NANCY 2 UNIVERSITE

UFR des Sciences du Langage

Ecole Doctorale Langage, Temps et Société - Lorraine

Descriptions Définies et Démonstratives : Analyses de Corpus pour la Génération de Textes

THÈSE

présentée et soutenue publiquement le 27 Novembre 2003

pour l'obtention du

Doctorat de l'université Nancy 2

(spécialité Sciences du Langage)

par

Hélène Manuélian

Composition du jury

Rapporteurs :	M. F. Corblin	Université de Paris 4 - Sorbonne
	Mme L. Danlos	Université de Paris 7 - Jussieu
	Mme C. Schnedecker	Université de Strasbourg 2 - M. Bloch
Examinateurs - directeurs de thèse :	M. P. Riley	Université de Nancy 2
	M. J.-M. Pierrel	Université de Nancy 1 - H. Poincaré
Invitée - encadrante :	Mme C. Gardent	CNRS - Loria

LORIA - UMR 7503 / UFR des Sciences du Langage - Nancy 2

Loria

Mis en page avec la classe thloria.

Remerciements

Au moment où cette thèse s'achève, je me rends compte du nombre immense de personnes qui m'ont aidée au cours de ces dernières années. Aussi, j'aimerais ici toutes les remercier, et j'espère n'oublier personne.

Je remercie tout d'abord les rapporteurs de cette thèse, Francis Corblin, Catherine Schnedecker et Laurence Danlos, qui m'ont fait l'honneur de juger mon travail, et dont les remarques, les critiques et les suggestions me permettent de définir des priorités dans les perspectives possibles à mes travaux de thèse.

Je remercie Philip Riley, qui a accepté la co-direction de cette thèse, et sans qui rien n'aurait été possible.

Je remercie également Jean-Marie Pierrel, dont les convictions sur la pluridisciplinarité m'ont permis de faire cette thèse sous sa direction au LORIA, dans des conditions privilégiées.

Je ne saurai exprimer toute ma gratitude à Claire Gardent qui a encadré cette thèse pendant deux ans. Je la remercie de m'avoir consacré autant de temps et d'avoir toujours été enthousiaste et encourageante. Il m'est ici impossible d'énumérer tout ce que j'ai appris à ses côtés, mais je la remercie de m'avoir transmis un peu de sa patience, de sa rigueur, et surtout de m'avoir appris à ne plus reculer devant l'ampleur de certaines tâches.

Merci à Laurent Romary, le responsable de l'équipe Langue et Dialogue du LORIA, pour son accueil chaleureux, sa compréhension et ses conseils avertis pendant les cafés du samedi matin.

Special thanks to Eric Y. Kow qui a réalisé les programmes informatiques qui ont servi pour l'analyse de corpus, pour sa patience et sa gentillesse, autant en tant qu'ingénieur de l'équipe qu'en tant qu'ami.

Bien sûr, je remercie tous les autres membres de l'équipe Langue et Dialogue qui ont contribué à la bonne humeur qui a régné dans mon environnement pendant cinq ans, et particulièrement Azim Roussanaly, Bertrand Gaiffe, Matthieu Quignard et Patrick Blackburn.

Centaines de mercis aux doctorants passés et actuels du LORIA, Susanne, Evelyne, Armelle, Fred, Benoît, Djamé, Huyen, Hacène, Tony et Yannick, avec qui j'ai pu partager mes angoisses comme mes espoirs autour de cafés toujours bienvenus et réconfortants.

Durant les deux dernières années de ma thèse, j'ai eu la chance d'avoir un poste d'ATER à l'Université de Metz, et je tiens à remercier les enseignants du département Sciences du Langage :
Brigitte Wiederspiel qui m'a accueillie dans son département, et avec qui c'est un plaisir de travailler,
Danièle Coltier avec qui j'ai aimé bavarder lorsque nous partagions le même bureau,
Anne-Marie Chabrolle-Cerretini pour ses « bons tuyaux » et les trajets en train,
Laurent Perrin, avec qui j'ai le plaisir de partager aujourd'hui un bureau,

Marceline Laparra qui m'a donné de précieux conseils et dont les encouragements ont une valeur inestimable.

Je dois des remerciements tout particuliers à Caroline Masseron qui m'a prise sous son aile lors de mon arrivée, qui, patiemment, s'est efforcée et s'efforce encore aujourd'hui de m'apprendre un métier, qui m'encourage sans cesse et qui est tout simplement une formidable collègue.

Bien entendu, je remercie aussi tous les autres enseignants du département qui sont toujours prêts à m'apporter l'aide et le soutien dont j'ai besoin.

En plus d'un environnement professionnel particulièrement chaleureux et stimulant, je n'aurais pu terminer cette thèse sans un entourage personnel aussi exceptionnel que le mien.

Je remercie mes parents, pour tout.
Je remercie Elise et Mélanie, mes soeurs, pour ce qu'elles sont et parce que rien ne serait si bien sans elles.
Merci à Pierre-Etienne d'être là,
Merci à Denis, pour ses encouragements silencieux,
Merci à mes grands parents, dont la confiance donne de la force,
Merci aussi à Jocelyne, Claire, David, Sandrine, Christian et Christine.

Je dois aussi remercier mes amis de loin, qui malgré la distance, m'ont apporté leur soutien :
Merci à Lorraine,
Merci à Sophie, qui depuis Toulouse a suivi presque quotidiennement l'histoire de ma thèse,
Merci à Anne-Gaëlle, qui a toujours tout compris sans jamais demander d'explications.

Sans les amis d'ici, ces quatre dernières années auraient été moins faciles et moins gaies, alors je remercie Jacques, Vincent, Sylvie et Renaud, qui m'ont offert leur précieuse amitié, pardonné mes absences et qui m'ont soutenue et surtout supportée jusqu'au bout.

Il y a eu aussi des participants involontaires à cette thèse : ma rédaction a été facilitée par une bande sonore indispensable, alors j'envoie quelques remerciements à Bernard Lenoir, Interpol, Joy Division, The Pastels, Elk City, Tom McRae, BRMC et Stereolab. Peu importe s'ils ne les reçoivent pas.

Enfin, je remercie celui qui a subi cette thèse au quotidien, celui avec qui j'ai partagé tous les jours moments d'angoisse et d'euphorie, moments d'espoir et de désespoir, celui qui a toujours respecté mes choix parfois même au détriment des siens, et ce bien avant le début de la thèse. Mille mercis à toi, Fabien.

À mes parents

Table des matières

Introduction

Désigner un objet de façon à ce que notre interlocuteur l'identifie est une opération fondamentale dans toute communication linguistique. Bien que cette opération paraisse simple, elle implique toute une série de mécanismes complexes pour que notre interlocuteur reconnaisse l'objet dont nous parlons. Pour les linguistes, l'identification et la modélisation de ces mécanismes restent des problèmes centraux et particulièrement en traitement automatique des langues (TAL). Pour que le processus de désignation et d'identification de l'objet aboutisse, il faut que nous partagions un ensemble de connaissances linguistiques ainsi que des représentations et des connaissances du monde. En effet, il ne s'agit pas uniquement de savoir comment nommer un objet et de connaître les mots à notre disposition dans la langue pour que le processus aboutisse, parce qu'il n'existe pas de relation univoque entre les objets du monde et le lexique à notre disposition. Pour permettre à notre interlocuteur d'identifier les objets dont nous parlons, il est nécessaire de savoir comment les nommer *dans la situation de communication*, et c'est le sujet des études sur le phénomène appelé *référence*, phénomène grâce auquel nous arrivons à relier une expression linguistique à notre représentation mentale d'un objet du monde extra-linguistique, que cet objet soit concret ou abstrait, réel ou imaginaire. L'objet auquel réfère une expression linguistique sera appelé *référent*.

Le processus de référence à un objet étant très lié à la situation de communication et au contexte, il existe par conséquent de très nombreuses manières de référer à un même objet. En d'autres mots, de nombreuses expressions linguistiques peuvent être utilisées pour le désigner. Ainsi, dans les exemples 1 à 5, toutes les expressions en gras peuvent être utilisées pour référer à la même personne, selon des points de vue différents ou dans des situations différentes :

(1) ***Paul*** *est heureux.*

(2) ***Un homme*** *est heureux.*

(3) ***Le professeur de mathématiques de Jeanne*** *est heureux.*

(4) ***Mon frère*** *est heureux.*

(5) ***Ce champion de judo*** *est heureux.*

Dans les exemples 6 à 10, les expressions en gras peuvent toutes être utilisées pour référer une seconde fois à *Paul*, mentionné explicitement dans la première phrase. Ici encore, il paraît évident que les expressions, bien qu'elles aient toutes la capacité de référer au même objet, seront employées dans des contextes différents.

(6) *Paul a gagné le premier prix.* ***L'homme*** *est heureux.*

(7) *Paul a gagné le premier prix.* ***Cet homme*** *est heureux.*

(8) *Paul a gagné le premier prix.* ***Il*** *est heureux.*

(9) *Paul a gagné le premier prix.* ***Le lauréat*** *est heureux.*

(10) *Paul a gagné le premier prix.* ***Ce professeur de mathématiques*** *est heureux.*

Dans les exemples 11 à 14, le référent de la deuxième expression en gras est identifié grâce à des inférences faites à partir de la première. Ceci signifie qu'on peut aussi faire référence à des objets dont la présence est seulement implicite dans le contexte.

(11) ***Un couple*** *est entré.* ***L'homme*** *a l'air heureux.*

(12) *Paul* ***a été assassiné.*** ***Le meurtrier*** *n'a pas été arrêté.*

(13) *Jeanne* ***est enceinte.*** *Elle a décidé de* ***le*** *garder.*

(14) *Nous sommes arrivés dans* ***un village.*** ***Cette église****, tout de même, quelle horreur !*

Les exemples que nous venons de présenter montrent que la production d'expressions référentielles peut être vue comme un problème de choix à deux niveaux : tout d'abord le contenu de la description utilisée pour référer à l'objet peut être très variable ; c'est le problème du choix du contenu sémantique de l'expression référentielle. Ensuite, la forme linguistique de l'expression peut elle aussi varier.

Choix du contenu sémantique de la description Le problème de la détermination du contenu d'une expression référentielle peut se poser de trois manières différentes, illustrées par nos trois séries d'exemples (1 à 14) :

Les exemples 1 à 5 montrent une première facette du problème du choix du contenu de la description : quelle est, au vu de la situation de communication, la meilleure expression pour que l'interlocuteur identifie le référent visé par le locuteur, et éventuellement interprète une information au sujet de ce référent ?

La deuxième série d'exemples (6 à 10) pose le problème en des termes différents : étant donné un référent déjà mentionné dans le discours, quelles informations peut contenir une nouvelle expression le désignant (une *expression coréférentielle*) ? En effet, si on observe les exemples 6, 7 et 9, on peut dire que l'information contenue dans l'expression coréférentielle est possible à inférer à partir du contenu de la phrase précédente (le prénom *Paul* permet d'inférer que le référent est un homme, et le verbe gagner permet d'inférer que le sujet de ce verbe peut être appelé *lauréat*). En revanche, l'expression *ce professeur de mathématiques* n'est pas inférable à partir de la phrase précédente, et pourtant, la coréférence entre le prénom *Paul* et cette expression ne fait pas de doute. Enfin, on a dans l'exemple 8 une expression pronominale qui n'a pas d'autre contenu sémantique que d'indiquer qu'on réfère à l'objet déjà mentionné dont le genre est masculin et le nombre singulier. Les problèmes posés par cette série d'exemples sont alors les suivants : comment peut-on choisir le contenu d'une expression pour qu'elle soit interprétée comme coréférant à (i.e. référant au même objet que) la première expression employée ? Ensuite, comment l'information nouvelle à propos du référent est-elle distribuée par rapport à l'information donnée ou

inférable concernant le référent ? En d'autres mots, quels sont les éléments du syntagme qui permettent d'établir la coréférence (en répétant l'information donnée) et quels sont les éléments du syntagme qui apportent l'information nouvelle ?

Pour terminer, la troisième série d'exemples (exemples 11 à 14) pose le problème de l'introduction d'un nouveau référent dans le discours, dont la présence est inférable parce qu'il est implicitement introduit par un autre segment discursif. Ce phénomène est appelé anaphore associative. Le deuxième référent est identifié grâce à la relation qu'il entretient avec le premier, mais quelle est cette relation, et quelles sont, de façon générale, les relations qui peuvent être impliquées dans ce phénomène ?

Choix de la forme linguistique de l'expression L'autre problème posé dans la production (ou génération) d'expressions référentielles est le choix de la forme linguistique que va prendre la description utilisée pour référer à un objet. En effet, nos exemples 1 à 14 montrent que plusieurs formes linguistiques peuvent être utilisées : tout d'abord, on peut utiliser un groupe nominal défini ou indéfini pour référer à un objet qui n'a pas encore été mentionné dans le discours. Quand l'objet a déjà été mentionné, on peut utiliser un groupe nominal défini ou démonstratif, ou un pronom. Enfin, les description définies et démonstratives ainsi que les pronoms peuvent être employées pour produire des anaphores associatives. Il est alors nécessaire d'identifier des contraintes sur l'apparition de ces diverses formes linguistiques, puisque dans chaque situation (première mention dans le discours, coréférence ou anaphore associative), il existe plusieurs possibilités.

La thèse présentée porte sur la génération automatique d'expressions référentielles. Autrement dit, comment faire en sorte qu'une machine puisse faire les bons choix de contenu sémantique et de forme linguistique pour désigner un objet ?

Il existe déjà des algorithmes permettant la génération d'expressions référentielles. Le plus connu est celui de Dale et Reiter [Dale, 1992, Dale et Reiter, 1995], qui est considéré comme l'algorithme standard et qui produit des descriptions définies coréférentielles contenant uniquement de l'information donnée, des pronoms et des *one-anaphors* (des expressions ayant pour tête nominale les pronoms indéfinis *un/une* comme par exemple dans le syntagme anaphorique *un/une rouge*, mis pour *un X rouge*, X ayant déjà été mentionné.). Diverses extensions de cet algorithme ont été réalisées, dont celle de Gardent et Striegnitz [Gardent et Striegnitz, 2003] pour les anaphores associatives. Cet algorithme traite la génération des anaphores associatives méronymiques, c'est à dire de celles qui réfèrent à une partie de l'objet déjà mentionné, comme dans l'exemple suivant, dans lequel on infère la présence du toit grâce à la mention explicite d'une maison.

(15) *La maison n'est plus habitable. Le toit s'est effondré.*

Les algorithmes de génération existants ne tiennent donc compte que d'une partie des choix disponibles pour référer à un objet. Il est donc possible de les étendre de façon à obtenir une génération automatique encore plus proche de l'utilisation naturelle des expressions référentielles, en introduisant d'une part la possibilité d'utiliser le déterminant démonstratif, et d'autre part la possibilité d'ajouter de l'information nouvelle sur le référent dans les expressions coréférentielles. Il semble aussi possible d'étendre le nombre des

relations impliquées dans l'anaphore associative, dans la mesure où les exemples 11 à 14 sont des illustrations de l'anaphore associative, mais n'impliquent pas uniquement des relations méronymiques. Nous présentons donc dans notre thèse des propositions d'extensions de ces algorithmes dans les trois directions suivantes :

1) Tout d'abord, nous proposons un traitement extensif des anaphores associatives en répondant à deux questions : quelles sont les relations sémantiques impliquées dans l'anaphore associative ? Quelles sont les sources de ces relations ? La réponse à ces questions permet d'étendre l'algorithme de Gardent et Striegnitz.

2) Ensuite, nous proposons une extension des algorithmes pour le choix du contenu sémantique des expressions coréférentielles en montrant qu'il est possible d'intégrer des descriptions ajoutant de l'information nouvelle à propos du référent.

3) Enfin, nous proposons d'étendre les algorithmes aux descriptions démonstratives, en donnant des critères de choix du déterminant. Pour cela, nous devons définir des contraintes sur l'apparition du défini et du démonstratif.

