation thématique. Application à la campagne DEFT'2006, Mémoire de Master, Université Joseph Fourier, Grenoble, 66p.

Kintsch, W. (2002). On the notions of theme and topic in psychological process models of text comprehension. In M. Louwerse & W. van Peer (Eds.) Thematics: Interdisciplinary Studies, Amsterdam, Benjamins, 157-170.

Kintsch, W., Van Djik, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review*, *85* (5), 363-394.

Kleiber, G. (1981). Problèmes de référence. Descriptions définies et noms propres. Paris, Klincksieck.

Kleiber, G. (1986). Pour une explication du paradoxe de la reprise immédiate, *Langue Française*, *12*, 54-79.

Kleiber, G. (1989). Référence, texte et embrayeurs, *Semen*, *4*, 13-50.

Kleiber, G. (1990). Quand *il* n'a pas d'antécédent, *Langage*, *97*, 24-50.

Kleiber, G. (1992). Cap sur les topiques avec le pronom il, *L'Information grammaticale, 54*, 15-25.

Kleiber, G. (1994). Anaphores et Pronoms. Louvain-la-Neuve : Duculot, 229p.

Kleiber, G. (1997a). Des anaphores associatives méronymiques aux anaphores associatives locatives, *Verbum*, *19*, 1-2, 25-66.

Kleiber, G. (1997b). Les anaphores associatives actancielles, *Scolia*, *10*, 89-120.

Kleiber, G. (2000). Typologie des anaphores associatives : le cas des anaphores associatives fonctionnelles », in Englebert, A., Pierrard, M., Rosier, L., Van Raemdonck, D. (éds) ; *Actes du XXIIe Congrès International de Linguistique et philologie romane*, Bruxelles, 7, Sens et fonctions, Tübingen, Niemeyer, 335-342.

Kleiber, G. (2001). L'anaphore associative. Paris : PUF.

Kleiber, G. (2002). Marqueurs référentiels et théorie du centrage, *Linx*, *47*, Université Marc Bloch.

Kleiber, G. (2003). Un « puzzle » référentiel en anaphore associative, *in* Fonseca, F.I. & Brito, A. M. (éds), Lingua portuguesa : estruturas, usos e contrastes, Porto, Centro Linguistica da Universidade do Porto, 97-110.

Kleiber, G., Schnedecker, C., Charolles, M., David, J. (éds). (1994). L'anaphore associative. Aspects linguistiques, psycholinguistiques et automatiques. Paris : Klincksieck, 343p.

Kolla, M. (2002). Automatic text summarization using lexical chains : algorithms and experiments, Thèse de Doctorat, Université de Lethbridge, Canada, 89p.

Kripke, S. (1972). Naming and Necessity, *in* Davidson, D., Harman, G. (dir.), Semantics of Natural Language. Dordrecht, Reidel, 253-355. [Trad. Française : 1982, La Logique des noms propres. Paris, Minuit]

Labadié, A. (2008). Segmentation thématique de texte linéaire et non-supervisée : Détection active et passive des frontières thématiques en Français. Thèse de Doctorat, Université Montpellier 2, 184p.

Labadié, A., Chauché, J. (2007). Segmentation thématique par calcul de distance thématique, *Actes des 7èmes journées francophones Extraction et Gestion des Connaissances*, 355-366.

Labadié, A., Prince, V. (2008a). Comparaison de méthodes lexicales et syntaxico-sémantiques dans la segmentation thématique de texte non supervisée, *Actes de TALN*, Avignon.

Labadié, A., Prince, V. (2008b). Intended boundaries detection in topic change tracking for text segmentation, *International Journal of Speech Technology*, *11* (3-4),167-180.

Labadié, A., Prince, V. (2008c). Finding text boundaries and finding topic boundaries: two different tasks ? *in* Advances in Natural Language Processing Lecture Notes in Computer Science, *5221*, 260-271.

Labadié, A., Prince, V. (2008d). Lexical and Semantic Methods in Inner Text Topic Segmentation: A Comparison between C99 and Transeg, *Proceedings of the 13th international conference on Natural Language and Information Systems: Applications of Natural Language to Information Systems*, edited by E. Kapetanios, V. Sugumaran, and M. Spiliopoulou, *5039*, Springer-Verlag, Berlin, Heidelberg, 347-349.

Labadié, A., Prince, V. (2008e). The impact of corpus quality and type on topic based text segmentation evaluation, *Proceedings of the International Multiconference on Computer Science and Information Technology*, *3*, 313-319.

Lafferty, J., McCallum, A., Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data, *Actes de ICML,* 282-289.

Laignelet, M. (2009). Analyse discursive pour le repérage automatique de segments obsolescents dans des documents encyclopédiques, thèse de Doctorat Université de Toulouse Le Mirail, 326p.

Lambrecht, K. (1994). Information structure and sentence form: Topic, focus, and the mental representation of discourse referents, Cambridge Studies in Linguistics, 71, Cambridge: Cambridge University Press.

Lambrecht, K., Michaelis, L.A. (1998). Silence accent in information questions: default and projection. *Linguistics and Philosophy, 21*, 477-544.

Lamprier, S., Amghar, T., Levrat, B., Saubion, F. (2008). Using an Evolving Thematic Clustering in a Text Segmentation Process, *Universal Computer Science*, 178-192.

Landauer, T.K., Foltz, P.W., Laham, D. (1998). Introduction to Latent Semantic Analysis, *Discourse Processes, 25*, 259-284.

Landragin, F. (2003). La saillance comme point de départ pour l'interprétation et la génération, actes de la *Journée d'étude de l'Association pour le Traitement Automatique des Langues sur la structure informationnelle*, Paris.

Landragin, F. (2004). L'utilisation de scores numériques en sémantique computationnelle, *Actes des Journées scientifiques de sémantique et Modélisation (JSM'04)*, Lyon.

Landragin, F. (2005). Traitement automatique de la saillance, *Actes de TALN*, Dourdan, 6-10 juin.

Landragin, F. (2007). Saillance. In: Godard, D., Roussarie, L. & Corblin, F. (Eds.), Dictionnaire de sémantique, GdR Sémantique et Modélisation, CNRS, http://www.semantique-gdr.net/dico.

Landragin, F. (2011). Une procédure d'analyse et d'annotation des chaînes de coréférence dans des textes écrits, *Corpus 10*, 61-80.