La méthodologie employée est une étude empirique basée sur une analyse semi-automatique de corpus, combinée avec une analyse de la littérature existant sur le sujet. Tout en les utilisant, il nous a semblé nécessaire de dépasser les analyses théoriques par introspection pour pouvoir parvenir à notre but, les contraintes définies par ces analyses ne semblant pas toujours généralisables, et surtout n'étant que peu souvent formalisables et opérationnelles pour le traitement automatique des langues. Ces analyses présentent cependant de nombreux éléments que nous pouvons réutiliser, et qui servent de base à notre analyse de corpus.

De plus, l'analyse de corpus présente une série d'avantages simples : tout d'abord, le linguiste n'a *a priori* plus à se poser la question de l'acceptabilité et de la couverture des exemples utilisés dans la démonstration. En effet, le corpus rassemble une série d'emplois *attestés* et souvent *variés* dont on peut supposer qu'ils sont représentatifs des emplois les plus fréquents et les plus spontanés. Bien entendu, utiliser un seul corpus, un seul type de texte peut faire penser que la couverture des exemples n'est pas aussi large qu'on pourrait le souhaiter. Cependant, il apparaît tout de même clairement que l'étude de corpus permet de rassembler un grand nombre d'exemples, ce qui permet de valider la théorie sur des données importantes, et non sur quelques exemples construits.

Ensuite, un autre problème se pose quand il existe des cas où plusieurs solutions sont acceptables, ce qui est fréquemment le cas pour le choix entre le défini et le démonstratif. Le corpus permet de dégager une solution préférée par les locuteurs, qui ne sont pas orientés par la théorie lors de la production de leur énoncé, que cette production de l'énoncé soit spontanée (orale ou écrite) ou non (discours oral préparé ou écrit).

Le corpus que nous avons étudié est le corpus PAROLE fourni par l'ATILF. Nous le présentons en détail dans le cinquième chapitre de notre thèse. Il contient environ soixante cinq mille mots, dont dix mille descriptions définies et démonstratives. Il présentait l'avantage d'être annoté au niveau morphosyntaxique, ce qui a facilité son annotation au niveau référentiel.

La plupart des exemples donnés dans cette thèse sont donc des exemples attestés, ex-

traits du corpus que nous avons étudié. Nous les différencions des exemples construits par leur numérotation, qui est préfixée par un « c ». Ils appuieront notre démonstration tout au long de cette thèse, et montreront que les cas envisagés dans les analyses théoriques des déterminants sont souvent simplifiés.

Notre thèse est donc située à la rencontre de trois domaines en linguistique (la référence, la génération de texte, et la linguistique de corpus), et toute la difficulté de notre travail repose sur l'harmonie de cette rencontre. Nous avons restreint notre champ d'étude à la génération de descriptions définies et démonstratives, qui posent en elles-mêmes, par leur apparente proximité d'emploi, des problèmes de choix importants. Nous devions confirmer et étendre les études connues sur ce type d'expressions référentielles en réalisant une étude de corpus, dans la mesure où les études théoriques ne répondent pas à toutes les questions qui se posent en génération automatique. De cette manière, nous avons pu établir une série de contraintes utilisables et formalisables en génération automatique de textes, dans le but de construire des algorithmes tenant compte de données linguistiques variées et les plus complètes possibles. Ainsi, notre thèse se présente de la manière suivante :

Dans la première partie, nous présentons les concepts qui ont servi notre travail.

Le premier chapitre expose les données connues sur la référence et les expressions référentielles, ainsi que les grandes théories sur la détermination en français, mais aussi en anglais. Nous définissons dans un premier temps les concepts généraux, puis nous présentons les divers emplois des description définies et démonstratives identifiés dans la littérature théorique, mais aussi dans la littérature empirique basée sur l'analyse de corpus. Nous concluons par une synthèse montrant les limites des analyses existantes.

Nous présentons ensuite (chapitre 2) la problématique de la génération automatique de textes, et plus particulièrement de la génération d'expressions référentielles. Nous présentons les algorithmes existants et nous justifions le choix que nous faisons en décidant d'étendre non pas l'algorithme standard de génération d'expressions référentielles, mais l'algorithme de [Gardent et Striegnitz, 2003].

Notre travail étant fondé sur une analyse de corpus, nous présentons dans le troisième chapitre les concepts liés à la linguistique de corpus et au traitement de corpus électroniques. Nous revenons sur les méthodes d'analyse de corpus, et sur le problème de la standardisation du format des données contenues dans les corpus.

Nous terminons la première partie par une synthèse exposant comment se lient les problèmes posés par les trois domaines abordés (la référence, la génération de textes et la linguistique de corpus), et les objectifs que doit réaliser une étude de corpus pour la génération d'expressions référentielles.

La deuxième partie du document présente le travail réalisé pendant la thèse.

Nous présentons dans le chapitre 5 le travail que nous avons réalisé sur le corpus PAROLE. Nous y décrivons la série de traitements réalisés pour arriver à une première annotation du corpus dont nous présentons les résultats et les limites.

Dans le sixième chapitre, nous exposons les résultats d'une étude approfondie des anaphores associatives annotées dans notre corpus. Cette étude a été un premier pas dans la démonstration de l'importance des connaissances du monde dans la génération

d'expressions référentielles et surtout de l'importance de la structuration du contexte dans un générateur. Nous concluons par une proposition pour l'extension de l'algorithme de [Gardent et Striegnitz, 2003].

Dans le chapitre sept, nous présentons une étude des description définies et démonstratives fondée sur une seconde annotation de notre corpus et originale dans le sens où nous montrons que non seulement l'inférence sur les connaissances du monde et la structuration du contexte sont importantes en génération, mais aussi que la notion d'informativité d'une expression référentielle est importante. Les observations que nous avons faites sur le corpus nous permettent de proposer une seconde extension de l'algorithme de [Gardent et Striegnitz, 2003].

Le chapitre huit concerne le choix du déterminant. Nous reprenons notre étude de corpus basée sur l'informativité des expressions référentielles, et nous établissons de contraintes sur le choix du déterminant ; nous montrons ensuite que ces contraintes sont de plusieurs types : il existe en effet des contraintes sémantiques, mais les contraintes syntaxiques ont elle aussi leur importance dans le choix du déterminant. Ces contraintes nous amènent à la rédaction d'un algorithme pour le choix du déterminant.

Première partie

Expressions référentielles, Génération automatique et Analyse de corpus

Chapitre 1

Descriptions définies et démonstratives

La référence (le fait de pouvoir relier une expression linguistique à un objet du monde) et les phénomènes liés au choix du déterminant dans les expressions référentielles sont des sujets vastes, qui ont déjà donné lieu à une abondante littérature, que nous essayons de résumer ici. Bien entendu, nous n'avons pas la prétention d'être exhaustive sur le domaine, tant il est vaste. Notre but est de donner dans ce chapitre tout le matériau qui sert de support à notre étude. Dans un premier temps, nous nous concentrons sur la référence et sur les processus d'identification des référents en fonction des formes linguistiques que prennent les expressions référentielles (section 1.1). Nous n'exposerons que brièvement les théories fondamentales dans le domaine, afin de pouvoir simplement expliquer l'utilisation des formes linguistiques permettant de référer. Nous présenterons tout d'abord les formes que prennent les expressions référentielles (section 1.1.1), puis les divers types de référence identifiés dans la littérature (sections 1.1.2 et 1.1.3). Nous terminerons cette section sur la référence par la description des processus d'identification des référents en lien avec la détermination des syntagmes (section 1.1.5). Nous exposerons ensuite plus longuement les divers emplois des expressions référentielles qui nous concernent (i.e. les syntagmes nominaux démonstratifs et définis), d'un point de vue théorique dans la section 1.2 et d'un point de vue empirique dans la section 1.3. Notre but étant de les générer automatiquement, il nous faut alors isoler non seulement les principes de fonctionnement des expressions référentielles, mais aussi leurs conditions d'apparition, et leurs divers emplois. En effet, la littérature sur la référence et les expressions référentielles est la plupart du temps située dans un point de vue d'analyse ou de compréhension des textes. La question dans cette perspective est alors de savoir comment les déterminants fonctionnent dans l'identification des référents. Notre perspective étant la production d'expressions référentielles, la nécessité est pour nous plus grande de connaître les différentes utilisations et les contextes d'apparition des déterminants. Nous terminerons donc ce chapitre par une synthèse des contraintes régissant l'utilisation des deux déterminants (section 1.4).

1.1 Référence et processus d'identification des référents

1.1.1 Les formes d'expressions référentielles

La référence est le processus qui permet de relier une expression linguistique à un objet appartenant au monde ou à un modèle mental du monde, qu'on appelle le *référent*. Le référent peut être abstrait ou concret, et peut aussi ne pas exister (cf. personnages de fiction, [Reboul et al., 1997]). De façon générale, on peut référer à des entités de tous types (i.e. des événements ou des objets). Les formes que prennent les expressions référentielles peuvent donc être variées.

Référence aux objets La référence aux objets est probablement le type de référence le plus étudié. On réfère aux objets par le biais de groupes nominaux ou de pronoms. Les groupes nominaux peuvent alors prendre la forme de noms propres, de syntagmes nominaux définis, indéfinis ou démonstratifs. Les pronoms ne constituant pas l'objet de notre étude, nous ne nous y attarderons pas. Signalons simplement que la référence pronominale constitue en elle-même un objet d'étude complexe, ne serait-ce que par la variété des pronoms qui existent (i.e. déictiques, personnels, mentionnels).

Référence aux événements La référence aux événements est un phénomène peu étudié comparativement à la référence aux objets [Danlos, 1999]. Elle constitue un objet d'étude complexe de par la variété des formes linguistiques qu'elle peut prendre. En effet, il est possible de faire référence à un événement par un syntagme nominal, mais aussi par une ou plusieurs phrase(s). L'étude de la référence aux événements devient alors difficile à restreindre, puisque toute expression linguistique devient potentiellement une expression référentielle.

Dire qu'on étudie les expressions référentielles de manière générale devient donc, suite à ce que nous venons d'exposer, assez vague. Par ailleurs, il est difficile, d'un point de vue linguistique, de pouvoir affirmer qu'on étudie toutes les expressions référentielles. Pour bien faire, il est nécessaire de restreindre quelque peu l'objet d'étude. Nous avons choisi de faire porter notre étude sur les expressions référentielles nominales, définies et démonstratives, que nous appellerons désormais *descriptions définies et démonstratives*. Notre choix se portant sur une forme linguistique particulière, nous ne distinguerons pas l'étude de la référence aux événements de celle de la référence aux objets, puisque les deux peuvent être impliquées dans des descriptions définies et démonstratives. Dans les deux paragraphes suivants, nous étudierons les divers types de référence identifiés dans la littérature et reliés à des formes linguistiques particulières, qui sont des premiers pas vers l'explication de la façon dont une expression linguistique saisit, isole et extrait son référent du contexte. En effet, bien que nous travaillions dans une perspective de génération, ces données fondamentales en analyse sont malgré tout nécessaires, parce qu'elles permettent de savoir comment le contexte doit être structuré lors de la génération des expressions référentielles (nous revenons sur ce point au chapitre 2).

1.1.2 Référence actuelle et référence virtuelle

Milner distingue deux types de référence pour le syntagme nominal : la référence virtuelle et la référence actuelle [Milner, 1982]. Nous exposons brièvement cette distinction, qui nous permettra d'expliquer la différence de fonctionnement entre le défini et le démonstratif notamment en section 1.2.3.2.

Pour pouvoir identifier le référent d'un groupe nominal, on doit faire appel à ces deux types de référence. Dans un premier temps, on utilise la référence virtuelle du syntagme qui est proche du sens lexical du syntagme nominal sans son déterminant. Cette référence virtuelle permet de savoir quelles conditions doit remplir le référent pour être désigné par la description utilisée. Ainsi, quand on parle de *chat*, on doit passer par la référence virtuelle du mot chat, c'est à dire par les conditions qui font qu'un chat est un chat (i.e. un chat est animal à moustache qui miaule, etc...). De la même façon, quand on parle d'un *chat noir* on doit passer par les conditions qui font qu'un chat noir est un chat noir.

La référence actuelle est ce qui permet d'attribuer à la séquence linguistique un référent du monde extralinguistique. La référence actuelle d'un syntagme nominal n'a d'existence que dans un contexte d'interprétation et si le nom est en présence d'un déterminant.

1.1.3 Descriptions attributives et référentielles

Donnellan distingue lui aussi deux types de descriptions définies : les descriptions attributives et les descriptions référentielles [Donnellan, 1966].

Descriptions attributives Les descriptions attributives possèdent les caractéristiques suivantes :

- Elles servent à dire quelque chose à propos d'un objet, quel qu'il soit, à partir du moment où il correspond à la description.
- Il y a une présupposition générale que quelque chose satisfait la description : si rien ne satisfait la description, la question *qui est le x qui y ?* n'a pas de réponse.
- Elles dénotent mais ne réfèrent pas, c'est à dire qu'elles ont un sens, une référence virtuelle, mais pas de référence actuelle.

Descriptions référentielles Les descriptions référentielles possèdent des caractéristiques parallèles à celles énumérées pour les descriptions attributives :

- Elles servent à ce que l'auditeur puisse identifier le référent dont le locuteur parle.
- Il y a une présupposition qui affirme qu'un objet particulier correspond à la description : si rien ne satisfait la description, la question *qui est le x qui y ?* a une réponse sous forme de description définie.
- Elles dénotent et elles réfèrent : ces expressions ont à la fois une référence actuelle et une référence virtuelle.

Par ailleurs, la différence entre l'utilisation attributive ou référentielle d'une description ne dépend pas des croyances du locuteur. Une description peut être utilisée attributivement même quand le locuteur pense qu'un objet particulier satisfait la description.

Ainsi, la description *l'assassin de Smith est fou* peut être vraie même si le locuteur pense que cet assassin est Jones, et qu'ensuite, il apprend que ce n'est pas le cas. Une description peut être utilisée référentiellement sans que quoi que ce soit ne corresponde à la description. Ainsi, la description *le roi* dans *le roi est au palais* peut être utilisée, même si le locuteur pense que la personne est un usurpateur.

Cette distinction est importante parce qu'elle ne semble s'appliquer qu'aux descriptions définies. Il pourrait alors s'avérer qu'elle soit une possibilité d'expliquer les différences d'emploi entre défini et démonstratif. En effet, s'il est impossible d'utiliser une description démonstrative de façon attributive, cette distinction pourrait s'avérer opérationnelle dans le choix du déterminant.

1.1.4 Coréférence et anaphore

Les expressions coréférentielles Deux expressions sont coréférentielles quand elles réfèrent au même objet, c'est à dire quand elles ont le même référent. Deux expressions coréférentielles peuvent être interprétées indépendamment l'une de l'autre (exemple 16). En d'autres mots, on n'a pas besoin d'avoir lu ou entendu la première expression référentielle pour interpréter correctement la deuxième.

Les expressions anaphoriques Une expression est anaphorique si la présence d'un segment linguistique antérieur est nécessaire pour identifier son référent (exemple 17). Une expression anaphorique peut être coréférente avec son antécédent, mais ça n'est pas obligatoire : on parle alors d'anaphore associative (exemple 18).

(16) ***Iggy Pop*** *a donné un concert hier soir.* ***L'ancien chanteur des Stooges*** *a interprété « China Girl ».*

(17) ***David Bowie*** *a donné un concert hier soir.* ***Il*** *a interprété « China Girl ».*

(18) *Iggy Pop a donné* ***un concert*** *hier soir.* ***La salle*** *était pleine.*

Ainsi, dans le texte 16, *Iggy Pop* et *L'ancien chanteur des Stooges* sont des expressions coréférentielles (i.e. elles réfèrent au même objet), mais peuvent être interprétées de façon indépendante. En revanche, dans le texte 17, le pronom *il* et le nom *David Bowie* coréfèrent, mais on dira que le pronom est anaphorique dans la mesure où son référent n'est identifiable que dans le contexte. De la même manière, la description définie *la salle* est interprétable grâce à la présence du syntagme *un concert* : on l'identifie alors comme *la salle où avait lieu le concert* (pour plus de détails sur l'anaphore associative, cf. section 1.2.4). Une expression anaphorique a besoin de s'appuyer sur un *antécédent* pour être interprétée, tandis qu'une expression simplement coréférentielle a un antécédent, mais n'a pas forcément besoin de lui pour être interprétée.

Les chaînes de référence La notion de chaîne de référence a été élaborée par Chastain [Chastain, 1979]. Cette notion a été ensuite étudiée par plusieurs auteurs dont [Corblin, 1995] et [Schnedecker, 1997], qui donnent les éléments de définition que nous

reprenons dans ce paragraphe. Pour commencer, une définition informelle peut être la suivante : *« il s'agit des expressions qui, tout au long des textes, indiquent que le locuteur fait référence à un même individu et qu'on appelle, pour cette raison, des expressions co-référentielles. »* [Schnedecker, 1997]. Nous considérerons donc ici qu'une chaîne de référence est l'ensemble de toutes les expressions référant au même objet dans un texte. La notion de chaîne de référence pose cependant quelques problèmes si on veut la définir de façon plus formelle : combien de maillons doit contenir la chaîne (en d'autres mots, combien d'expressions référentielles doit contenir la chaîne) ? Si les maillons sont très espacés, peut-on considérer qu'ils font partie de la même chaîne, ou doit-on décider que la chaîne s'interrompt pour reprendre plus loin dans le texte ?

Concernant le nombre de maillons d'une chaîne, le consensus porte sur le minimum : pour constituer une chaîne de référence, il est nécessaire d'avoir au moins trois expressions coréférentielles. Ceci permet alors de dépasser la notion d'anaphore qui est une relation binaire (entretenue par seulement deux expressions référentielles). En revanche, il ne semble pas exister de nombre maximal de maillons pour une chaîne de référence. Cette question reste en suspens dans l'ouvrage de Schnedecker. En revanche, il semble clair qu'on doit considérer que la chaîne redémarre lorsqu'on utilise une redénomination, ou un nouveau syntagme nominal plein, interprétable indépendamment de la première mention. Il n'est donc pas possible de donner un critère numérique, mais on peut borner les chaînes en fonction des unités lexicales qui les composent. Par ailleurs, Corblin donne des indications sur la composition des chaînes de référence importantes dans le cadre de notre travail : une description démonstrative ne peut pas se trouver en position initiale d'une chaîne. En revanche, une description définie pourra soit constituer le maillon initial, soit un maillon subséquent dans la chaîne [Corblin, 1995].

La distance entre les maillons pose en elle-même plusieurs problèmes : tout d'abord, pour calculer la distance « physique » entre chaque élément de la chaîne, on ne sait pas exactement quelle unité de mesure utiliser : faut-il compter les mots, les propositions, les phrases ? Plusieurs hypothèses sont avancées, mais Schnedecker ne tranche pas réellement en faveur de l'une ou de l'autre.

Notre thèse portant sur les diverses utilisations des descriptions définies et démonstratives, elle portera donc à la fois sur les phénomènes de coréférence et d'anaphore, à l'intérieur des chaînes de référence. Rappelons que dans une perspective de génération, nous devons nous intéresser aux formes utilisées pour produire ces phénomènes, et donc aux contextes d'apparition de ces formes. C'est pourquoi nous allons maintenant nous intéresser plus directement aux formes linguistiques qui nous concernent : les descriptions définies et démonstratives. Nous allons donc maintenant présenter la façon dont les déterminants donnent des instructions sur la façon d'extraire les référents des expressions référentielles, avant de passer à notre section concernant les emplois des descriptions dans leur totalité.

1.1.5 Processus d'identification des référents en fonction du déterminant

La forme linguistique que prend une expression référentielle peut être vue comme une instruction donnée à l'interlocuteur pour lui permettre d'identifier le référent visé. Dans les expressions référentielles que nous étudions, le déterminant est le principal porteur de ces instructions. Il nous est donc nécessaire de définir avec précision les instructions qu'ils donnent sur la manière d'identifier le référent désigné par le syntagme qu'ils déterminent. Précisons dès à présent que les deux déterminants étudiés ici peuvent avoir des interprétations génériques (le syntagme nominal singulier est alors utilisé pour désigner l'ensemble des objets correspondant à la description). Ces interprétations n'entrant pas dans notre champ d'étude, nous ne les traiterons pas dans cette partie.

1.1.5.1 Extraction dans un ensemble

Pour permettre l'identification d'un référent, l'expression référentielle doit permettre un contraste entre le référent visé et les autres objets du contexte. Le défini est généralement reconnu comme déterminant un syntagme dont le référent est identifiable dans le contexte grâce à la référence virtuelle du groupe nominal. Le contraste provoqué par l'expression référentielle doit être *externe à sa classe*, c'est à dire qu'on oppose le référent visé aux autres objets du contexte en utilisant le nom de la classe d'objets à laquelle il appartient [Corblin, 1987]. Au contraire, le démonstratif permet d'identifier un objet par un contraste *interne à la classe* d'objets dénotée par le nom. En d'autres mots, le démonstratif désigne un objet particulier parmi un ensemble d'objets appartenant à la même classe.

Nous reprenons les exemples 19 et 20 à Corblin :

(19) *Ce garçon est stupide.*

(20) *Le garçon est stupide.*

Pour Corblin, l'exemple 19 sera interprété comme « le garçon dont je parle par opposition aux autres garçons ». En revanche, l'exemple 20 sera interprété comme « le garçon par opposition aux autres individus mentionnés qui ne sont pas des garçons ». Pour reprendre les termes exacts de Corblin, avec le démonstratif *l'individu qui fait l'objet du jugement est classifié comme un N opposable à d'autres (N)*, alors qu'avec le défini *l'individu qui fait l'objet du jugement est saisi par opposition aux autres individus mentionnés qui ne sont pas des N.*

Ces principes, énoncés par Corblin, ont été schématisés pour un modèle computationnel de résolution de la référence par [Gaiffe, 1992], dont nous reproduisons les axiologies (figure 1.1). Dans ces schémas, Q représente le contexte, N la classe d'objets dénotée par le nom, et P les propriétés utilisées pour différencier le référent visé parmi la classe des N. Avec le défini, le contexte est partitionné en deux ensembles, d'un côté l'individu appartenant à la classe des N, et de l'autre, les individus n'appartenant pas à la classe des N : le garçon par opposition aux « non-garçons ». Avec le démonstratif, le contexte est aussi divisé en deux ensembles, mais différemment : d'une part on aura l'individu appartenant à la classe des N auquel on réfère (propriété P), et d'autre part on aura les autres individus de la classe des N, dont on ne parle pas (n'ayant pas la propriété P). Ainsi, dans l'exemple 19, on oppose le garçon qu'on est en train de juger aux autres garçons qu'on n'est pas en train de juger.

FIG. 1.1 – Principes d'extraction des référents

1.1.5.2 Le paradoxe de la reprise immédiate

Le mode d'extraction des référents attribué à chaque déterminant nous amène directement à ce que Kleiber [Kleiber, 1986, Kleiber, 1988] et Corblin [Corblin, 1995] appellent *le paradoxe de la reprise immédiate*. Si le constat est simple : le défini comme le démonstratif peuvent servir à faire une reprise contenant le même nom tête, la motivation du choix du déterminant reste assez obscure, et les notions de contraste interne ou externe ne semblent pas pleinement satisfaisantes. Nous livrons ici les explications que Kleiber donne dans ses articles de 1986 et 1988 à ces constatations. Considérons les exemples suivants (nous les reprenons à Kleiber, ainsi que ses jugements d'acceptabilité[1]) :

(21) *J'ai vu une voiture. ?La voiture roulait vite.*

(22) *J'ai vu une voiture. Cette voiture roulait vite.*

(23) *J'ai vu un camion et une voiture. La voiture roulait vite.*

(24) *J'ai vu un camion et une voiture. ? Cette voiture roulait vite.*

(25) *J'ai vu un camion et une voiture. Le camion roulait vite.*

(26) *J'ai vu un camion et une voiture. ? Ce camion roulait vite.*

(27) *Un avion s'est écrasé hier. L'avion venait de Miami.*

(28) *Un avion s'est écrasé hier. ?Cet avion venait de Miami.*

(29) *Un avion s'est écrasé hier. ? L'avion relie habituellement Miami à New York.*

(30) *Un avion s'est écrasé hier. Cet avion relie habituellement Miami à New York.*

Dans les exemples 21 à 26, Kleiber confirme la validité de la thèse de l'opposition notionnelle développée par Corblin. En effet, si le démonstratif est meilleur en 22 qu'en 21, c'est parce que l'objet *voiture* ne peut pas être opposé à des objets d'un autre type dans ce contexte. En revanche, le défini est meilleur en 23 et 25 parce qu'on peut opposer les référents entre eux grâce à la classe d'objets dénotée par le nom. Le problème soulevé par les exemples 27 à 30 est le suivant : comment, avec la même phrase introductive, est-il possible que le défini soit meilleur dans un cas et le démonstratif meilleur dans un autre cas ? Pour Kleiber, l'explication réside dans les faits suivants : le défini s'interprète dans les *circonstances d'évaluation* fournies par la phrase introductive, alors que le démonstratif est

[1]Tout au long de la thèse, nous utiliserons les notations usuelles pour les jugements d'acceptabilité : ? pour les énoncés douteux et * pour les énoncés non acceptables.

un désignateur direct, un *connecteur anaphorique*. Dans les phrases 27 et 28, le prédicat de la seconde phrase exige que le syntagme soit interprété dans les circonstances d'évaluation constituées par la première phrase. Il y a une forme de continuité événementielle, qui n'apparaît pas dans les phrases 29 et 30, dans lesquelles le syntagme nominal sera plus facilement interprété si le démonstratif est utilisé. Le démonstratif permet un détachement des circonstances d'évaluation disponibles dans la phrase introductive, ce qui lui permet de provoquer une rupture dans la progression événementielle. Le rôle du prédicat dans la phrase où s'effectue la reprise est donc crucial.

Ajoutons par ailleurs que Kleiber affirme que le démonstratif est toujours possible à utiliser en reprise immédiate, ce qui n'est pas le cas du défini. En effet, le démonstratif est plus sûr pour assurer le lien coréférentiel. Ceci expliquerait pourquoi nous acceptons personnellement, de façon complètement indifférente les phrases 27 et 28, alors que nous sommes d'accord sur le jugement donné pour les phrases 29 et 30.

1.1.5.3 Unicité du référent

Une expression référentielle doit permettre d'identifier sans ambiguïté le référent visé. Pour ce faire, le moyen le plus simple est de construire une description du référent assez précise et qui le différencie des référents potentiels appartenant au contexte. Bien sûr, ceci n'est pas toujours possible : lorsque le référent est unique dans le contexte, on utilisera la plupart du temps le déterminant défini. On peut aussi utiliser le démonstratif, mais d'autres paramètres entrent alors en jeu, et en particulier la saillance du référent. Lorsque le contexte contient plusieurs objets correspondant à la description, on utilisera le démonstratif pour référer à l'objet le plus saillant correspondant à la description [Corblin, 1987]. L'unicité du référent correspondant à la description dénotée par le syntagme est donc nécessaire à l'utilisation du défini, alors qu'elle ne l'est pas pour le démonstratif.

1.2 Utilisations des déterminants définis et démonstratifs

Dans cette section, nous étudions les divers emplois des deux déterminants. Nous noterons que leurs emplois sont en apparence très proches, mais que les conditions des ces utilisations sont différentes. Nous étudierons dans un premier temps les emplois en première mention, selon la terminologie de [Fraurud, 1990], c'est-à-dire les emplois de groupes nominaux déterminés par le défini ou le démonstratif référant pour la première fois dans le texte à un objet. Nous étudierons ensuite les emplois en mention subséquente ou coréférentielle, c'est à dire les emplois des syntagmes nominaux désignant un objet qui a déjà été mentionné dans le texte. Nous différencierons les reprises directes des reprises indirectes, c'est à dire les mentions subséquentes utilisant la même tête nominale qu'en première mention, des mentions subséquentes utilisant une tête nominale différente de celle utilisée en première mention.

1.2.1 Emplois en première mention

1.2.1.1 Défini en première mention

Vieira dans [Vieira, 1998] propose une classification des utilisations des descriptions définies identifiées dans la littérature anglophone. Nous reprenons ici les deux types d'utilisation en première mention qu'elle donne pour le défini, c'est à dire les utilisations situationnelles (situational uses) et les utilisations non-familières (unfamiliar uses).

Utilisations situationnelles Ces utilisations sont de deux types. Elles se rapportent soit au contexte immédiat du dialogue, soit au contexte plus général de la communication. Pour les premières, il faut que le référent visé soit physiquement présent dans le contexte. Ainsi, à table, on pourra employer le syntagme nominal défini *le sel* pour désigner la salière présente sur la table ; le syntagme nominal défini *le chien* pourra être présent sur une pancarte à l'entrée d'une maison (*attention au chien*) pour désigner le chien de la maison. Ces utilisations sont appelées *utilisation en situation visible* et *utilisation en situation immédiate* par Hawkins, *utilisations pragmatiques* par Löbner ou *utilisations évoquées situationnellement* par Prince [Hawkins, 1978, Löbner, 1985, Prince, 1981].

Pour les secondes, les expressions qui se rapportent au contexte général de la communication, comme *la mariée* pendant un mariage, ou *Le premier Ministre*, *le temps*, *l'échafaud*, le référent n'est pas forcément visible, ou présent physiquement dans la situation de communication, mais sont des objets connus, identifiables uniquement par les deux locuteurs au moment où se passe leur échange. Ces utilisations sont appelées *utilisation en situation plus large* par Hawkins, ou *utilisations inférées situationnellement* par Prince.

Utilisations non familières Les utilisations non familières sont aussi appelées *containing inferrable* par Prince. Il s'agit de descriptions définies permettant d'identifier uniquement le référent parce qu'elles en donnent une définition complète, ou en lien avec un autre référent connu. Les exemples donnés par Vieira sont les suivants : *le fait que..., la couleur rouge, la femme avec qui Bill sort, le fond de la mer, les mêmes secrets, la première personne qui a....* Pour [Corblin, 1987], tous les noms modifiés par des génitifs ou des relatives restrictives peuvent être utilisés en première mention. Nous considérons qu'ils font partie de cette catégorie d'utilisation du défini.

1.2.1.2 Démonstratif en première mention

Selon Wiederspiel, le démonstratif peut avoir deux types d'utilisation en première mention : la première est appellée exophore a-mémorielle, la seconde est l'exophore mémorielle [Wiederspiel, 1994].

Exophore a-mémorielle La première utilisation, la plus largement reconnue et rencontrée, est ce que [Wiederspiel, 1994] appelle *exophore a-mémorielle* ou *deixis in praesentia*. Il s'agit de cas identiques aux utilisations situationnelles (situation visible ou immédiate) pour les définis, c'est à dire que le référent est présent physiquement dans la situation de communication. Généralement, ces emplois du démonstratif sont accompagnés

d'un geste. Ainsi, on peut imaginer une situation où le locuteur entre dans l'appartement de l'interlocuteur, désigne un fauteuil et dit :

(31) *J'aime bien ce fauteuil.*

Exophore mémorielle La seconde utilisation du démonstratif en première mention est probablement plus rare, et elle est appelée par Wiederspiel *exophore mémorielle* ou *deixis in absentia*. L'expression référentielle renvoie alors à un objet présent uniquement en mémoire du locuteur et présuppose la connaissance de l'objet par son interlocuteur. Wiederspiel donne pour ce type d'emploi l'exemple suivant :

(32) *Je me souviens, tu sais, de ce dîner à Etretat.*

1.2.2 Utilisations coréférentielles directes

1.2.2.1 Définis en reprises directes

La littérature anglo-saxonne ne distingue pas systématiquement les emplois coréférentiels directs (avec la même tête nominale dans l'antécédent et dans l'anaphore) des emplois coréférentiels indirects (avec une tête nominale différente dans l'antécédent et l'anaphore). Nous choisissons de les distinguer afin de cerner le plus précisément possible le phénomène de reprise par un syntagme nominal défini. Fraurud, Hawkins et Prince ne font pas la distinction, et parlent respectivement d'emploi en mention subséquente, d'emplois anaphoriques ou de référents *évoqués textuellement (textually evoked)* [Fraurud, 1990, Hawkins, 1978, Prince, 1981]. Clark, Sidner et Strand différencient les reprises directes des autres, et parlent de relation d'identité, de co-spécification ou de coréférence avec une tête identique [Clark, 1977, Sidner, 1979, Strand, 1997]. La littérature française distingue trois types de reprises directes (ou fidèles, les deux termes sont employés de façon strictement synonymes) :

- les reprises totalement fidèles : l'antécédent et la reprise contiennent le même nom et les mêmes modifieurs,
- les reprises directes par un défini nu : la reprise ne comporte pas de modifieurs et utilise le même nom tête que la première mention,
- les reprises par un défini modifié : la reprise et la première mention ont la même tête nominale, mais les modifieurs varient d'une mention à l'autre.