Landragin, F. (2012). La saillance : questions méthodologiques autour d'une notion multifactorielle, *Faits de Langues, 39*, Peter Lang, Berne, 15-31.

Lang, J., Qin, B., Liu, T., Li, S. (2009). Unsupervised Coreference Resolution with HyperGraph Partitioning, *Computer and Information Science*, 2 (4), 55-63.

Lappin, S., Leass, H. J. (1994). An algorithm for pronominal anaphora resolution, *Computational Linguistics, 20* (4), 535-561.

Lassalle, E., Denis, P. (2013). Apprentissage d'une hiérarchie de modèles à paires spécialisées pour la résolution de la coréférence, *Actes de TALN*, 17-21 juin, Les Sables d'Olonne, 118-131.

Lee, H., Chang, A., Peirsman, Y., Chambers, N., Surdeanu, M., and Jurafsky, D. (2013). Deterministic Coreference Resolution Based on Entity-Centric, Precision-Ranked Rules, *Computational Linguistics, 39,* (4), uncorrected proof.

Legallois, D. (2004). Cohésion lexicale et réseaux phrastiques dans la construction du texte expositif. In Porhiel, S., Klinger, D. (Eds.), L'unité texte : Association perspectives.

Legallois, D. (2006). Des phrases entre elles à l'unité réticulaire du texte, *Langages*, *163*, 56-70.

Legallois, D. (2011). Système, norme, Parole : application au lexique, au texte, à la phrase et aux constructions grammaticales, mémoire d'habilitation à diriger les recherches, Université de Caen.

Le Pesant, D. (2008). « Des descriptions linguistiques, des dictionnaires électroniques et un analyseur syntaxique pour la résolution des anaphores non pronominales en français ». *Cahiers du Cental, 5*, 187-202, *in* Constant, M. Dister, A., Emirkanian, L., Piron, S. (dir.). Louvain : Presses Universitaires de Louvain.

Leroy, S. (2001). Entre identification et catégorisation, l'antonomase du nom propre en français, Thèse de Doctorat, Université Montpellier III.

Levelt, W. J. M. (1989). Speaking: From intention to articulation. Cambridge, Mass.: MIT Press.

Lewis, D. D. (1992). An Evaluation of Phrasal and Clustered Representations on a Text Categorization Task. *In Proceedings of the Annual International A CM SIGIR Conference on Research and Development in Information Retrieval*, 37-50.

Lin, C-Y., Hovy, E. (1997). Identifying topics by position, *Actes de la Fifth Conference on Applied Natural Language Processing* (ANLP-97).

Litman, D. J., Passonneau, R. J. (1995). Combining multiple knowledge sources for discourse segmentation, *Proceedings of ACL*, 108-115.

Longacre, R. E. (1979). The paragraph as a grammatical unit, Syntax and Semantics, (Discourse and Syntax), T. Givón (éd.), New York, Academic Press, 115-134.

Longo, L. Todirascu, A. (2009). Une étude de corpus pour la détection automatique des thèmes, *actes des 6èmes journées de linguistique de corpus,* 10-12 septembre, Lorient.

Ludovic, J.-L. (2011). Approches supervisées et faiblement supervisées pour l'extraction d'événements complexes et le peuplement de bases de connaissances, Thèse de Doctorat, Université Paris 11 - Paris Sud, 192p.

Luo, X. (2005). On coreference resolution performance metrics, *Proceedings of HLT-EMNLP*, Vancouver, Canada, 25-32.

Luo, X., Ittycheriah, A., Jing, H., Kambhatla, N., Roukos, S. (2004). A mention-synchronous coreference resolution algorithm based on the Bell tree, *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, Stroudsburg, PA, USA.

Luquet, G. (1994). Remarques sur la structure des suffixes formateurs de noms d'agent et d'instruments en espagnol. *Recherches en linguistique hispanique*, *22*, 339-348.

Lyngfelt, B. (2000). « Optimality Theory Semantics and Control », Rutgers Optimality Archive. Disponible sur <http://roa.rutgers.edu/files/411-0800/roa-411-lyngfelt-3.pdf> consulté le 20/07/2012.

Lyons, J. (1970). Linguistique générale. Introduction à la linguistique théorique, F. Dubois-Charlier et D. Robinson (trad.), Paris, Librairie Larousse.

Maisonnasse, L., Tambellini, C. (2005). Dépendances syntaxiques et méthodes de détection de passages pour une segmentation sur le locuteur et le thème, *atelier DEFT, TALN*, Dourdan.

Makhoul, J., Kubala, F., Schwartz, R., Weischedel, R. (1999). Performance measures for information extraction, *Proceedings of the DARPA Broadcast News Workshop*, Virginie, USA, 249-252.

Manning, C. D., Raghavan, P., Schütze, H. (2008). *Introduction to Information Retrieval*, Cambridge University Press, 544p.

Manuélian, H. (2003). Descriptions définies et démonstratives : analyses de corpus pour la génération de textes, Thèse de doctorat de l'Université de Nancy 2, 200p.

Marandin, J. -M. (1988). À propos de la notion de thème en discours. Eléments d'analyse dans le récit, *Langue Française*, *78*, 67-87.

Marandin, J. -M. (2007). Thème (topic) de discours. *In* D. Godard, L. Roussarie et F. Corblin (éd.), *Sémanticlopédie: dictionnaire de sémantique*, GDR Sémantique & Modélisation, CNRS, http://www.semantique-gdr.net/dico/.

Marcus, M. P., Marcinkiewicz, M. A., Santorini, B. (1993). Building a large annotated corpus of English: the penn treebank. *Computational Linguistics*, *19* (2), 313-330.

Martin, R. (1983). La logique du sens. Paris, PUF.

Martin, R. (2006). « Définir l'indéfinition », in *Indéfini et prédication*, Corblin *et al.* (dir.). Paris : PUS, 11-24.

Masson, N. (1995). An automatic method for document structuring, *Proceedings of the 18th ACM-SIGIR*, Seattle, USA, 372-373.

Mathesius, V. (1942). O soustavném rozboru gramatickém. [About systematic grammatical analysis], Slovo a slovesnost, 88-92.

Mathesius, V. (1975). A Functional analysis of Present Day English on a General Linguistic Basis, The Hague, Mouton, 228p.

Maynard, D., Tablan, V., Cunningham, H., Ursu, C., Saggion, H., Bontcheva., K., Wilks, Y. (2002). Architectural Elements of Language Engineering Robustness, *Journal of Natural Language Engineering – Special Issue on Robust Methods in Analysis of Natural Language Data*, *8*, 257–274.