Reprise totalement fidèle La notion de reprise fidèle est la même que la notion de reprise directe, le nom tête du syntagme doit être le même que dans l'antécédent. Une reprise totalement fidèle est une reprise qui non seulement répète le nom contenu dans l'antécédent, mais répète aussi l'intégralité des modifieurs. [Theissen, 2001] présente une série d'emploi du défini en reprise totalement fidèle. Elle note que quasiment systématiquement, la reprise fidèle sans la répétition des modifieurs est possible. Il semble que la justification des cas de reprise totalement fidèle soit en partie informationnelle : si la propriété dénotée par les modifieurs a une valeur dans l'argumentation, ou est importante à souligner dans la prédication où apparaît la seconde mention du référent, la reprise totalement fidèle est possible. Cependant, elle admet que cette explication n'est pas suffisante et que la relation

entre le nom et les modifieurs est à prendre en compte. Notons par ailleurs que son étude ne porte que sur des noms modifiés par des adjectifs.

Reprise par un défini nu Pour Corblin, *le défini est indépendant de la notion de reprise* [Corblin, 1987]. En effet, un défini n'est pas systématiquement interprété comme un syntagme anaphorique. Pourtant, la seconde mention d'un objet par une description définie sans modifieurs (ou défini nu) peut être interprétée comme une reprise (exemple 33).

Reprise par un défini modifié Si les définis nus (i.e les définis non modifiés) peuvent être interprétés comme des reprises d'une mention antérieure, Corblin montre qu'il est difficile de faire une telle interprétation si on ajoute des modifieurs dans la seconde mention. Ainsi, l'exemple 33 peut être interprété comme illustrant la capacité de reprise, tandis que l'exemple 34 ne le permet pas, car il est difficile d'établir une relation de coréférence entre les deux syntagmes.

(33) *Cette rose (rouge) me gêne. Je vais jeter la rose.*

(34) **Cette rose rouge me gêne. Je vais jeter la rose fanée.*

L'explication de Corblin est alors que le défini ne renvoie pas à une mention antérieure, mais à un domaine d'interprétation (un fragment de discours antérieur), et est interprété comme *le N relativement à C* où C est le contexte. Nous verrons plus loin (chapitre 2) que ceci est très proche de l'interprétation que font [Gardent et Striegnitz, 2003] du défini, en utilisant la notion d'ancre.

La reprise par un défini nu est donc attestée en français. Cependant, nous avons vu, avec ce que Kleiber appelle le paradoxe de la reprise immédiate, que la reprise par un défini nu n'est pas possible dans tous les contextes, si l'antécédent est un indéfini. Par ailleurs, la reprise par un défini modifié et dont les modifieurs sont différents de ceux de l'antécédent paraît impossible. Corblin n'exclut pas qu'il existe des cas où il soit possible d'établir la coréférence entre les deux syntagmes, mais en revanche, exclut le fonctionnement anaphorique de ce type de reprise. Cette constatation reviendrait à dire que l'apport d'information sur le référent dans la reprise n'est pas possible.

Nous verrons dans le chapitre 7 que l'ajout d'information dans une reprise définie est attesté en corpus, et qu'il est possible de produire une description définie en seconde mention avec des modifieurs qui ne sont pas présents dans la première mention.

1.2.2.2 Démonstratifs en reprises directes

Pour Corblin, le statut de reprise d'une mention est définitoire du démonstratif. Kleiber [Kleiber, 1986] affirme que *« l'adjectif démonstratif se révèle être un vrai connecteur anaphorique »*. Pour Kleiber, la reprise fidèle démonstrative a un statut déictique. Le démonstratif permet de référer à un objet présent dans le contexte linguistique, et peut être équivalent à un déictique désignant un objet du contexte extra-linguistique. Le choix du démonstratif dans la reprise fidèle relèverait alors autant de son caractère déictique que de la notion « d'opposition interne à la classe ».

1.2.2.3 Synthèse

La reprise directe est attestée avec les deux déterminants. Le problème est donc de savoir exactement dans quelles conditions apparaissent les deux déterminants. Malheureusement, ils n'apparaissent pas exactement en distribution complémentaire dans ces configurations (cf. exemples 27 et 28). Les critères donnés par Kleiber et Corblin, les deux principaux chercheurs ayant travaillé sur le problème en français, sont les suivants :

- Le défini apparaît lorsqu'on peut établir un contraste avec un autre référent du contexte, c'est à dire quand le nom utilisé dans la description permet d'opposer le référent aux autres référents potentiels du contexte.
- Le démonstratif est utilisable à chaque fois, sauf dans les cas où le référent est explicitement opposable à un autre de par sa catégorie (i.e. quand l'ensemble des entités du contexte forment un ensemble hétérogène du point de vue de la classe des objets, cf. par exemple les cas de coordination 23 à 26 où le défini est nettement meilleur, mais d'autres configurations sont possibles).
- Le démonstratif est meilleur que le défini dans les cas où le prédicat associé à la reprise anaphorique constitue une rupture dans la continuité événementielle.

1.2.3 Utilisations coréférentielles indirectes

1.2.3.1 Reprise définie indirecte

La reprise définie infidèle est assez peu mentionnée dans la littérature française. Pour [Corblin, 1987], il semble que seuls quelques noms de qualité puissent être employés en reprise indirecte avec le défini (i.e. *le salaud, l'imbécile, le jeune homme...*). La reprise indirecte du référent par une description définie donne alors lieu à deux types d'interprétations :

- La reclassification est interprétée comme occasionnelle si la reprise occupe la fonction de sujet grammatical dans la phrase.
- La reclassification est interprétée comme permanente si la reprise occupe une autre fonction grammaticale que la fonction sujet.

Dans la littérature anglo-saxonne, on trouve de nombreuses mentions de la reprise définie indirecte : Fraurud, Hawkins et Prince ne font pas de distinction particulière entre reprise définie fidèle ou infidèle, mais Clark et Strand en font plusieurs. Pour Clark, une reprise définie indirecte peut être une *pronominalisation*, c'est-à-dire un usage où le nom employé dans la reprise est plus générique que son antécédent (*un ordinateur ... la machine*), ou un usage où la reprise donnera le type d'un objet mentionné en premier lieu par un nom propre (*Pinkerton Inc... l'entreprise*) ; il s'agira d'un épithète quand les noms n'auront pas de relation sémantique (*Un homme ... le salaud*). Pour Strand, il s'agira respectivement de cas de *généralisations* quand la reprise est un hypéronyme de l'antécédent et de *redescriptions* si l'antécédent est un nom propre ou s'il n'y a pas de relation sémantique entre les deux noms. Strand ajoute à ces emplois des *spécifications*, quand la reprise est un hyponyme de l'antécédent (*la voiture ... la berline*) et des *élargissements* dans un cas comme : *un homme et une femme... le couple.*

Les reprises définies indirectes apparaissent dans le même type de contexte que les reprises directes. Un moyen de les classer est d'utiliser la relation lexicale qui unit le nom contenu dans le syntagme de reprise et le nom servant d'antécédent. Même s'il existe des cas où il n'y a pas de relation lexicale connue entre l'antécédent et l'anaphore, il semble important (contrairement au démonstratif) de pouvoir établir un lien entre les deux syntagmes. Ce lien passe soit par le lexique, soit par des connaissances du monde dans le cas des redescriptions et des épithètes par exemple. Il semble tout de même que l'utilisation des connaissances du monde soit plus restreinte avec le défini qu'avec le démonstratif, dans la mesure où le défini semble permettre moins d'utilisations métaphoriques par exemple.

1.2.3.2 Reprise démonstrative indirecte

Pour Corblin, la référence virtuelle du syntagme n'est pas prise en compte dans l'interprétation d'un syntagme démonstratif [Corblin, 1987]. Ceci lui confère alors une capacité de reclassification importante, c'est à dire qu'on peut re-nommer un antécédent pratiquement à volonté, à partir du moment où l'on force la coréférence en utilisant le déterminant démonstratif. Il faut bien entendu que l'interlocuteur arrive à reconstituer un lien entre l'antécedent et l'anaphore pour qu'il puisse accepter et comprendre l'énoncé. Les exemples donnés par Corblin sont les suivants : métaphores (35), ou attribution de propriétés (36). Corblin ne précise pas un point qui sera important pour nous dans le chapitre 7 : l'attribution de propriété doit-elle consister en un apport d'information ? En effet, dans un exemple comme 36, on peut penser que *être un admirateur de classiques* est une propriété de Pierre nouvelle pour l'interlocuteur, mais ça n'est pas nécessairement le cas. Nous considérerons donc le terme *reclassification* dans son acception la plus large possible, c'est à dire comme une *reprise indirecte, n'entretenant pas de relation lexicale avec l'antécédent et apportant ou n'apportant pas d'information nouvelle sur le référent.*

(35) ***Deux arbres*** *encadraient l'entrée et* ***ces sentinelles*** *dormaient.*

(36) ***Pierre*** *s'est confié à moi.* ***Cet admirateur de classiques*** *était très lucide.*

Un autre problème qui nous semble laissé en suspens par Corblin est le suivant : le démonstratif semble forcer la coréférence et permettre ainsi à l'interlocuteur d'établir un lien de coréférence entre deux groupes nominaux. Cependant, quand le lien n'est pas purement lexical, et quand plusieurs antécédents sont possibles, on ne sait pas comment l'interlocuteur peut choisir entre les antécédents concurrents. Des paramètres de focalisation ou de récence de l'antécédent entrent-ils en jeu ?

Considérons les exemple suivants (inspirés d'un extrait du corpus) :

(37) ***Nelson Picket*** *a doublé Ayrton Senna dans le dernier tour du Grand Prix.* ***Ce Brésilien de 25 ans*** *est un champion en devenir.*

(38) *Nelson Picket a doublé* ***Ayrton Senna*** *dans le dernier tour du Grand Prix.* ***Ce Brésilien de 35 ans*** *devrait s'inquiéter de l'arrivée de jeunes sur le circuit.*

Dans l'exemple 37, il semble fort probable que l'antécédent de l'expression *Ce Brésilien de 25 ans* soit Nelson Picket. En revanche, dans l'exemple 38, l'antécédent est plus

probablement Ayrton Senna. Ici, deux noms propres sont en concurrence pour être identifiés comme antécédent d'une reprise, et seules les connaissances du monde et les prédicats impliqués dans les deux phrases permettent de désambiguïser la relation anaphorique. Il existe donc des paramètres supplémentaires pour établir le lien entre une description démonstrative et son antécédent, qui ne sont pas forcément des paramètres de saillance ni de récence, puisqu'avec une même phrase introductive, l'interprétation d'une description peut changer en fonction du verbe avec lequel elle est employée.

Enfin, on note des emplois appelés *simplification de l'antécédent* chez Wiederspiel, pour qui la tendance générale de l'anaphore est à la simplification de l'antécédent. La simplification de l'antécédent peut être de deux types : elle peut être conceptuelle, si le terme employé dans la reprise est plus générique que celui employé dans l'antécédent, ou formelle, si l'antécédent est un groupe verbal, une proposition et la reprise un groupe nominal. Parmi les exemples qu'elle donne, voici celui qui nous semble le plus représentatif. Notons que nous excluerons ce type d'exemple de notre étude dans un deuxième temps, l'antécédent du SN anaphorique n'étant pas nominal :

(39) *La jeune fille reconnaissante* ***changeait de chambre chaque jour****. Elle devait penser que* ***ce nomadisme*** *la préservait du péché d'infidélité. (Orsenna, 'Grand amour', cité dans [Wiederspiel, 1994] p.171)*

1.2.3.3 Synthèse

Alors que les reprises définies indirectes sont fréquentes, elles ne sont pas étudiées distinctement des reprises définies directes. Les contraintes pesant sur l'utilisation du défini semblent alors être les mêmes quels que soient ses emplois (extraction d'un référent en utilisant une propriété qui l'oppose aux autres référents potentiels du contexte). Le démonstratif semble lui avoir deux types d'emplois. Le premier emploi est purement anaphorique, c'est dans ce cas qu'on l'emploie en reprise directe. Par ailleurs son caractère purement anaphorique lui confère un emploi de re-classification des référents, c'est à dire qu'il permet de re-nommer les référents, en utilisant un nom sans lien lexical avec le premier, et ceci pour deux raisons : la référence virtuelle du nom utilisé n'a que peu d'importance pour l'interprétation du syntagme nominal, et la coréférence est forcée par sa présence, bien que des paramètres de focalisation de l'antécédent et les connaissances encyclopédiques entrent en jeu dans de nombreux cas.

1.2.4 L'anaphore associative

L'anaphore associative est un phénomène référentiel étudié depuis [Clark, 1977]. Il s'agit de cas où l'anaphore et l'antécédent ne sont pas coréférents. Un des exemples les plus cités est le suivant :

(40) *Nous arrivâmes dans* ***un village****.* ***L'église*** *était située sur une hauteur.*

Dans cet exemple, on interprète spontanément l'église comme étant l'église du village. Sans la mention antérieure du village, on ne pourrait pas utiliser le défini, et identifier le référent du syntagme *l'église*. L'anaphore associative est donc le phénomène qui permet

d'introduire un référent nouveau en s'appuyant sur un antécédent. Ce référent nouveau doit entretenir une relation précise avec son antécédent, qui provient en général des connaissances encyclopédiques des locuteurs. La relation prototypique impliquée dans l'anaphore associative est la relation de méronymie (appelée aussi relation partie-tout), mais elle n'est pas la seule. Pour plus de détails sur les relations associatives, nous renvoyons le lecteur au chapitre 6.

Du point de vue de l'utilisation des déterminants, le phénomène d'anaphore associative est intéressant : en effet, on a longtemps considéré que l'anaphore associative n'était possible qu'avec une description définie. Pourtant, on trouve dans [Gundel et al., 2000] et [Apothéloz et Reichler-Béguelin, 1999] des mentions d'anaphores associatives avec le démonstratif. On peut citer l'exemple suivant, emprunté par [Apothéloz et Reichler-Béguelin, 1999] à [Charolles, 1990] :

(41) *Nous arrivâmes dans* ***un village****.* ***Cette église****, tout de même, quelle horreur !*

La tournure exclamative permet ici de réaliser une anaphore associative avec le démonstratif. Il semble qu'une partie de ces utilisations associatives du démonstratif ait pour fonction de créer un effet d'empathie ou de traduction de pensées, ou pour éviter de mauvaises interprétations [Apothéloz et Reichler-Béguelin, 1999]. Nous n'irons pas plus loin sur les anaphores associatives et sur les raisons de l'utilisation des démonstratifs dans ce cas, car de nombreux exemples donnés par [Apothéloz et Reichler-Béguelin, 1999] nous semblent être plutôt des exemples de simplification des antécédents que de réels cas d'anaphore associative. Notons simplement que ces cas existent, et qu'ils doivent être pris en compte dans une étude des phénomènes référentiels impliquant des descriptions définies et démonstratives.

1.3 Etudes de corpus

Les travaux que nous venons d'étudier sur le déterminant tentent de classer les utilisations des déterminants et d'en définir les principes de fonctionnement. Selon nous, ils présentent essentiellement un défaut : ils ne sont pas validés sur corpus. En effet, la plupart des exemples donnés sont construits, et il est parfois difficile de juger de l'acceptabilité de certains de ces exemples. Valider ces théories sur corpus semble alors un bon moyen de lever les doutes. Par ailleurs, si certains phénomènes sont remarquables par leur fonctionnement, il semble, dans une perspective de génération, indispensable d'en connaître la fréquence d'apparition. Dans un premier temps, il nous semble que nous ne devons pas viser la génération de tous les syntagmes nominaux trouvés en corpus. Si la génération des phénomènes les plus fréquents est possible, nous considérerons que notre but est atteint. Nous présentons donc dans cette section les études de corpus déjà menées dans le domaine des descriptions définies et démonstratives.

1.3.1 Le défini

Les principaux travaux sur les descriptions définies fondés sur des corpus importants en traitement automatique des langues sont les travaux de Poesio et Vieira [Poesio et Vieira, 1998].

Ils ont été effectués dans le but de déterminer si une annotation des groupes nominaux définis est possible en fonction de leur interprétation, sur un corpus comprenant uniquement des articles de presse. Les constats faits par Poesio et Vieira sont les suivants (ils n'utilisent que les classifications données par la littérature anglo-saxonne présentées dans le début de ce chapitre) :

Tout d'abord, les annotateurs ne sont pas d'accord entre eux sur la classification des descriptions définies quand ils doivent les classer selon les classifications de Hawkins et Prince. En revanche, le taux d'accord est beaucoup plus important avec la classification de Fraurud, qui ne comporte que deux classes, les premières mentions et les mentions subséquentes. Par ailleurs, le taux d'accord sur l'expression servant d'antécédent n'est pas bon. Ceci pose la question intéressante de savoir sur quoi se base exactement un auditeur pour interpréter une expression coréférentielle : quel est le rôle de l'antécédent ? Quels éléments autres permettent la résolution de la coréférence ?

Les travaux de Poesio et Vieira se fondent sur deux expériences : la première utilise les catégories issues de la littérature :

- La catégorie *anaphorique - même tête* correspond à ce que nous appelons ici les reprises directes.
- Les *utilisations associatives* regroupent les utilisations pour lesquelles on reconnaît un antécédent (nominal ou non) à l'expression, relié par des associations lexicales ou des connaissances du monde. Il s'agit donc des reprises indirectes et des anaphores associatives.
- Les *utilisations non familières* et les *utilisations en situation plus large* sont regroupées. Pour une définition, on se reportera à la section 1.2.
- La catégorie *idiomatique* permet de classer les descriptions définies entrant dans des expressions figées.

La seconde expérience a été réalisée suite aux résultats peu satisfaisants de la première. La catégorie *anaphorique - même tête* a été élargie à une catégorie où toutes les reprises sont comprises, que la tête de la reprise soit identique ou non à celle de la première mention. La classe *utilisations associatives* ne change pas, et les auteurs ont décidé de séparer les *utilisations non-familières* des *utilisations en situations plus large* pour voir s'il était difficile ou non de les différencier.

Nous ne nous attarderons pas sur les conclusions qui intéressent Poesio et Vieira (i.e. la possibilité pour des annotateurs d'obtenir un accord sur la catégorie d'une expression), mais sur les résultats proprement linguistiques de l'expérience : dans quelle mesure le défini permet-il de désigner un objet en première mention, ou en mention subséquente ? Quels sont ses emplois majoritaires ? Les résultats en termes de fréquence d'apparition des phènomènes sont donnés dans les tableaux reproduits en figure 1.2 et en figure 1.3 (nous donnons une fourchette de résultats, les résultats variant d'un annotateur à l'autre) :

Ce qui nous paraît intéressant dans ces résultats sont les faits suivants : dans la première expérience, la catégorie majoritaire est celle pour laquelle les descriptions définies n'ont pas d'antécédent. Les reprises directes sont nombreuses, comme les utilisations associatives, mais ne sont pas majoritaires (42 à 47%). En revanche, quand la définition des utilisations coréférentielles devient plus lâche, le nombre de situations repérées augmente, il passe de 49

catégorie	fourchette de pourcentage
anaphorique même tête	28-32
associatif	14-15
non-familier - situation plus large	52-53
idiomatique	0,1-3,75
doute	0,09-0,67

FIG. 1.2 – Expérience 1 - Poesio et Vieira

catégorie	fourchette de pourcentage
coréférentiel	43-45
associatif	6-11
situation plus large	20-25
non-familier	18-26
doute	0-6

FIG. 1.3 – Expérience 2 - Poesio et Vieira

à 56%. Nous ne pouvons pas affirmer que cette différence est significative statistiquement, mais elle peut amener à la conclusion suivante : le problème des descriptions définies n'est peut être pas d'identifier leur antécédent, mais d'être sûr qu'elles ont un antécédent. En effet, avec une définition plus large de la coréférence, la proportion d'utilisations en première mention chute (52-53% dans la première expérience pour 38-52% dans la seconde). Il semble donc que l'interprétation d'une description définie comme coréférentielle n'est pas évidente, et varie d'un lecteur à l'autre.

1.3.2 Le démonstratif

Nous utiliserons dans cette étude des travaux sur corpus concernant le démonstratif : les travaux de [Errenati, 2001] et les travaux de [Vieira et al., 2002]. Ces travaux sont effectués dans des optiques différentes, et nous essaierons, après une brève présentation, d'en extraire des conclusions plus générales.

1.3.2.1 Démonstratif et type du référent

Les travaux de [Errenati, 2001] ont pour but de déterminer si le type du référent (événement, objet, humain, non humain) influence la forme et les possibilités de reprise démonstrative, dans un corpus composé d'articles de presse et de textes littéraires. Ces travaux entrent dans le cadre d'une étude plus vaste menée par [Danlos, 1999, Danlos et Gaiffe, 2000] sur la coréférence événementielle. Ils ne sont donc pas centrés sur notre problématique directement, mais manipulent la notion d'apport d'information qui sera centrale dans notre étude (chapitres 7 et 8).

Ce qui est intéressant pour nous dans le travail de [Errenati, 2001], c'est d'étudier

la répartition des reprises démonstratives selon deux critères : la nature de l'entité à laquelle les syntagmes réfèrent (entité, humain, événement) ou la forme de l'antécédent (nom propre, présentatif) ; l'autre critère de classification est la relation entretenue par l'antécédent et l'anaphore d'un point de vue sémantique.

Nous reproduisons en figure 1.4 les chiffres obtenus dans l'analyse de corpus.

Reprise d'un nom propre	Référent humain Référent non humain	7% 5,3%	12,3%
Emploi présentatif			12,6%
Reprise d'un référent humain	Même tête tête différente	2,3% 5,6%	9,1%
Reprise d'une entité	Même tête tête différente particularités - métalinguistiques	11,9% 11,1% 9%	31,1%
Reprise d'un événement	propositions Etat Evénements Situation complexe	4,45% 4,7% 16,7% 8,9%	34,8%

FIG. 1.4 – Résultats de l'annotation de Errenati

Nous ne rapporterons pas d'autres chiffres concernant cette étude de corpus. Les constats qui nous semblent importants ici sont les suivants :

- Les démonstratifs servent dans une proportion importante à la reprise de noms propres, surtout quand le référent est humain (12,3%).
- Dans ce corpus, les reprises directes apparaissent en petit nombre. On trouve là une confirmation de la capacité de reclassification du démonstratif.
- Les reprises d'entités ou d'événements semblent en nombre aussi important l'une que l'autre.

Ensuite, [Errenati, 2001] donne d'autres conclusions, basées sur des chiffres contenus dans son mémoire que nous ne reproduirons pas ici : la possibilité d'apport d'information du syntagme de reprise semble dépendre du type du référent. Plus précisément, elle est fréquente lorsque l'antécédent est un nom propre référant à un humain. En revanche, elle est plus rare pour la reprise de noms communs et pour les événements et les propositions.

Enfin, une tentative de permutation des démonstratifs avec le défini a été faite, et donne les résultats suivants : les configurations les plus favorables à la permutation sont les reprises de noms propres et les reprises par hypéronymie alors qu'elle apparaît plus difficile en reprise directe avec perte des modifieurs.

1.3.2.2 Forme syntaxique des reprises et type de l'antécédent

L'étude de [Vieira et al., 2002] porte sur les groupes nominaux démonstratifs en français et en portugais, sur un corpus de questions et de réponses de députés européens aux commissaires européens. Elle s'articule en trois points : tout d'abord, l'étude de la struc-

ture syntaxique des syntagmes démonstratifs, ensuite les types d'utilisation (coréférence directe, indirecte, ou autre), l'analyse des antécédents (leur nature grammaticale essentiellement) et pour finir, la distribution entre référents abstraits et référents concrets des descriptions démonstratives.

Nous ne présenterons pas ici les chiffres présentés par les auteurs, nous nous contenterons de donner leurs conclusions sur le français :

- La plupart des utilisations du démonstratif ne contiennent pas de modifieurs. Ceci est expliqué par les auteurs par la saillance des référents repris par des expressions démonstratives.
- La majorité des emplois coréférentiels ont été classés dans la catégorie *autres*, c'est à dire qu'ils n'ont pas été classés. Ceci serait lié à la difficulté de repérer un antécédent pour la description démonstrative.
- Parmi les emplois classés par les annotateurs, les plus nombreux sont les emplois en reprise directe.
- Les antécédents nominaux sont les plus nombreux, très loin devant les antécédents phrastiques ou sous forme de syntagmes verbaux.
- Les démonstratifs préfixent majoritairement des noms abstraits (par opposition aux noms référant à des objets concrets : le nom *but* est un nom abstrait, alors que le nom *chaise* est concret).
- La relation sémantique majoritaire entre les reprises démonstratives et leurs antécédents nominaux est l'hypéronymie.

Une partie de cette étude a été ensuite menée sur des définis, [Salmon-Alt et Vieira, 2002]. La comparaison entre définis et démonstratifs amène aux conclusions suivantes :

- La prédominance des définis nus est beaucoup moins flagrante que celle des démonstratifs nus.
- La majorité des définis est employée en première mention, et comme pour les démonstratifs, les reprises les plus nombreuses sont les reprises directes.

1.3.2.3 Synthèse

Les études de corpus que nous venons de présenter exhibent toutes des résultats cruciaux dans la compréhension des phénomènes de détermination. Toutes présentent cependant la lacune de ne pas faire de comparaison approfondie entre le défini et le démonstratif : ainsi, pour la génération de textes, peu de critères de choix entre les déterminants sont disponibles. Les études de corpus confirment cependant certaines des données théoriques que nous avons vues précédemment : tout d'abord, le démonstratif est effectivement un déterminant de reprise, qui semble mieux supporter la reprise directe immédiate que le défini. L'hypothèse de simplification du référent est aussi très nette lorsqu'on observe aussi bien le type que la forme des antécédents des descriptions qui réfèrent à des événements sous des formes non nominales. L'étude de [Errenati, 2001] montre aussi la capacité de reclassification du démonstratif, moins visible chez [Vieira et al., 2002] et [Salmon-Alt et Vieira, 2002].

1.4 Contraintes sur l'utilisation des déterminants

Dans une perspective de génération, une étude comparative des déterminants est nécessaire, afin de déterminer leurs conditions d'emploi. Comme nous le verrons au chapitre 2, la génération peut être vue comme une tâche de choix entre des formes différentes. Il est donc nécessaire, pour pouvoir décider de la génération d'un défini ou d'un démonstratif, d'avoir des critères de sélection assez rigides, quitte à ce qu'ils soient un peu trop puissants pour décrire fidèlement la réalité.

Dans cette section, nous allons énumérer une série de contraintes régissant l'utilisation des déterminants, que nous testerons dans la seconde partie de notre thèse sur un corpus, afin de nous assurer que les contraintes énoncées sont suffisantes pour la génération.

1.4.1 Contraintes liées aux principes d'interprétation des déterminants

Première mention et reprise Le démonstratif en première mention est quasiment exclu de la génération de texte non multimodale (i.e. sans gestes ou équivalents de gestes de monstration). En effet, les cas d'exophore mémorielle sont trop rares dans les types de textes générés automatiquement pour être pris en compte. En revanche, le défini est beaucoup trop employé en mention subséquente pour qu'on ignore son utilisation en génération de texte. On peut donc, en traitement automatique de la langue, fonder partiellement le choix du déterminant sur la position de la mention dans la chaîne anaphorique, dans le sens où on peut exclure le démonstratif en première mention.

Unicité du référent Il semble clair que la contrainte d'unicité du référent correspondant à la description dénotée par le syntagme nominal joue différemment avec le défini et le démonstratif, mais reste valable pour les deux. Elle est impérative pour l'utilisation du défini, alors qu'elle ne l'est pas pour le démonstratif. En effet, si la description démonstrative peut convenir pour plusieurs référents, elle réfèrera à l'élément le plus saillant du contexte correspondant à la description (en général le dernier mentionné). Ceci revient tout de même à dire que le référent d'une description démonstrative doit pouvoir être identifié uniquement dans un contexte peut-être plus restreint que le défini. Par ailleurs, dans de nombreux cas d'utilisation du démonstratif, la question de la saillance ne se pose pas, car le démonstratif peut aussi servir à préfixer une description identifiante uniquement dans le contexte. La caractéristique d'unicité n'est donc pas systématiquement suffisante pour le choix du déterminant. Nous nous retrouvons donc face à trois cas possibles, dont deux sont identiques pour l'utilisation du défini et du démonstratif :

1. Le défini préfixe une description uniquement identifiante du référent visé, et est interprété comme « l'unique X correspondant à la description ».
2. Le démonstratif préfixe une description uniquement identifiante du référent visé, et est interprété comme « l'unique X correspondant à la description ».
3. Le démonstratif préfixe une description non uniquement identifiante et est interprété comme « le X correspondant à la description le plus saillant dans le contexte ».

Saillance du référent Une caractéristique du démonstratif semble être que son antécédent doit être saillant. Effectivement, pour une utilisation en première mention, le référent doit être saillant dans la situation de communication. De la même façon, en cas de description non identifiante dans le contexte, le référent de la description démonstrative doit être le plus saillant. Cependant, la notion de saillance n'est pas pertinente dans le cas de la reprise immédiate. Ceci peut sembler vrai dans les exemples 21 à 26, en revanche, cette explication ne vaut pas pour les exemples 27 à 30.

Description attributive Si la description n'a pas de référent identifié, on dit que c'est une description attributive par opposition à une description référentielle [Donnellan, 1966]. Seules les descriptions définies ont cette capacité de décrire des référents non identifiés strictement. On peut donc dire que l'une des contraintes d'utilisation du démonstratif est que le référent doit obligatoirement être identifié par le locuteur.

1.4.2 Contraintes liées à la forme de la reprise

Une reprise est la plupart du temps directe Les données provenant des analyses de corpus montrent la prédominance des reprises directes sur toutes les autres. Les analyses théoriques le montrent aussi, parce qu'elles traitent essentiellement de la reprise directe. Pour autant, la reprise indirecte n'est pas rare quel que soit le déterminant étudié. Les études de corpus citées précédemment ne donnent pas de points de comparaison entre les reprises indirectes définies et démonstratives, tandis que les analyses théoriques ne donnent pas d'explication totalement satisfaisante sur l'utilisation des déterminants en reprise indirecte. En effet, Corblin affirme que le démonstratif permet d'attribuer des propriétés au référent par le biais de la reprise indirecte, mais les études de corpus prouvent que cela est possible aussi avec le défini. Les utilisations du démonstratif en reprise indirecte avec un nom propre pour antécédent relevées dans [Errenati, 2001] nous paraissent intéressantes, mais ne donnent pas de point de comparaison avec le défini.

Les modifieurs Il existe peu d'études sur la présence des modifieurs dans les reprises. Si l'on sait que certains modifieurs sont utilisés avec le défini pour créer des descriptions uniquement identifiantes (relatives restrictives, génitif), on a peu de données sur les modifieurs présents dans les reprises. Nous avons vu que les reprises définies totalement fidèles sont possibles dans certains cas seulement, mais ne semblent jamais indispensables. Le problème est alors de savoir quel est le contenu des modifieurs dans les reprises, et quelles sont les contraintes régissant l'ajout de modifieurs dans les reprises. De la même manière, on peut se demander si le contenu et la forme grammaticale des modifieurs ont une influence sur le choix du déterminant.

Dans cette section après une revue de l'état de l'art sur la référence, et particulièrement sur les phénomènes de référence impliqués lors de la production de descriptions définies et démonstratives, nous avons établi une série de contraintes sur l'utilisation des déterminants. Nous avons déjà vu dans ce chapitre que la plupart des études menées dans

ce domaine étaient dirigées par des perspectives de compréhension du langage et non de génération.