McCallum, A., Wellner, B. (2003). Toward Conditional Models of Identity Uncertainty with Application to Proper Noun Coreference, *Computer Science Department Faculty Publication Series.*

McCallum, A., Li, W. (2003). Early Results for Named Entity Recognition with Conditional Random Fields, Feature Induction and Web-Enhanced Lexicons, *Proceedings of the Seventh Conference on Natural Language Learning (CoNLL).*

McCarthy, J. F. (1997). A Trainable Approach To Coreference Resolution For Information Extraction, thèse de Doctorat, Université du Massachusetts, 198p.

McCarthy, J. F., Lehnert, W. G. (1995). Using decision trees for coreference resolution, *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, Montréal, Canada, 1050-1055.

McCord, B. A., Lappin, S., Zadrozny, W. (1992). Natural Language Processing within a Slot Grammar Framework, *International Journal on Artificial Intelligence Tools*, *1*, 229-277.

McDonald, D. D. (1996). Internal and External Evidence in the Identification and Semantic Categorisation of Proper Names, *in* Boguraev, B. and J. Pustejovsky (eds.) Corpus Processing for Lexical Acquisition, Cambridge, MIT, 32-43.

Miller, G. (1995). WordNet: A Lexical Database for English. *Communications of the ACM*, *38* (11), 39-41.

Milner, J.-C. (1982). Ordres et raisons de langue. Paris : Seuil.

Minel, J.-L., Descles, J.-P., Cartier, E., Crispino, G., Ben Hazez, S., Jackiewicz, A. (2001). Résumé automatique par filtrage sémantique d'informations dans des textes, Présentation de la plate-forme FilText, *Technique et Science Informatiques*, *3*, Paris : Hermès.

Mitkov, R. (1998). Robust pronoun resolution with limited knowledge, *Proceedings of the 18th International Conference on Computational Linguistics (COLING'98)/ACL'98*, Montréal, Canada, 867-875.

Mitkov, R. (1999). Anaphora resolution: the state of the art, Working paper, (Based on the COLING'98/ACL'98 tutorial on anaphora resolution), University of Wolverhampton, Wolverhampton.

Mitkov, R. (2000). Towards a more consistent and comprehensive evaluation of anaphora resolution algorithms and systems, *Proceedings of the Discourse Anaphora and Reference Resolution Conference (DAARC2000)*, 96-107.

Mitkov, R. (2001). *Outstanding issues in anaphora resolution*. In Al. Gelbukh (Ed.). Computational Linguistics and Intelligent Text Processing, 110-125.

Mitkov, R., Evans, R., Orasan, C., Dornescu, I., and Rios, M. (2012). Coreference Resolution: To What Extent Does It Help NLP Applications ? Text, *Speech and Dialogue. Lecture Notes in Computer Science*, *7499*, 16-27.

Mondada, L. (1994). Verbalisation de l'espace et fabrication du savoir : Approche inguistique de la construction des objets de discours, Lausanne : Université de Lausanne, 670p.

Morris, J., Hirst, G. (1991). Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text, *Computational Linguistics*, *17* (1), 21-42.

Muzerelle, J., Schang, E., Antoine, J.-Y., Eshkol, I., Maurel, D., Boyer, A., Nouvel, D. (2012). Annotations en chaînes de coréférences et anaphores dans un corpus de discours spontané en français, *Actes du 3ème Congrès Mondial de Linguistique Française (CMLF)*, 2497-2516.

Muzerelle, J., Lefeuvre, A., Antoine, J.-Y., Schang, E., Maurel, D., Villaneau, J., Eshkol, I. (2013). ANCOR, premier corpus de français parlé d'envergure annoté en coréférence et distribué librement, *Actes de TALN*, 555-563.

Namer, F. (2000). "Flemm : Un analyseur Flexionnel du Français à base de règles", Traitement automatique des langues pour la recherche d'information, *revue T.A.L*, (Ch. Jacquemin éd.), Paris.

Nand, P. (2012). Resolving Co-reference Anaphora Using Semantic Constraints, thèse de Doctorat, AUT University, 233p.

Ng, V. (2005). Supervised ranking for pronoun resolution: some recent improvements, *Proceedings of AAAI*, 1081-1086.

Ng, V. (2008). Unsupervised Models for Coreference Resolution, *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, 640-649.

Ng, V. (2010). Supervised Noun Phrase Coreference Research: The First Fifteen Years, *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 1396-1411.

Ng V., Cardie C. (2002). Improving machine learning approaches to coreference resolution, *Proceedings of ACL (Association For Computational Linguistics),* Morristown, 104-111.

Nicolov, N., Salvetti, F., Ivanova, S. (2008). Sentiment analysis: Does coreference matter ?, *Proceedings of the Symposium on Affective Language in Human and Machine*, Aberdeen, UK.

NIST. (2004). The ace evaluation plan. www.nist.gov/speech/tests/ace/index.htm [consulté le 16/04/13].

Nøklestad, A. (2009). A Machine Learning Approach to Anaphora Resolution Including Named Entity Recognition, PP Attachment Disambiguation, and Animacy Detection, PhD Thesis, University of Oslo, 295p.

Nølke, H., (1997). « Anaphoricité et focalisation : le cas du pronom personnel disjoint », in De Mulder W., Tasmowski-De Ryck L., Vetters C., *Relations anaphoriques et (in)cohérence*, Amsterdam, Rodopi, 55-66.

Nomoto, T., Matsumoto, Y. (1996). Exploiting Text Structure for Topic Identification, *Proceedings of the 4th Workshop on Very Large Corpora*, 101-112.

Nouioua, F. (2007). Heuristique pour la résolution d'anaphores dans les textes d'accidents de la route, *Actes de la Journée d'étude de l'Association pour le Traitement Automatique des LAngues (ATALA) sur La résolution des anaphores en Traitement Automatique des Langues*, Paris.

Oliveira Santos, C. S. (2006). ALEXIA - Acquisition of Lexical Chains for Text Summarization, thèse de Doctorat, University of Beira Interior, Portugal, 115p.

Passerault, J.-M., Chenet, D. (1991). Le marquage des paragraphes : son rôle dans la gestion des traitements pendant la lecture, *Psychologie française*, *36*, 159-165.

Passonneau, R. J., Litman, D. J. (1993). Intention-based segmentation: human reliability and correlation with linguistic cues, *Proceedings of the Association for Computational Linguistics*, 148–155.