S'il est clair pour nous que nombre des conclusions de ces études sont réutilisables en génération, elles ne revêtent cependant pas toutes la même importance relative en analyse qu'en génération. Pour nous, les contraintes liées à la forme que prennent les expressions référentielles seront sans doute plus déterminantes dans le processus de génération qu'elles ne le sont dans le processus d'analyse. Par ailleurs, il nous faut les formaliser de façon à ce qu'elles soient utilisables en génération.

Dans le chapitre suivant, nous présentons les travaux réalisés en génération d'expressions référentielles, afin de cerner la problématique qui nous occupe. Nous verrons quels sont les paramètres fondamentaux pris en compte en génération, et ceux qui ne le sont pas encore, afin de déterminer les éléments manquant à l'étude des expressions référentielles pour en permettre la génération automatique.

Chapitre 2

La génération d'expressions référentielles

La génération de textes est une tâche qu'on peut considérer comme la conversion de données non linguistiques en texte. Il s'agit par conséquent d'une tâche complexe que nous présentons dans ce chapitre. Nous mettrons en avant la génération d'expressions référentielles, qui est la sous-tâche du processus global de génération sur laquelle porte cette thèse. Nous présenterons en premier lieu ce qu'est la génération de textes, en montrant les différences entre génération et compréhension de textes (section 2.1.1). Nous aborderons ensuite l'architecture des systèmes de génération (section 2.1.2), puis plus particulièrement le module appelé *module de micro-planification* dans lequel a lieu la génération d'expressions référentielles (section 2.1.3). Nous poursuivrons dans les sections suivantes en présentant les algorithmes de génération d'expressions référentielles existants (section 2.2 et 2.3).

2.1 La génération automatique de textes

2.1.1 Génération et compréhension de textes

La génération et la compréhension de textes se différencient tout d'abord du point de vue de la démarche adoptée. Bien qu'en apparence, elles semblent être des problèmes inverses (comment passer d'un contenu à sa réalisation pour la génération ou comment passer d'une réalisation linguistique à son interprétation pour la compréhension), on montre très vite que ce n'est pas le cas. La compréhension peut être vue comme une tâche de gestion d'hypothèses alors que la génération se problématise plutôt en termes de choix entre des possibilités [Reiter et Dale, 2000]. La génération de textes doit répondre à la question suivante : *Etant donné les divers moyens à disposition pour parvenir au but fixé, lesquels doivent être utilisés ?* La plus grosse différence entre la génération et la compréhension est liée aux données linguistiques traitées. Autant la compréhension doit pouvoir tenir compte d'énoncés complexes et parfois même mal formés, autant la génération peut se contenter de traiter de textes simples, et en tous cas, les textes générés peuvent être plus simples que les textes compris. Le problème principal en génération est donc posé en deux termes simples [Danlos, 1985, Danlos, 1987, Levelt, 1989] :

– Quelle est l'information à communiquer ? - *Quoi dire ?*
– Comment générer un texte contenant cette information qui soit compréhensible pour un être humain ? - *Comment le dire ?*

Après cette courte présentation des différences entre génération et compréhension de textes, nous présentons plus en détail comment fonctionne un système de génération.

2.1.2 Architecture d'un système de génération

Pendant longtemps, on a considéré qu'un système de génération pouvait se composer de deux modules, appelés respectivement *Quoi dire ?* et *Comment le dire ?* [Danlos, 1985]. Plus tard, les recherches ont montré que le processus était plus complexe, et en particulier qu'il devait pouvoir y avoir des allers-retours entre ces modules.

Les systèmes de génération se doivent d'être modulaires, au moins pour des raisons techniques (facilités de débogage, facilitation du travail en équipe et réutilisation des modules d'une application à une autre). Ces modules doivent bien entendu être interfacés de façon à pouvoir communiquer entre eux. Le consensus sur ce que devrait être l'architecture d'un système de génération n'est pas encore complet. Nous présenterons ici la vision de E. Reiter et R. Dale, exposée dans leur dernier ouvrage [Reiter et Dale, 2000].

2.1.2.1 Entrée et sortie d'un système de génération

Le processus de génération de langue naturelle est souvent vu comme un processus dirigé par des buts communicatifs. En effet, pour produire un texte, il faut connaître le but de la communication, c'est à dire le but que veut atteindre le locuteur en produisant l'énoncé. En entrée du générateur, plusieurs choses sont donc nécessaires :

Base de connaissances Pour produire un énoncé à propos de quelque chose, il est nécessaire d'en avoir une représentation. Celle-ci va servir à faire des descriptions, à donner des informations, etc. Il semble en effet difficile de pouvoir produire des énoncés sur des éléments qu'on ne connaît pas. Les savoirs nécessaires sont de trois types : en premier lieu le système a besoin de connaissances linguistiques pour produire du langage ; ensuite, il est nécessaire qu'il ait des connaissances encyclopédiques, c'est à dire des connaissances générales sur le monde ; enfin, le système doit aussi avoir des connaissances du domaine qui concerne l'application dont il fait partie. La base de connaissances du système peut être vue comme une base de données, dont le contenu dépend de l'application, particulièrement en ce qui concerne les connaissances du domaine.

Buts communicatifs Il s'agit, nous l'avons vu, de pouvoir formaliser le but de la communication, c'est à dire la raison pour laquelle l'énoncé est produit. Les raisons peuvent être multiples : faire faire une action, donner une information, ou demander une information à l'utilisateur du système.

Modèle de l'utilisateur Le modèle de l'utilisateur, comme la base de connaissances, est lui aussi dépendant du système. Certains systèmes n'en possèdent pas. Ce type de base

de connaissances permet au système d'avoir une représentation de ce que l'utilisateur sait, afin de ne donner que les informations nécessaires, ou de ne demander que des actions que l'utilisateur peut faire.

Historique du discours L'historique du discours permet au système de garder une trace de ce qui a été dit. Par exemple, il peut garder en mémoire les entités qui ont déjà été mentionnées dans le discours, ce qui est fondamental pour la génération d'expressions référentielles. En effet, on ne désigne pas de la même façon une entité dont on n'a pas encore parlé et une entité dont on a déjà parlé (différence entre expression référentielle en première mention ou en mention anaphorique ou coréférentielle).

2.1.2.2 Survol de l'architecture

Historiquement, on a considéré que les systèmes de génération devaient avoir une architecture minimalement bipartite, avec un module de détermination du contenu (*Quoi dire ?*) et un module de réalisation linguistique du contenu (*Comment le dire ?*). Depuis plusieurs années maintenant, un accord a abouti à une architecture générique tripartite [Reiter et Dale, 2000, Danlos et Roussarie, 2000].

Cette architecture tripartite est composée de trois modules : le module de planification, le module de microplanification, et le module de réalisation, que nous décrivons maintenant.

Module de planification du document Ce module a pour tâche la détermination du contenu, et la structuration du document. Il doit décider de ce qui est dit dans le document produit par le système, et de la façon dont doivent être regroupées les informations dans le document.

Module de microplanification ou de planification de phrase Ce module est le module intermédiaire entre le *Quoi dire ?* et le *Comment le dire ?*. Certaines décisions relèvent en effet à la fois des deux modules. Parfois, certaines décisions sur le regroupement des informations peuvent relever de connaissances linguistiques. C'est pourquoi ce module contient les sous-modules de lexicalisation (choix des mots), de génération d'expressions référentielles (choix de l'expression utilisée pour désigner un objet) et d'agrégation (choix de la façon dont on regroupe les informations en phrases ou en paragraphes, choix de l'ordre d'apparition des informations). Nous reviendrons plus en détails dans la section suivante sur la génération d'expressions référentielles en elle-même.

Module de réalisation Ce module est celui qui décide finalement de la façon dont les informations vont être réalisées linguistiquement. Il convertit les représentations abstraites en phrases et les structures abstraites en paragraphes.

La plupart du temps, les trois modules sont organisés en *pipeline*. L'architecture en pipeline est l'architecture la plus répandue parmi les systèmes existants. Ces architectures sont unidirectionnelles, c'est à dire que chaque module traite les données à son tour, et aucun

retour sur les décisions prises dans un module intervenant antérieurement n'est possible. De nombreux travaux ont cependant contesté ce type d'architecture, arguant que des interactions étaient possibles entre certains choix de détermination du contenu, et certains choix de réalisation de ce contenu. Les architectures proposées contiennent alors des modules interdépendants, [Danlos et Roussarie, 2000]. Certaines propositions d'architectures sont encore plus radicales dans leur opposition au modèle en pipeline. Les systèmes ne sont pas modulaires du tout, et tous les processus de décision sont vus comme des satisfactions de contraintes. Un seul composant de raisonnement est utilisé pour trouver une solution. Ainsi, on ne suppose pas un ordonnancement des tâches à accomplir. On peut citer comme exemple de cette approche le système KAMP, proposé par [Appelt, 1985], qui utilise une représentation uniforme et des mécanismes de raisonnement pour générer des textes courts qui satisfont plusieurs buts communicatifs.

2.1.3 Microplanification et génération d'expressions référentielles

La tâche qui nous concerne dans cette thèse est la tâche de génération d'expressions référentielles. Le monde (et les représentations du monde) sont constitués d'objets, dont on veut parler (auxquels on veut faire référence). Pour référer aux objets, on produit des expressions référentielles, qui permettent à l'interlocuteur d'identifier précisément les entités dont on parle.

Il y a en effet de nombreuses façons de référer à un objet. Tout d'abord, en première mention, il peut y avoir plusieurs *perspectives* pour référer à un objet. Pour reprendre l'exemple donné par [Reiter et Dale, 2000], une même personne peut être vue comme un linguiste, un visiteur venu d'outre-atlantique, ou un adepte du tai-chi. Le choix du mode de présentation utilisé pour référer à un objet va donc dépendre des buts communicatifs. En mention subséquente, il va falloir désigner le référent de façon à ce qu'il soit distingué sans ambiguïté parmi toutes les autres entités mentionnées dans le discours. La forme que va prendre l'expression référentielle va donc dépendre hautement du contexte, c'est à dire de l'historique du discours du point de vue de son contenu (combien d'entités mentionnées, quel type d'entité a déjà été mentionné), et de la forme prise par le discours antérieur.

Ces éléments expliquent pourquoi la génération d'expressions référentielles est située dans le module de microplanification, le module à la frontière entre la détermination du contenu et la réalisation de surface. En effet, nous le verrons tout au long de cette thèse, le problème de la génération d'expressions référentielles fait appel à la fois à des problèmes de détermination du contenu : quelle information doit-on donner/répéter lorsqu'on génère une expression référentielle, pour que l'interlocuteur identifie l'objet désigné sans ambiguïté ? et à des problèmes de réalisation : quel type de structure doit-on utiliser (un pronom est-il suffisant, un groupe nominal est-il nécessaire, avec quel déterminant ...) ?

2.2 L'algorithme standard de Dale and Reiter

L'algorithme de Dale et Reiter [Dale, 1992, Dale et Reiter, 1995] repose sur des principes fondamentaux qui n'ont pas été remis en question depuis. Dans cette section, nous

présenterons dans un premier temps ces notions, et dans un deuxième temps l'algorithme qui a découlé de leurs travaux.

2.2.1 Une notion fondamentale : la notion de description distinguante / identifiante

2.2.1.1 Attention, adéquation, efficacité

L'algorithme de Dale et Reiter repose sur des principes proches des maximes de Grice [Grice, 1975]. Ces principes sont les principes d'attention (*sensitivity*), d'adéquation et d'efficacité.

Le principe que nous appelons ici principe d'attention affirme que pour construire une expression référentielle, le locuteur doit tenir compte de ce que son interlocuteur sait, et utiliser un langage qu'il connaît et des expressions qu'il comprend. Par exemple, le locuteur doit savoir si l'entité dont il parle a déjà été mentionnée dans le discours, et utiliser en conséquence le déterminant défini ou indéfini.

Le deuxième principe est le principe d'adéquation, qui veut que l'expression référentielle permette d'identifier de façon non ambiguë le référent. En d'autres termes, l'expression doit apporter suffisamment d'information pour que l'interlocuteur identifie le référent visé.

Enfin, le principe d'efficacité vise à ce que l'expression référentielle n'apporte pas plus d'information que nécessaire pour l'identification du référent. En effet, une description plus précise que nécessaire peut amener l'interlocuteur à faire des inférences inutiles et à se construire une représentation fausse du domaine.

Ces principes sont donc effectivement très proches des maximes de Grice : on reconnaît les maximes de pertinence pour le premier, de qualité pour le second, de quantité et de manière pour le dernier. Ils amènent par ailleurs naturellement aux notions de *pouvoir discriminant (discriminatory power)*, et de *description distinguante (distinguishing description)* décrites par Dale dès 1992 [Dale, 1992].

2.2.1.2 Description distinguante et pouvoir discriminant d'une expression

Afin de distinguer une entité x des autres entités présentes dans le contexte, on doit utiliser un ensemble de propriétés de x qui permettent de la décrire, et qui la différencient des autres entités du contexte. Une description distinguante est alors une expression référentielle qui exprime assez de propriétés distinctives de l'entité x pour que l'interlocuteur puisse l'identifier, comme le préconise le principe d'adéquation. Le principe d'efficacité impose à la description distinguante utilisée d'être la description distinguante minimale, c'est à dire la description dénotant le plus petit ensemble de propriétés nécessaires à l'identification de x dans le contexte.

Les propriétés dénotées par l'expression référentielle doivent être celles qui ont le plus grand pouvoir discriminant, c'est à dire celles qui, une fois exprimées, éliminent le plus grand nombre d'entités du contexte parmi les référents potentiels de l'expression référentielle.

Il peut bien entendu résulter de ce raisonnement plusieurs expressions référentielles dont les capacités distinguantes sont équivalente. Ceci pose un ensemble de problèmes

complexes sur lesquels nous ne reviendrons pas, et qui sont les suivants :

- le but communicatif doit permettre de déterminer quelle est l'expression la plus adaptée [Reiter, 1990],
- les propriétés doivent être sélectionnées selon un certain ordre, déterminé par des données psychologiques - on sait par exemple que la couleur d'un objet est une propriété plus importante que sa taille [Reiter et Dale, 1992],
- les propriétés doivent être ordonnées entre elles en fonction du contexte.

2.2.2 Algorithme de génération

L'algorithme prend en entrée un référent cible t et un ensemble de faits décrivant t et les autres entités présentes dans le contexte. Il doit construire une expression qui identifie sans ambiguïté t dans le contexte, ou alors, il échoue. Il commence par une description qui est un ensemble vide, et ajoute de manière incrémentale des propriétés de façon à ce que la description élimine petit à petit toutes les entités du contexte auxquelles on ne veut pas référer mais qui satisfont encore la description, et que nous appellerons désormais des distracteurs. La description sera étendue jusqu'à ce que tous les distracteurs soient éliminés.

Une restriction est ensuite imposée à l'algorithme : on exclut toutes les propriétés déjà intégrées dans la description, qui, selon une hiérarchie de concepts, sont plus générales que les propriétés nouvellement ajoutées à la description.

La sortie de l'algorithme est alors une liste de propriétés qui identifient de façon unique le référent dans le contexte. En d'autres mots, l'algorithme donne en résultat le contenu sémantique de l'expression référentielle à générer. Il est alors possible que ce contenu sémantique ne soit pas verbalisable de façon acceptable. Afin d'éviter ce problème, on vérifie immédiatement que la propriété choisie peut être intégrée à un arbre syntaxique complet. En cas d'arbre syntaxique incomplet, on inclura alors une propriété qui n'est pas distinguante, mais qui permettra de générer un groupe nominal complet. Prenons l'exemple suivant : dans un ensemble contenant un carré rouge et un carré bleu, on veut désigner le carré rouge. La propriété distinguante entre les objets est alors la couleur. Cependant, on ne peut pas générer *le rouge* en première mention, il manque le nom-tête du syntagme. On va donc devoir générer la réalisation d'une propriété non distinctive de l'objet, qui soit réalisable sous forme nominale. On ajoutera alors à la description l'élément *carré*, afin de générer le syntagme nominal complet *le carré rouge*.