Passonneau, R. J., Litman, D. J. (1997). Discourse Segmentation by Human and Automated Means, *Computational Linguistics*, *23*, 103-139.

Péry-Woodley, M-P. (2000). Une pragmatique à fleur de texte : approche en corpus de l'organisation textuelle. Mémoire d'HDR, *Carnets de grammaire N°8 (juillet 2000)*, Université de Toulouse-LeMirail : ERSS, 164 p.

Péry-Woodley, M.-P. (2001). Modes d'organisation et de signalisation dans des textes procéduraux, *Langages, 141*, 28-46.

Péry-Woodley, M.-P., Asher, N., Enjalbert, P., Benamara, F., Bras, M., Fabre, C., Ferrari, S., Ho-Dac, L.-M., Le Draoulec, A., Mathet, Y., Muller, P., Prévot, L., Rebeyrolle, J., Tanguy, L., Vergez-Couret, M., Vieu, L., Widlöcher, A. (2009). ANNODIS : une approche outillée de l'annotation de structures discursives. In *Actes de la 16ème Conférence Traitement Automatique des Langues Naturelles (TALN'09)*, session poster, Senlis, France.

Péry-Woodley, M.-P., Afantenos, S. D., Ho-Dac, L.-M., Asher, N. (2011). La ressource ANNODIS, un corpus enrichi d'annotations discursives, *TAL*, *52* (3), 71-101.

Piérard, S., Degand, L., Bestgen, Y. (2004). Vers une recherche automatique des marqueurs de la segmentation du discours. In G. Purnelle, C. Fairon, & A. Dister (Eds.), *Actes des 7èmes Journées internationales d'Analyse statistique des Données Textuelles*, Louvain-la-Neuve : Presses universitaires de Louvain, 859-864.

Piérard, S., Bestgen, Y. (2005a). Deux indices pour l'étude des marqueurs de la continuité thématique dans de grands corpus, *Actes de JCL*, 11-119.

Piérard, S., Bestgen, Y. (2005b). Identification automatique des marqueurs globaux du discours par l'analyse des expressions récurrentes, *Phraseology*, Louvain-la-Neuve.

Piérard, S., Bestgen, Y. (2006a). Adverbiaux temporels et expressions référentielles comme marqueurs de segmentation : emploi simultané ou exclusif ?, *Schedae*, prépublication n°3, fascicule n°1, 23-28.

Piérard, S., Bestgen, Y. (2006b). Validation d'une méthodologie pour l'étude des marqueurs de la segmentation dans un grand corpus de textes, *TAL*, *47* (2).

Piérard, S., Bestgen, Y. (2007). Deux indices pour l'étude des marqueurs de la continuité thématique dans de grands corpus. *Actes des Troisièmes journées de la linguistique de corpus*, Williams Geoffrey (ed.), Lorient, 111-119.

Pimm, C. (2008). *Plus-value linguistique* pour la segmentation automatique de texte, *in* Constant, M. et al. (Eds.). Description linguistique pour le traitement

automatique du français, *Cahiers du Cental*, 5, UCL, Presses Universitaires de Louvain, 203-219.

Pincemin, B. (1999). Sémantique interprétative et analyses automatiques de textes : que deviennent les sèmes ?, *Sémiotiques*, *17*, 71-120.

Poesio, M., Ishikawa, T., Schulte im Walde, S., Vieira, R. (2002). Aquiring Lexical Knowledge for Anaphora Resolution, *Proceedings of LREC*, Las Palmas, Spain, 1220-1224.

Poesio, M., Ponzetto, S. P., Versley, Y. (2010). Computational Models of Anaphora Resolution: A Survey. Disponible sur < http://wwwusers.di.uniroma1.it/~ponzetto/pubs/poesio10a.pdf>. (Consulté le 25/06/13)

Ponzetto, S. P., Strube, M. (2006). Exploiting semantic role labeling, WordNet and Wikipedia for coreference resolution, *Proceedings of HLT-NAACL*, New York, 192-199.

Popescu-Belis, A. (1999). Modélisation multi-agent des échanges langagiers : application au problème de la référence et son évaluation, Thèse de Doctorat, Université de Paris XI, 343p.

Popescu-Belis, A. (2000). Évaluation numérique de la résolution de la référence : critiques et propositions, *T.A.L. : Traitement automatique de la langue*, *40* (2), 117–146.

Popescu-Belis, A., Robba, I. (1998). Evaluation of Coreference Rules on Complex Narrative Texts, *Proceedings of DAARC2 (Discourse Anaphora and Anaphor Resolution Colloquium)*, Lancaster, UK, 178-185.

Popescu-Belis, A., Robba, I., Sabah, G. (1998). Reference Resolution Beyond Coreference: a Conceptual Frame and its Application, *Actes COLING-ACL'98*, Montréal, Canada, 1046-1052.

Poon, H, Domingos, P. (2008). Joint Unsupervised Coreference Resolution with Markov Logic, *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (EMNLP), 650-659.

Porhiel, S. (1998). Les indicateurs d'intérêt. Thèse de Doctorat, Université Paris 13. Presses du Septentrion, 450p.

Porhiel, S. (2001). Linguistic expressions as a tool to extract thematic information, *Corpus Linguistic 2001*, Lancaster University.

Porhiel, S. (2004). Les introducteurs de cadres thématiques, *Cahiers de Lexicologie*, *85* (2), 9-45.

Porhiel, S. (2005a). Les marqueurs de thématisation : des thèmes phrastiques et textuels, *Travaux de linguistique*, *51* (2).

Porhiel, S. (2005b). Les séquences thématiques, *Langue Française, 148*, 111-126.

Porhiel, S. (2006). Le détachement en position initiale: rôle phrastique ou discursif/textuel ? Exemple du syntagme à propos de X, *Linguistik online*, *26* (1), 99-126.

Pottier, B. (1964). Vers une sémantique moderne, *Travaux De Linguistique Et De Philologie*, T.II, Université de Strasbourg.

Poudat, C. (2004). Recension et présentation comparative d'étiqueteurs pour le français et l'anglais. *Texto!* [en ligne], *9* (4). Disponible sur : <http://www.revue-texto.net/Corpus/Publications/Poudat_Taggers.html>. (Consulté le 10/10/2013).