Nous reprenons maintenant l'exemple donné par [Gardent et Striegnitz, 2003] pour illustrer le fonctionnement de l'algorithme de [Reiter et Dale, 1992]. Supposons que l'algorithme ait pour entrée, la liste de faits R suivante :

R : {*lapin*(r_1), *lapin*(r_2), *lapin*(r_3), *chapeau*(h_1), *chapeau*(h_2), *chapeau*(h_3), *blanc*(r_1), *noir*(r_2), *blanc*(r_3), *dans*(r_1, h_1), *dans*(r_2, h_2)}

Le référent cible est r_1.

Le tableau 2.1 illustre les étapes suivies par l'algorithme.

La première colonne du tableau contient la liste des buts à atteindre, c'est à dire la liste des entités à décrire. Au début, seule l'entité r_1 est dans la liste des buts, la description est vide, et tous les objets de R sont des distracteurs. Dans la deuxième ligne du tableau,

Buts	Description L	Distracteurs	Action
[r_1]	∅	{r_1, r_2, r_3, h_1, h_2, h_3}	étendre L
[r_1]	{lapin(r_1)}	{r_1, r_2, r_3}	étendre L
[r_1]	{lapin(r_1), blanc(r_1)}	{r_1, r_3}	étendre L
[h_1,r_1]	{lapin(r_1), blanc(r_1), dans(r_1, h_1)}	h_1	empiler buts
[r_1]	{lapin(r_1), blanc(r_1), dans(r_1, h_1)}	r_1	empiler buts
[ ]	{lapin(r_1), blanc(r_1), dans(r_1, h_1)}		retourner L

FIG. 2.1 – Etapes successives de la recherche de description distinguante

la description n'est plus vide, et l'ensemble des distracteurs ne contient plus que trois objets. Dans la troisième ligne, la propriété ajoutée à L permet d'éliminer r_2. L'ajout de la propriété dans*(r_1, h_1)* ajoute alors un nouveau but, qu'il faut réaliser. Ce nouveau but vient au dessus de la pile, on doit donc le réaliser en premier. Comme l'ensemble des distracteurs ne contient pas d'autre entité contenant un lapin blanc, h_1 est alors éliminé de la pile. De la même façon, r_1 est alors identifié uniquement par la description *(lapin(r_1)*, *blanc(r_1)*, *dans(r_1, h_1)*, ce qui permet de réaliser la description *le lapin blanc dans le*, qui n'est pas satisfaisante syntaxiquement, mais qui sera « réparée » de la façon décrite précédemment, en décrivant h_1 par sa propriété nominalisable, même si elle ne le distingue pas de h_2 et h_3. On aura alors la description *le lapin blanc dans le chapeau*.

2.2.3 Les extensions de ces algorithmes

Dans cette section, nous présentons deux extensions de l'algorithme de Dale et Reiter qui répondent à une déficience immédiatement visible : il peut arriver qu'il n'y ait pas de description distinguante obtenue de cette façon. Pour y remédier, deux solutions ont été proposées, que l'on pourrait résumer ainsi : (1) donner la possibilité à l'algorithme de faire des descriptions négatives et de référer à des ensembles disjonctifs, (2) permettre à l'algorithme de désigner un objet grâce à une expression multimodale (incluant une expression linguistique et un geste). De très nombreuses autres extensions ont été proposées pour cet algorithme et particulièrement pour les descriptions définies vagues [Van Deemter, 2000] ou relationnelles [Krahmer et al., 2001], mais nous ne reviendrons pas dessus.

2.2.3.1 Extension aux descriptions booléennes

Afin de remédier au problème de ne pas trouver de description uniquement identifiante pour un référent, [Van Deemter, 2001] montre qu'il est possible d'étendre l'algorithme à des descriptions « négatives », en utilisant des expressions booléennes et des opérations de disjonction sur des ensembles. Pour plus de clarté, nous montrerons simplement comment l'algorithme est étendu, en reproduisant l'exemple donné par Van Deemter dans son article. L'entrée de l'algorithme est constituée de la base de connaissances suivante :

R = {*chien(a), chien(b), chien(c), chien(d), chien(e) caniche(a), caniche(b) noir(a), noir(b), noir(c), blanc(d), blanc(e)*}

Aucun des individus n'est uniquement identifiable selon l'algorithme de Dale et Reiter. Le terme *chien* est vrai pour tous, il ne permettra donc en aucun cas de les différencier. En revanche, les autres propriétés ne sont pas vraies pour tous les individus de la base de connaissances, et on doit pouvoir les différencier autrement. Van Deemter montre en effet qu'il est possible de désigner le chien c par une expression du type *le chien noir qui n'est pas un caniche*. Pour y parvenir, il faut donc arriver à la définition d'une propriété négative *ne pas être un caniche*.

Par ailleurs, si on veut référer en même temps aux chiens qui sont blancs et à ceux qui sont des caniches, ce qui est tout à fait possible linguistiquement avec l'expression *les chiens blancs et les caniches*, on doit pouvoir faire référence à des ensembles disjoints dans la même expression référentielle, ce que ne permet pas l'algorithme de base.

Nous n'entrerons pas dans les détails sur la méthode utilisée par Van Deemter, notons simplement que son algorithme est une extension de l'algorithme de base permettant de résoudre les problèmes mentionnés précédemment, avec des caractéristiques informatiques raisonnables.

2.2.3.2 Adaptation aux expressions multimodales

Une autre façon de remédier au problème de l'impossibilité d'obtenir une description uniquement identifiante dans le contexte est de générer une expression multimodale, c'est à dire de coupler l'expression linguistique avec un indice permettant de distinguer le référent visuellement. L'algorithme présenté par [Van der Sluis, 2001] permet de faire cela. Le principe est relativement simple : lorsque le « pointage » sur un référent est possible et quand l'expression linguistique est inefficace, par un calcul sur la focalisation et la saillance des référents potentiels de l'expression linguistique, on génère la possibilité d'attirer visuellement l'attention de l'utilisateur sur un référent. Cette solution est effectivement une bonne solution, mais elle n'est utilisable que pour des applications où les référents sont visibles à l'écran. Pour des applications d'un autre type (génération de textes résumant des données numériques, par exemple), cette solution n'est pas valable.

2.3 Extension de l'algorithme aux anaphores associatives

Les critiques émises par Van Deemter, Krahmer et Van der Sluis sont totalement justifiées. La critique que nous pouvons cependant faire rapidement est que les extensions qu'ils proposent, si elles permettent de faciliter la référence à des objets non identifiables uniquement, n'étendent pas le nombre de phénomènes linguistiques couverts par l'algorithme. Les expressions référentielles générées sont essentiellement des premières mentions sous forme de descriptions définies ou indéfinies, ou des mentions subséquentes sous forme de reprises coréférentielles fidèles ou de pronoms. Par ailleurs, les référents de ces expressions doivent obligatoirement être présents explicitement dans le contexte.

Le constat fait par Gardent et Striegnitz est le suivant : les descriptions définies ont des emplois bien plus variés que ceux emplois générés par les algorithmes de génération [Gardent et Striegnitz, 2000, Gardent et Striegnitz, 2003]. Très rapidement, il devient difficile de générer des descriptions définies sans une base de connaissances plus large que

la simple description des entités du domaine, et sans raisonnement par inférences sur cette base de connaissances. Elles montrent que pour générer toutes les utilisations des descriptions définies recensées dans la littérature, et plus particulièrement les anaphores associatives (exemple 42) et les reprises coréférentielles indirectes (exemple 43), on doit s'appuyer sur des connaissances du monde à partir desquelles on peut raisonner (faire des inférences). Gardent et Striegnitz proposent une extension pour remédier à cela, et montrent comment étendre l'algorithme de base à toutes les descriptions définies, anaphores associatives incluses.

(42) ***Ce restaurant*** *est excellent.* ***Le cuisinier*** *a été formé en France.*

(43) ***Une actrice*** *entra en scène.* ***La femme*** *portait un grand chapeau.*

Dans cette section, nous présenterons en détail les travaux de Gardent et Striegnitz ([Gardent et Striegnitz, 2000] et [Gardent et Striegnitz, 2003]). Nous commencerons par présenter les contraintes linguistiques et contextuelles permettant de générer des descriptions définies. Ensuite, nous présenterons la structuration des bases de connaissances en entrée du générateur nécessaires à la génération d'anaphores associatives. Enfin, nous décrirons l'algorithme qui permet de générer tous les types de descriptions définies.

2.3.1 Familiarité et Unicité

On reconnaît en général deux propriétés aux descriptions définies : la familiarité [Heim, 1982] et l'unicité [Russell, 1905].

Unicité Cette propriété est liée au fait que pour être identifié, le référent d'une description définie doit être le seul à correspondre à la description dénotée par le syntagme nominal dans le contexte. Ainsi, si on utilise la description définie *le lapin blanc dans le chapeau*, il doit y avoir dans le contexte, un seul objet x correspondant à la description `lapin(x)` $\wedge$ `blanc(x)` $\wedge$ `dans-chapeau (x)`.

Familiarité Pour qu'une description définie soit employée, le référent de l'expression doit être connu de l'interlocuteur. Le référent peut être connu parce qu'il a déjà été mentionné dans le discours, ou parce qu'il est présent ou inférable dans la situation d'énonciation.

On reconnaît aux référents quatre sortes de descriptions familières qui emploient le défini :

- les anaphores associatives
- les descriptions coréférentielles directes ou indirectes
- les utilisations familières indirectes (appelées aussi non familières) : cette catégorie couvre les descriptions définies avec des modifieurs sous forme propositionnelle (le fait que Jean vienne) ou faisant référence à un élément connu du (familier au) locuteur (le frère de ta voisine)
- les utilisations faisant référence à une situation plus large (larger situation uses) : la description réfère à un objet non mentionné dans le discours, mais connu de l'interlocuteur parce qu'il fait partie du monde ou de la situation de discours *(Le soleil, le premier ministre...)*

2.3.2 Descriptions définies et inférences

2.3.2.1 Anaphores associatives

Les anaphores associatives nécessitent un raisonnement à la fois sur le contexte et les connaissances encyclopédiques. En effet, l'interlocuteur doit pouvoir reconstruire le lien entre le référent de l'antécédent (entité mentionnée dans le discours) et celui de l'anaphore (entité nouvelle dans le discours). [Gardent et Striegnitz, 2003] ne considèrent que les relations méronymiques ou *relations par association* de [Clark, 1977], c'est-à-dire les relations entre une partie-nécessaire, une partie-probable ou une partie-inférable et un tout (cf chapitre 6).

2.3.2.2 Reprises indirectes

Les reprises indirectes sont les cas où deux descriptions différentes réfèrent à la même entité. On réfère à une entité connue de l'interlocuteur, en utilisant une description qui est nouvelle pour lui. Si l'interlocuteur peut établir qu'il s'agit de deux expressions référant au même objet, c'est parce que ses connaissances du monde lui permettent de lier les deux expressions au même objet. Ainsi, dans l'exemple 43, la propriété dénotée par la seconde expression référentielle est impliquée par les connaissances du monde, dans la mesure où *femme* est un hypéronyme de *actrice*. De même, dans l'exemple 44, on peut faire le lien entre *Jean* et *l'imbécile* parce qu'*imbécile* peut être une propriété d'un être humain auquel on réfère grâce au nom propre *Jean*.

(44) ***Jean*** *a oublié de venir à la réunion.* ***L'imbécile*** *a encore perdu son agenda.*

Une fois encore, seules les connaissances du monde permettent d'établir le lien entre les deux groupes nominaux.

2.3.3 Familiarité et unicité dans la génération d'anaphores associatives

2.3.3.1 Structuration du contexte

Dans l'algorithme standard pour la génération d'expressions référentielles, le contexte est un ensemble non structuré de faits. Pour pouvoir générer des anaphores associatives (et des expressions référentielles de manière générale), il est nécessaire d'avoir un modèle structuré du contexte et de formaliser les notions de familiarité et d'unicité en fonction de ce modèle. C'est ce que nous exposons rapidement dans cette section.

Le contexte est divisé en trois parties :

- Le modèle du discours : il s'agit d'un ensemble de formules atomiques représentant le discours déjà produit. (abrégé DM pour *Discourse Model* dans les exemples et le pseudo-code qui suivent)
- Les connaissances du monde : il s'agit d'un ensemble de règles qu'on considère comme un savoir partagé entre le locuteur et l'interlocuteur. (abrégées WKL pour *World Knowledge* dans les exemples et le pseudo-code qui suivent)
- Le modèle du locuteur : il s'agit d'un ensemble de formules atomiques représentant le savoir additionnel du locuteur, i.e. ce qu'il peut vouloir inclure dans son

discours.(abrégé SM pour *Speaker Model* dans les exemples et le pseudo-code qui suivent)

2.3.3.2 Unicité et Familiarité dans l'algorithme

Dans les algorithmes de génération standard, le référent cible de la description définie (l'entité à décrire) est obligatoirement connu de l'interlocuteur. La familiarité est donc une caractéristique donnée de la description, ce qui restreint la génération à la génération de descriptions coréférentielles.

Les anaphores associatives réfèrent à des référents non familiers reliés par une relation inférable à un référent familier. [Gardent et Striegnitz, 2003] proposent alors d'élargir la notion de familiarité en tenant compte de ces relations inférables. Ainsi, pour générer une anaphore associative, [Gardent et Striegnitz, 2003] étendent l'algorithme en introduisant non seulement des cibles correspondant à des entités déjà mentionnées, mais aussi à des cibles reliées par une relation associative à d'autres entités déjà mentionnées. Pour étendre la notion de familiarité à ces entités, les auteurs introduisent les notions d'ancres souhaitées et d'ancres potentielles.

Ancres souhaitées Dans la perspective du locuteur, étant donné une cible t, les ancres souhaitées $IA(t)$ sont les entités auxquelles le locuteur veut relier la cible. Ces ancres souhaitées sont donc déjà connues de l'interlocuteur, qui les relie avec la cible soit par identité (coréférence) soit par association (anaphore associative).

Ancres potentielles Dans la perspective de l'auditeur, étant donné une cible t et une description L, les ancres potentielles $PA(t, L)$ sont les entités faisant partie des connaissances partagées auxquelles l'auditeur peut, étant donné la description L, relier la cible soit par identité, soit par association sur la base des connaissances partagées.

Condition de familiarité La description L de la cible t satisfait la condition de familiarité si et seulement si toutes les ancres potentielles de t sont aussi des ancres souhaitées de t. Plus formellement, la familiarité peut s'exprimer ainsi :

$$IA(t) \subseteq PA(t,L)$$

Condition d'unicité L'unicité requiert que l'ensemble des ancres potentielles soit inclus dans l'ensemble des ancres souhaitées et que la cible soit unique par rapport à l'ensemble des ancres potentielles. Il doit donc n'y avoir qu'un seul objet correspondant à la description du locuteur et qui puisse être relié à l'ancre. La condition d'unicité peut alors s'exprimer formellement ainsi si a est une ancre potentielle :

$$PA(t,L) \subseteq IA(t)$$
$$\forall\ a\ \exists\ PA(t,L) : t \text{ est unique par rapport à } a \text{ étant donné } L.$$

2.3.3.3 Trois exemples illustrant les conditions d'unicité et de familiarité

Etudions maintenant un exemple satisfaisant les conditions de familiarité et d'unicité.

```
DM = (restaurant (r))
WKL = (∀x (restaurant(x)) → ∃y(cook(y) ∧ part-of(x,y)))
SM = (cook(c), part-of(c,r))
Cible = c
Description = cook(c)
```

Dans ce contexte, la phrase suivante pourra alors être générée :

(45) *Jean a emmené Jacques* **au restaurant**. **Le cuisinier** *portait une toque blanche.*

En effet, l'ensemble des ancres souhaitées de c est constitué uniquement de r, de même que l'ensemble de ses ancres potentielles. Les deux conditions sont respectées.

Voici maintenant un exemple violant la contrainte de familiarité :

```
DM = (zoo (z))
WKL = ()
SM = (cook(c), part-of(c,z))
Cible = c
Description = cook(c)
```

La phrase alors générée peut alors être la suivante :

(46) *Jean a emmené Jacques au zoo. Le cuisinier portait une toque blanche.*

Ici, la contrainte de familiarité est violée dans la mesure où l'ensemble des ancres potentielles de c est vide. Rien en effet dans la base des connaissances du monde ne dit qu'un zoo possède un cuisinier.