Pradhan, S., Ramshaw, L., Marcus, M., Palmer, M., Weischedel, R., Xue, N. (2011). CoNLL-2011 Shared Task: Modeling Unrestricted Coreference in OntoNotes, *Proceedings of the 15th Conference on Computational Natural Language Learning: Shared Task*, Portland, Oregon, 23-24 June, 1-27.

Pradhan, S., Moschitti, A., Xue, N., Uryupina, O., Zhang, Y. (2012). Conll-2012 shared task: Modeling multilingual unrestricted coreference in ontonotes, *Joint Conference on EMNLP and CoNLL - Shared Task*, Jeju Island, Korea, Association for Computational Linguistics, 1-40.

Prévost, S. (1998). La notion de Thème : flou terminologique et conceptuel, *Cahiers de praxématique*, *30*, 13-35.

Prévost, S. (2001). La postposition du sujet en français aux XVe et XVIe siècles. Analyse sémantico-pragmatique. Paris : CNRS éditions, 325p.

Prévost, S. (2003). Les compléments spatiaux : du topique au focus en passant par les cadres, *Travaux de linguistique, 47*, 51-78.

Prince, A., Smolensky, P. (1993). Optimality Theory: Constraint Interaction in Generative Grammar. Rutgers University Center for Cognitive Science Technical Report 2.

Prince, E. F. (1981). Toward a Taxonomy of Given-New Information, *In* Cole, P., (éd). Radical Pragmatics, Academic Press, New York, 223-255.

Prince, V., Labadié, A. (2007). Text Segmentation Based on Document Understanding for Information Retrieval, *Proceedings of NLDB'07*, 295-304.

Qiu, L., Kan, M. Y., Chua, T. S. (2004). A public reference implementation of the RAP anaphora resolution algorithm, *ArXiv Computer Science e-prints*, 291-294.

Rahman, A., Ng, V. (2009). Supervised Models for Coreference Resolution, *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, 968-977.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, *66* (336), 846-850.

Rastier, F. (1996). La sémantique des thèmes ou le voyage sentimental, revue *Texto.*

Reboul A., Balkanski C., Briffault X., Gaiffe B., Popescu-Belis A., Robba I., Romary L., Sabah G. (1997). *Le projet CERVICAL : Représentations mentales, référence aux objets et aux événements,* Rapport interne, Loria-CNRS/Limsi, France.

Reboul, A., Gaiffe, B. (1999). Représentations mentales et reference. Disponible sur : http://hal.archives-ouvertes.fr/docs/00/02/91/08/PDF/Reboul-Gaiffe.pdf. (Consulté le 21/08/13)

Recasens, M. (2010). *Coreference: Theory, Annotation, Resolution and Evaluation.* PhD Thesis. University of Barcelona, 259p.

Recasens, M., Màrquez, L., Sapena, E, Martí, M.A., Taulé, M., Hoste, V., Poesio, M., Versley, Y. (2010). SemEval-2010 Task 1: Coreference Resolution in Multiple Languages, *Proceedings of the ACL International Workshop on Semantic Evaluation* (SemEval-2010), 1-8, Uppsala, Sweden.

Recasens, M., Hovy, E. (2011). BLANC: Implementing the Rand Index for coreference evaluation. *Natural Language Engineering*, *17*(4), 485-510.

Reichler-Béguelin, M.-J. (1989). Anaphores, connecteurs, et processus inférentiels, *in* C. Rubattel (éd.). *Modèles du discours. Recherches actuelles en Suisse Romande*, Berne : P. Lang, 302-336.

Reinhart, T. (1981). Pragmatics and linguistics: An analysis of sentence topics, *Philosophica*, *27*, 53-94.

Reinhart, T. (1982). Pragmatics and linguistics: An analysis of sentence topics. Bloomington: Indiana University Linguistics Club.

Retoré, C., Danlos, L., Moot, R., Prost, J.-P., Van de Cruys, T. (2013). Présentation de l'atelier Mixeur, *Actes de TALN*, Les sables d'Olonne, France.

Reynar, J.-C. (1994). An automatic method of finding topic boundaries, *Proceedings of the Student Session of the 32nd Annual Meeting of the Association for Computational Linguistics*, 331-333.

Reynar, J.-C., (1998). Topic Segmentation: Algorithms and Applications. Thèse de doctorat, University of Pennsylvania, 187p.

Richard, E. (2000). La répétition : syntaxe et interpretation. Thèse de doctorat, Université de Bretagne occidentale, 349p.

Riedl, M., Biemann, C. (2012). How text segmentation algorithms gain from topic models, *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 553–557.

Riegel, M. (1985). L'adjectif attribut. Paris, PUF.

Riegel, M., Pellat, J.-C., RIOUL, R. (2002). Grammaire méthodique du français. Paris : PUF.

Rimmon-Kenan, S. (1985). Qu'est-ce qu'un thème ?, *Poétique*, 397-406.

Roget, P. (1977). Roget's International Thesaurus, Fouth Edition, Harper and Row Publishers Inc.

Rossignol, M. (2005). Acquisition sur corpus d'informations lexicales fondées sur la sémantique différentielle, thèse de Doctorat, Université de Rennes 1, 220p.

Rossignol, M., Sébillot, P. (2003). Extraction statistique sur corpus de classes de mots-clés Thématiques, *TAL (Traitement automatique des langues)*, *44* (3): 217-246.

Royauté, J. (1999). Les groupes nominaux complexes et leurs propriétés : application à l'analyse de l'information. Thèse de Doctorat, Université Henri Poincaré Nancy I.

Salmon-Alt, S. (2001). Référence et Dialogue finalisé : de la linguistique à un modèle opérationnel, Thèse de Doctorat, Université H. Poincaré, Nancy, 270p.

Salmon-Alt, S. (2002). Le projet ANANAS : Annotation Anaphorique pour l'Analyse Sémantique de Corpus, *Actes de TALN*, Nancy, 24-27 juin, 163-172.

Salmon-Alt, S. (2004). Résolution automatique d'anaphores infidèles en français : Quelles ressources pour quels apports ?, *Actes de TALN*, session poster, Fès, 19-21 avril.

Salmon-Alt, S., Bick, E., Romary, L., Pierrel, J.-M. (2004). La FREEBANK : vers une base libre de corpus annotés, *Actes de TALN*, Fès, 19-21 avril. <http://aune.lpl.univ-aix.fr/jep-taln04/proceed/actes/taln2004-Fez/SalmonAlt2.pdf> (consulté le 20/06/13)

Salton, G., Allan J., and Burckley, C. (1993). Approaches to Passage Retrieval in Full Text Information Systems. In Korfhage *et al.*, 49-58.

Salton, G., Singhal, A., Buckley, C., Mitra, M. (1996). Automatic Text Decomposition Using Text Segments and Text Themes, *Proceedings of the seventh ACM conference on Hypertext*, 53-65.

Santamaría, C., Gonzalo, J., Artiles, J. (2010). Wikipedia as sense inventory to improve diversity in web search results, *Proceedings of ACL*, 1357-1366.

Schang, E., Boyer, A., Muzerelle, J., Antoine, J.-Y., Eshkol, I., Maurel, D. (2011). Coreference and anaphoric annotations for spontaneous speech corpus in French, *Proceedings of DAARC'2011*, Faro, Portugal.

Schlobinski, P., Schütze-Coburn, S. (1992). On the Topic of Topic and Topic Continuity, *Linguistics*, *30* (1), 89-121.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of the International Conference on New Methods Language Processing*, Manchester, UK, 44-49.

Schiffrin, D. (1992). Conditionals as Topics in Discourse, *Linguistics*, *30* (1), 217-241.

Schnedecker, C. (1997). Nom propre et chaînes de référence. *Recherches Linguistiques*, *21.* Paris : Klincksieck.

Schnedecker, C. (2001). Adverbes ordinaux et introducteurs de cadre : aspects linguistiques et cognitifs, *Linguisticae Investigationes, 24* (2), 257-287.

Schnedecker, C. (2002). « Les corrélats anaphoriques *l'un / l'autre, le premier / le second* : aspects cohésifs de la référence "en stéréo" », *in* Andersen, H. L., Nølke, H. (éds), 2002, *Macro-syntaxe et macro-sémantique. Actes du colloque international d'Århus, 17-19 mai*, Berne, Peter Lang, 257-283.

Schnedecker, C. (2005). Les chaînes de référence dans les portraits journalistiques : éléments de description, *Travaux de linguistique, 51* (2), Duculot, 85-133.

Schnedecker, C. (2006). De l'un à l'autre et réciproquement : aspects sémantiques, discursifs et cognitifs des pronoms anaphoriques corrélés, Louvain : De Boeck/université, 376p.

Schnedecker, C. (2009). La notion de « saillance » : problèmes définitoires et avatars, *Colloque Saillance*, Université de Genève, 3-6.

Schnedecker, C., Bianco, M. (1995). Antécédents « dispersés » et référents conjoints ou la construction mentale et la reprise pronominale des entités plurielles, *Sémiotiques*, *8*, 79-108.

Schnedecker, C., Bianco, M. (2000). Référence « Pro-nominale » Plurielle, *Verbum, 22, 4.*

Schnedecker, C., Longo, L. (2012). Impact des genres sur la composition des chaînes de référence : le cas des faits divers, *Actes du Congrès Mondial de Linguistique Française (CMLF), SHS Web of Conferences 1*, 1957-1972.

Seddah, D., Chrupała, G., Cetinoglu, O., Genabith, J., Candido, M. (2010). Lemmatization and Lexicalized Statistical Parsing of Morphologically-Rich

Languages: the Case of French*, Proceedings of the NAACL-HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages (SPMRL 2010)*, Los Angeles, 85-93.

Siblot, P. (2000). Qu'est-ce que poser un thème ? *La thématisation dans les langues.* Éd. par Claude GUIMIER. Neuchatel : Peter Lang.

Sidner, C. (1983). Focusing in the comprehension of definite anaphora, *in* Grosz, B. *et al.* (eds.) (1983). Readings in natural language processing, Morgan Kaufmann, P., 362-394.

Sigogne, A. (2010). HybridTagger : un étiqueteur hybride pour le français, *Actes de MajecSTIC* (MAnifestation des JEunes Chercheurs en Sciences et Technologies de l'Information et de la Communication), Bordeaux, disponible à : http://majecstic2010.labri.fr/Version_Finale2/Bioinformatique_et_Systeme_d_information/Sigogne.pdf (consulté le 27/07/2011).

Simon, A., Gravier, G., Sébillot, P. (2013). Un modèle segmental probabiliste combinant cohésion lexicale et rupture lexicale pour la segmentation thématique, *Actes de TALN*, 17-21 juin, Les Sables d'Olonne, 202-214.

Siouffi, G., Van Raemdonck, D. (1999). 100 fiches pour comprendre la linguistique, Bréal.

Sitbon, L. (2004). Fusion d'approches non-supervisées et génériques pour la segmentation thématique, mémoire, 15p.

Sitbon, L., Bellot, P. (2004). Evaluation de méthodes de segmentation thématique linéaire non supervisées après adaptation au français, *Actes de TALN*, Fez, Maroc, 441-450.

Smolczewska, A., Lallich-Boidin, G. (2004). Validation par prototypage d'un modèle de segmentation des documents techniques composites. In Enjalbert P. et Gaio, M., éds. Approches sémantiques du document numérique, *Actes du 7ème Colloque International sur le Document Electronique (CIDE.7)*, La Rochelle, 75-92.

Sobha, Lalitha Devi., Pattabhi, RK Rao., Vijay Sundar Ram, R., Malarkodi, CS., Akilandeswari, A. (2011). Hybrid Approach for Coreference Resolution, *Proceedings of Computational Natural Language Learning: Shared Task*, *23-24,* Portland, Oregon, 93-96.

Soon, W. M., Ng, H. T., Lim, D. (2001). A machine learning approach to coreference resolution of noun phrases, *Computational Linguistics, 27 (4)*, 521-544.

Spinoza, B. (1670). *Traité des autorités théologique et politique*, *Œuvres complètes.* Paris : Bibliothèque de la pléiade.

Stairmand, M.A. (1996). A Computational Analysis of Lexical Cohesion with applications in Information Retrieval, thèse de Doctorat, Department of Language Engineering, UMIST Computational Linguistics Laboratory.

Stark, H.A. (1988). What do paragraph markings do ? *Discourse Processes*, *11*, 275-303.

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufiş, D., Varga, D. (2006). The JRC-Acquis: A multilingual aligned parallel corpus with 20+languages, *Proceedings of LREC*, Italy.

Steinberger, J., Poesio, M., Kabadjov, M. A., Jeek, K. (2007). Two uses of anaphora resolution in summarization. *Information Processing and Management: an International Journal*, *43* (6), 1663-1680.

Stokes, N. (2003). Spoken and written news story segmentation using lexical chaining, *Proceedings of the Student Workshop at HLT-NAACL, Companion Volume*, 49-54.

Stokes, N., Carthy, J., Smeaton, A.F. (2004). « Select: a lexical cohesion based news story segmentation system », *AI Communications*, *17* (1), 3-12.

Stoyanov, V., Gilbert, N., Cardie, C., Riloff, E. (2009). Conundrums in noun phrase coreference resolution: Making sense of the state-of-the-art, *Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing*, Singapore, 656–664.

Stoyanov, V., Cardie, C., Gilbert, N., Riloff, E., Buttler, D., and Hysom, D. (2010). Coreference Resolution with Reconcile, *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, Short Paper, 156-161.

Sutton, C., McCallum, A. (2006). An Introduction to Conditional Random Fields for Relational Learning, In L. Getoor & B. Taskar, Eds., Introduction to Statistical Relational Learning. MIT Press.

Swart H. de, Winter Y. et Zwarts J. (2007). Bare nominals and reference to capacities, *Natural language & Linguistics Theory, 25*, 195-222.

Tardif, O. (2010). Algorithme de résolution de la coréférence d'entités nommées dans des textes journalistiques en français, thèse de Doctorat, Université de Provence.

Teh, Y. W., Jordan, M., Beal, M., Blei, D. (2006). Hierarchical Dirichlet Processes, *Journal of the American Statistical Association, 101*, 1566-1581.

Todirascu, A., Gledhill, C. (2008). Collocations en contexte : extraction et analyse contrastive, *revue électronique Texte et corpus*, *3*, Actes des Journées de la linguistique de Corpus 2007, 137-148.

Todirascu, A., Ion, R., Navlea, M., Longo, L. (2011). French Text Preprocessing with TTL, *Actes de l'Académie Roumaine*, *Series A*, Ed. Romanian Academy, Publishing House of the Romanian Academy, *12* (2), 151-158.

Tomassone, R. (2002). A propos du thème, *les revues pédagogiques de la Mission Laïque française, connaissance du français*, *44*.

Touratier, C. (2010). La sémantique, 2ème édition. Cursus, 287p.

Trávnícek, F. (1962). O tak zvaním aktualnim cleneni vetním (on so-called functional sentence perspective), in *Slovo a slovesnost*, *22*, Prague, 163-171.

Trouilleux, F. (2001). Identification des reprises et interprétation automatique des expressions pronominales dans des textes en français. Thèse de doctorat, GRIL, Université Blaise-Pascal, Clermont- Ferrand.

Trouilleux, F., Gaussier, E., Bès, Gabriel G., Zaenen, A. (2000). Coreference Resolution Evaluation Based on Descriptive Specificity. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC 2000)*, Athènes, Grèce.

Turco, G., Coltier, D. (1988). Des agents doubles de l'organisation textuelle, les marqueurs d'intégration linéaire, *Pratiques*, *57*, 57-79.

Tutin, A., Clouzot, C., Antoniadis, G. (2000). Un corpus d'anaphores discursives pour les études en TAL. Rapport non publié, http://w3.u-grenoble3.fr/tutin/Publis/rapport.pdf (consulté le 14/11/2011).

Tyrkkö, J. (2011). Fuzzy Coherence: Making sense of Continuity in Hypertext Narratives, Thèse de Doctorat, Department of Modern Languages, Université d'Helsinki, Suède.

Ulland, H. (1993). *Les nominalisations agentives et instrumentales en français moderne*. Berne : P. Lang.

Uryupina, O. (2006). Coreference resolution with and without linguistic knowledge, *Proceedings of LREC*, Genoa, Italy, 893–898.

Uryupina, O. (2007). Knowledge Acquisition for Coreference Resolution, thèse de Doctorat, Saarland University, Saarbrücken, Germany, 280p.

Uryupina, O. (2010). Corry: A System for Coreference Resolution, *Proceedings of the 5th International Workshop on Semantic Evaluation*, ACL, Uppsala, Sweden, 15-16 July, 100-103.

Utiyama, M., Isahara, H. (2001). A Statistical Model for Domain-Independent Text Segmentation, *ACL*, 491-498.

Vallette d'Osia, E. (2011). Evaluation numérique d'un module de détection de chaînes de référence, mémoire de Master 2, Université de Strasbourg, 57p.

van Dijk, T. A. (1977a). Sentence Topic and Discourse Topic, *Papers in Slavic Philology*, *1*, 49-61.

van Dijk, T. A. (1977b). Text and Context. London: Longman.

van Dijk, T. A. (1980). The semantics and pragmatics of functional coherence in discourse, in Ferrara, A., Speech act theory: Ten years later, Special issue of Versus, Milano, *26/27*.

van Dijk, T. A. (1985). Handbook of Discourse Analysis, vol. 2. London: Academic Press.

van Dijk, T. A. (1995). On macrostructures, mental models and other inventions. A brief personal history of the Kintsch-van Dijk Theory, in Weaver III, C., Mannes, S., Fletcher, C. R. (Eds.), Discourse comprenhension, Essays in honor of Walter Kintsch, Hillsdale, NJ: Erlbaum, 383-410.

van Dijk, T. A., Kintsch, W. (1983). Strategies of discourse comprehension. New York: Academic Press.

van Rijsbergen, C. J. (1979). Information Retrieval. London: Butterworths, 2ème édition.

Vendler, Z. (1957). Verbs and times, *The Philosophical Review*, Vol. 66/2., 143-160.

Versley, Y., Ponzetto, S.P., Poesio, M., Eidelman, V., Jern, A., Smith, J. Yang, X., Moschitti, A. (2008). Bart: A modular toolkit for coreference resolution, *Proceedings of the ACL-08: HLT Demo Session*, Columbus, Ohio, 9-12.

Vicedo, J. L., Ferrandez, A. (2006). Coreference in Q&A. In Strzalkowski, T. and Harabagiu, S., editors, Advances in Open Domain Question Answering, 32, *Text, Speech and Language Technology*, 71–96, Springer-Verlag, Berlin.

Victorri, B. (2005). Le calcul de la référence, dans Enjalbert, P. (éd.), *Sémantique et traitement automatique des langues*, Hermès, 133-172.

Victorri, B. (2011). ANALEC download Web page, UU http://www.lattice.cnrs.fr/Telecharger-Analec

Vigier, D., Hernandez, N., Charolles, M., Descles, J.-P. (2004). Text organization by combining fine-grained linguistic markers with global statistical measures, *DOCUMENT DESIGN Conference*, Tilburg University, The Netherlands, 22nd-24th January.

Vilain, M., Burger, J., Aberdeen, J. Connolly, D., Hirschman, L. (1995). A Model-Theoretic Coreference Scoring Scheme, *Proceedings of MUC-6*, 45-52.

Virtanen, T. (1992). Discourse Functions of Adverbial Placement in English: Clause-Initial Adverbials of Time and Place in Narratives and Procedural Place Descriptions, Abo Akademi University Press.

Voutilainen, A., Heikkilä, J. (1992). An English constraint grammar (ENGCG) A surface-syntactic parser of English, in Fries, U., Tottie, G., Schneider, P. (Eds.). *Creating and using English language corpora*, Rodopi: Amsterdam and Atlanta, 189-199.

Wagstaff, K. L. (2002). *Intelligent Clustering with Instance-Level Constraints*, these de Doctorat, Cornell University, 140p.

Walker, M., Joshi, A. Prince, E. (1998). Centering Theory in Discourse. Clarendon Press, Oxford.

Weissenbacher, D. (2008). Influence des annotations imparfaites sur les systèmes de Traitement Automatique des Langues, un cadre applicatif : la résolution de l'anaphore pronominale, thèse de Doctorat, Université Paris-Nord – Paris XIII, 184p.

Weissenbacher, D., Nazarenko, A. (2007). Un classifieur bayésien pour la résolution des anaphores, *Actes de TALN*, *1*, 12-15 juin, Toulouse, 47-56.

Widlöcher, A., Bilhaut, F., Hernandez, N., Rioult, F., Charnois, T., Ferrari, S., Enjalbert, P. (2006). Une approche hybride de la segmentation thématique : collaboration du traitement automatique des langues et de la fouille de texte, *Actes de DEfi Fouille de Texte (DEFT'06)*, Semaine du Document Numérique (SDN'06), Fribourg, Suisse.

Widlöcher, A., Mathet, Y. (2009). La plate-forme Glozz : environnement d'annotation et d'exploration de corpus, *Actes de TALN 2009*, session poster, Senlis, France.

Wiggins, D. (2001). Sameness and Substance Renewed. Cambridge University Press, 257p.

Wilmet, M. (1986). La détermination nominale, Paris, PUF.

Wilson, D. (1998). Discourse, coherence and relevance: A reply to Rachel Giora, *journal of Pragmatics*, *29*, 57-74.

Winograd, T. (1972). Understanding Natural Language, *Cognitive Psychology*, *3* (1), Academic Press, 191p.

Wolters, M.K. (2001). *Towards Entity Status*, Ph.D. Thesis, Bonn University.

Yang, X., Zhou, G., Su, J., Tan, C. L. (2003). Coreference resolution using competition learning approach, *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, *1*, Stroudsburg, PA, USA, 176-183.

Yang, X., Su, J., Zhou, G., Tan, C. L. (2004). An NP-cluster based approach to coreference resolution, *Proceedings of the 20th International Conference on Computational Linguistics*, 226-232.

Yang, X., Su, J., Lang, J., Tan, C. L., Li, S. (2008). An entity-mention model for coreference resolution with inductive logic programming, *Proceedings of ACL-08: HLT*, 843-851.

Youmans, G. (1991). A New Tool for Discourse Analysis: The Vocabulary-Management Profile, *Language*, *67* (4), 763-789.

Zampa, V. (2003). Les outils dans l'enseignement : conception et expérimentation d'un prototype pour l'acquisition par expositions à des textes, Thèse de Doctorat, Université Pierre-Mendès-France Gronoble II, 270p.

Zheng, J., Chapman, W. W., Crowley, R. S., Savova, G., K. (2011). Coreference resolution: A review of general methodologies and applications in the clinical domain, *Journal of Biomedical Informatics*, *44*, 1113-1122.

Zhou, G., Su, J., Zhang, J., Zhang, M. (2005). Exploring Various Knowledge in Relation Extraction, *Proceedings of ACL*, 427-434.

Table des matières

UNIVERSITÉ DE STRASBOURG

Laurence LONGO

RBS

Vers des moteurs de recherche « intelligents » : un outil de détection automatique de thèmes

Méthode basée sur l'identification automatique des chaînes de référence

Résumé : Cette thèse se situe dans le domaine du Traitement Automatique des Langues et vise à optimiser la classification des documents dans les moteurs de recherche. Les travaux se concentrent sur le développement d'un outil de détection automatique des thèmes des documents (ATDS-fr). Utilisant peu de connaissances, la méthode hybride adoptée allie des techniques statistiques de segmentation thématique à des méthodes linguistiques identifiant des marqueurs de cohésion. Parmi eux, les chaînes de référence – séquence d'expressions référentielles se rapportant à la même entité du discours (*e.g. Paul…il…cet homme*) – ont fait l'objet d'une attention particulière, car elles constituent un indicateur important dans la détection des thèmes (*i.e.* ce sont des marqueurs d'introduction, de maintien et de changement thématique). Ainsi, à partir d'une étude des chaînes de référence menée dans un corpus issu de genres textuels variés (analyses politiques, rapports publics, lois européennes, éditoriaux, roman), nous avons développé un module d'identification automatique des chaînes de référence *RefGen* qui a été évalué suivant les métriques actuelles de la coréférence.

Mots-clés : Détection automatique de thèmes, chaînes de référence, traitement automatique des langues, sémantique lexicale, coréférence, genres textuels, segmentation thématique, marqueurs linguistiques, cohésion, linguistique de corpus

Abstract: This thesis in the field of Natural Language Processing aims at optimizing documents classification in search engines. This work focuses on the development of a tool that automatically detects documents topics (ATDS-fr). Using poor knowledge, the hybrid method combines statistical techniques for topic segmentation and linguistic methods that identify cohesive markers. Among them, reference chains - sequences of referential expressions referring to the same entity (*e.g.* Paul ... he ... this man) - have been given special attention as they are important topic markers (*i.e.* they are markers of topic introduction, maintenance and change). Thus, from a study of reference chains extracted from a corpus composed of various textual genres (newspapers, public reports, European laws, editorials and novel) we developed *RefGen*, an automatic reference chains identification module, which was evaluated according to current coreference metrics.

Keywords: Topic detection, reference chains, natural language processing, lexical semantics, coreference, textual genre, topic segmentation, linguistic markers, cohesion, corpus linguistics