Enfin, pour terminer, voici un exemple violant la contrainte d'unicité :

```
DM = (restaurant(r1), restaurant(r2), italien(r1), chinois(r2))
WKL = (∀x (restaurant(x)) → ∃y(cook(y) ∧ part-of(x,y)))
SM = (cook(c), part-of(c,r1))
Cible = c
Description = cook(c)
```

La phrase générée pourrait alors être la suivante :

(47) *Il y a un restaurant italien au coin de la rue, et un restaurant chinois dans la rue à droite. Le cuisinier est excellent.*

Enfin, dans cet exemple, la contrainte d'unicité est violée parce que l'ensemble des ancres potentielles de c est constitué de { r_1 *et* r_2 }.

2.3.4 Présentation de l'algorithme

L'algorithme de Gardent et Striegnitz est réellement une extension de l'algorithme de Dale et Reiter. Nous reproduisons figure 2.3.4 le pseudo-code donné dans
[Gardent et Striegnitz, 2003]

L'idée principale derrière cette extension est d'utiliser la relation entre les ancres potentielles et les ancres souhaitées pour contrôler l'algorithme, qui démarre avec une description vide et l'étend jusqu'à ce que la condition d'unicité soit satisfaite, et tant que la condition de familiarité est satisfaite. Plus techniquement, l'algorithme fonctionne jusqu'à ce que l'ensemble d'ancres potentielles soit égal à l'ensemble d'ancres souhaitées. L'entrée de l'algorithme est constituée de l'entité cible et de la représentation du contexte, structurée comme cela a été décrit plus haut. Il construit simultanément le contenu sémantique et la forme de surface de l'expression référentielle.

Entrée :
WKL (connaissances du monde) : ensemble de règles reliant les entités les unes aux autres
DM (modèle de discours) : ensemble de formules atomiques
SM (modèle du locuteur) : ensemble de formules atomiques
t : entité cible, t appartient aux termes de SM et de DM.

Initialisation :
1. buts ← pile avec l'élément t
2 N ← structure syntaxique initiale avec une place vide pour un nom

Check success :
3. Si buts est vide, alors retourner <uniquely identifying, N>
4. but-courant ← but en sommet de pile
5. Si IA(but-courant) $\nsubseteq$ PA(but-courant, L(N)), alors retourner <unfamiliar, N>
6. Si PA(but-courant, L(N)) = IA(but-courant) et $\forall$ a $\in$IA(t) : t est unique selon a étant donné L(N) alors empiler but en sommet de pile ; aller en 4.

Etendre la description :
7. Si but-courant $\in$ terms(DM) alors R ← DM sinon R ← DM + SM
8. Essayer de sélectionner une formule atomique p applicable tel que R+WKL $\models$ p
9. S'il n'existe pas de tel p alors retourner <non identifying, N>
10 pour chaque o $\in$ termes(p) - termes(L(N)) dépiler(o, buts)
11. N ← N' tel que L(N') = L(N) $\cup$ {p}
12. Aller en 4.

FIG. 2.2 – algorithme de Gardent et Striegnitz

N est l'arbre syntaxique partiel ; on suppose qu'on a accès aux places libres dans l'arbre

et que la fonction L donne l'ensemble des propriétés que N verbalise. La sortie de l'algorithme est un arbre syntaxique et une classification de la description définie en *uniquement identifiante, non-uniquement identifiante* ou *non familière.*

La structure globale de l'algorithme étendu est identique à celle de l'algorithme standard. Une pile de buts indique les entités qu'on doit décrire. Après les initialisations (1-2), on entre dans la boucle principale qui réussit quand la pile de buts est vide (3). L'algorithme examine le but qui est au sommet de la pile (4-6) et si nécessaire étend la description (7-12).

La stratégie principale de l'algorithme standard est d'ajouter des informations dans la description jusqu'à ce que tous les distracteurs soient éliminés. Si ce n'est pas possible, aucune description définie n'est construite. La stratégie de l'algorithme étendu est d'ajouter de l'information jusqu'à ce que la condition d'unicité soit satisfaite. La description est étendue jusqu'à ce que toutes les ancres potentielles qui ne sont pas des ancres souhaitées soient éliminées et s'il est consistant avec le modèle du locuteur que la cible soit unique en fonction de l'ancre (6).

L'algorithme échoue et retourne la description et l'information *description non identifiante uniquement* si la condition d'unicité n'est pas satisfaite ou si la description ne peut être spécifiée plus (8-9). L'algorithme échoue donc dans les mêmes conditions que l'algorithme standard. L'algorithme étendu peut aussi échouer si la description ne satisfait pas la condition de familiarité. Dans ce cas, l'ensemble d'ancres potentielles n'inclut pas l'ancre souhaitée (5).

Comme dans l'algorithme standard, la définition et le choix des propriétés applicables sont assez libres. Il y a cependant des prérequis minimaux pour que la propriété p soit applicable à l'objet : le premier indique que la propriété p concerne un objet qui a déjà été mentionné, et le second que p soit une propriété complètement nouvelle dans la description. En d'autres mots, on ne doit pas pouvoir la déduire de l'union des connaissances du monde, du modèle de discours, et de la description déjà produite. Plus formellement, ces pré-requis se représentent ainsi :

$$\text{termes(p)} \cap \text{termes(L)} \neq \emptyset$$
$$\text{WKL} + \text{DM} + \text{L(N)} \nvdash \text{p}$$

Ces prérequis sont les mêmes que les deux premiers des trois mentionnés dans l'algorithme standard. L'algorithme étendu requiert en plus les restrictions conditionnelles suivantes à la place du troisième prérequis de l'algorithme standard :

- S'il y a des trous syntaxiques, remplir l'un d'entre eux (seules les propriétés pouvant en remplir un sont applicables)
- N'utiliser une propriété qui n'assure pas de la condition de familiarité que si un trou syntaxique doit être rempli et qu'il n'y a pas d'autre solution
- Parmi les propriétés qui assurent la satisfaction de la condition de familiarité, préférer celles qui vont dans le sens de la condition d'unicité en éliminant des distracteurs. S'il n'existe pas de telles propriétés, alors choisir une relation d'anaphore associative qui lie la cible à une entité $o \in$ *termes (DM).*

Comme dans l'algorithme standard, on peut ensuite imaginer différentes stratégies pour choisir parmi les propriétés. On peut soit suivre [Dale et Reiter, 1995] en imaginant un ordre prédéfini entre les propriétés, soit suivre [Dale, 1992] et choisir la propriété qui élimine le plus d'ancres potentielles.

2.4 Conclusion

L'algorithme de Gardent et Striegnitz a donc les mêmes capacités de génération des descriptions définies coréférentielles que l'algorithme standard de Dale et Reiter. De plus, il autorise les inférences sur les connaissances du monde en permettant les reprises indirectes coréférentielles et les anaphores associatives. Il ne couvre certes pas certains cas d'utilisation des descriptions définies (descriptions non familières), mais les auteurs donnent des pistes pour résoudre le problème. Par ailleurs, l'algorithme ne couvre que des relations *partie - tout* dans l'anaphore associative, ce qui est loin de couvrir tous les cas possibles. Les auteurs affirment d'ailleurs que pour résoudre ces problèmes, une étude empirique serait nécessaire.

Notre thèse propose des extensions à l'algorithme de Gardent et Striegnitz. Notre choix se porte sur cet algorithme car il permet de tenir compte du rôle des connaissances du monde dans la génération des anaphores, et autorise les inférences sur ces connaissances. Nous montrerons en effet dans la suite de cette thèse que les phénomènes référentiels reposent en grande partie sur ces éléments.

Nous nous proposons donc de mener notre étude de la façon suivante : après une analyse de corpus (chapitre 5 partie 2), nous soumettrons des pistes pour étendre l'algorithme de Gardent et Striegnitz dans les directions suivantes :

Caractérisation des liens impliqués dans l'anaphore associative Ceci permettrait d'élargir les possibilités de générer ce type d'anaphore, en permettant un raisonnement plus complexe sur les bases de connaissances données en entrée du générateur. Nous présentons nos résultats dans cette voie au chapitre 6, partie 2. L'algorithme de [Gardent et Striegnitz, 2003] ne propose que la génération d'anaphores associatives impliquant la relation de méronymie (relation partie / tout). Nous présenterons dans la deuxième partie de cette thèse une étude de corpus dont le but est d'identifier précisément les relations impliquées dans l'anaphore associative ainsi que la source de cette relation. En effet, si l'on considère l'exemple 48, on ne peut pas dire qu'une convalescence est une partie d'opération. Pourtant, on a clairement affaire à une relation associative dans ce cas. Il nous semble nécessaire de pouvoir introduire ce type d'anaphore associative à l'algorithme de génération, afin de permettre des textes naturels. Il est alors nécessaire de pouvoir déterminer exactement le type de la relation entretenue par les noms *opération* et *convalescence*, ainsi que de la source de connaissance permettant de construire la relation.

(48) *Jean a subi une* ***opération***. ***La convalescence*** *sera longue.*

Notre étude de corpus permettra aussi d'évaluer la faisabilité de l'automatisation de la construction des relations identifiées, et de la construction des inférences nécessaires à partir des diverses bases de connaissances du générateur.

Localiser les sources d'inférence impliquées dans les reprises Nous montrons au chapitre 7 que les reprises définies ou démonstratives impliquent très souvent des processus inférentiels. Comme pour les anaphores, il est nécessaire de localiser la source de ces inférence, afin de permettre la génération. Nous montrons aussi dans le chapitre 7 que les reprises peuvent ajouter de l'information. Nous menons une étude de corpus pour vérifier et affiner ces observations, afin de les utiliser dans l'algorithme de génération.

Introduction de la génération de déterminants démonstratifs Nous souhaitons réaliser ceci pour deux raisons. Tout d'abord, la reprise indirecte est dans de nombreux cas meilleure avec un démonstratif ; de plus, le démonstratif peut être moins ambigu que le défini dans des cas où deux référents ne sont pas possibles à distinguer par leurs propriétés. Ainsi, dans les exemples suivants, le démonstratif semble meilleur que le défini (exemple 49, emprunté à Kleiber), et permet de produire une reprise moins ambiguë dans l'exemple 50, ou le syntagme démonstratif réfère clairement au dernier *chat* mentionné.

(49) *J'ai vu une voiture. Cette voiture roulait très vite.*

(50) *Un premier chat entra. Il regarda autour de lui. Un second chat entra. Ce chat avait l'air affamé.*

Nous renvoyons au paradoxe de la reprise immédiate et aux capacités reclassifiantes du démonstratif dans le chapitre précédent pour plus de détails sur les phénomènes en eux-mêmes. Ajoutons simplement qu'introduire la distinction entre les deux déterminants dans un algorithme de génération agrandirait ses capacités à référer à des objets en autorisant des formes que la simple génération de descriptions définies ne permet pas. Nos résultats dans ce domaine sont présentés au chapitre 8 dans la deuxième partie de cette thèse.

Chapitre 3

Analyse de corpus annotés

L'utilisation de corpus électroniques en linguistique française n'est pas un phénomène récent. L'INALF (Institut national de la langue française) s'est donné pour mission depuis les années soixante de constituer une immense base de données textuelles destinées à la rédaction d'un dictionnaire, le *Trésor de la langue française*. Cette base de données textuelles est la base FRANTEXT, consultable librement à des fins de recherche sur le site internet de l'ATILF[2]. Le dictionnaire auquel elle a donné naissance est aujourd'hui en ligne, à la disposition des chercheurs, sous le nom de TLFI (Trésor de la Langue Française Informatisé)[3].

Un phénomène plus récent est la mise à disposition de corpus[4] annotés, c'est à dire de bases de données textuelles enrichies d'analyses linguistiques de tout ordre (phonologiques, morphologiques, syntaxiques ou sémantiques et référentielles) [Habert et al., 1997, Traum et al., 2003]. Plus précisément, un corpus annoté est un corpus dans lequel, à certains segments de mots ou de texte, on a associé des informations permettant leur analyse ou le repérage de phénomènes linguistiques. Ce chapitre est composé des éléments suivants : tout d'abord, nous expliquons l'intérêt d'une étude de corpus, et comment il est nécessaire de la planifier (section 3.1). Nous présentons ensuite (section 3.2) les problèmes posés par l'annotation de corpus. Enfin, nous terminons sur l'utilité d'utiliser des formats de fichier et d'annotation standard, autant pour la commodité des traitements informatiques que pour permettre la réutilisation des annotations d'un point du vue plus théorique (section 3.3).

[2]Laboratoire *Analyse et Traitement Informatique de la Langue Française*, UMR 7118, CNRS, Université de Nancy 2

[3]le Trésor de la Langue Française est consultable librement à l'URL : http ://atilf.atilf.fr/tlf.htm

[4]Nous utilisons le terme corpus dans une acception plus large de celle de [Habert et al., 1997], pour qui un corpus est « un ensemble de données langagières qui sont sélectionnées et organisées selon des critères linguistiques explicites pour servir d'échantillon du langage ». Pour nous, il s'agira simplement d'un ensemble de textes rassemblés sous forme électronique. Un corpus peut en effet servir à l'étude d'un seul type de texte, ou d'un seul sujet, auquel cas il ne sera pas forcément un échantillon représentatif de la langue.

3.1 Planifier une étude de corpus

3.1.1 L'intérêt des corpus annotés

La linguistique américaine [Boas, 1940, Fries, 1952] a longtemps utilisé des masses de données linguistiques attestées pour appuyer ses études. Chomsky remet en cause ce principe dès 1957, et privilégie l'introspection, c'est-à-dire l'intuition du locuteur sur sa propre langue [Chomsky, 1956, Chomsky, 1957]. L'attitude de Chomsky, bien que dominante, n'est pourtant pas générale et de nombreux travaux restent basés sur des exemples attestés [Imbs, 1971, Michea, 1964, Quirk, 1960]. Ces travaux sont plus tard portés par le courant de la linguistique de corpus (*corpus linguistics*, [Aarts, 1990, Leech, 1991]). A l'heure actuelle, l'utilisation des corpus est considérée comme un outil de base en recherche linguistique, et plus particulièrement l'utilisation des corpus annotés [Veronis, 2000]. Pour [Habert et al., 1997], *l'utilisation de corpus annotés permet d'observer finement les phénomènes et de remettre en question des postulats linguistiques.*

Les explications à cela sont diverses : tout d'abord, les corpus permettent de rendre compte dans le détail de la variation langagière, du point de vue du registre de langue, du point de vue sociolinguistique, ou tout simplement pour la distinction entre écrit et oral, sans en faire des caricatures.

Les corpus permettent aussi d'étudier l'articulation de la compétence et de la performance. Les corpus contiennent des énoncés qui pourraient être jugés agrammaticaux et qui pourtant ont été produits sciemment, en l'état. Les corpus remettent donc en cause la distinction entre énoncé acceptable et énoncé inacceptable.

De plus, les corpus permettent d'évaluer l'importance relative de certains phénomènes linguistiques. Ce comptage permet de ne pas éliminer des phénomènes linguistiques existant d'une étude, mais peut justifier par exemple qu'un système de TAL n'en tienne pas compte.

Enfin, un dernier argument de poids est lié au fait que les corpus dispensent le linguiste d'avoir à formuler lui même de jugement d'acceptabilité. Ceci rend plus facile l'étude d'états antérieurs de la langue. Nous pensons aussi que cela permet d'éviter d'argumenter sur des cas limites au sujet desquels les locuteurs d'une même langue ont des avis différents ou des hésitations.

[Habert et al., 1997] expliquent le regain d'intérêt pour les corpus annotés dans la communauté du TAL par son besoin de ressources de grande taille et par le besoin de combiner des méthodes statistiques à des méthodes symboliques.

Dans le cadre de notre thèse, l'étude de corpus n'a pas de lien avec les méthodes statistiques, qui permettent essentiellement un apprentissage automatique par les machines. D'ailleurs, à ce sujet, [Reiter et Sripada, 2002] montrent qu'il n'est pas judicieux d'utiliser les corpus comme moyen d'apprentissage pour les générateurs, dans la mesure où il existe une grande variation entre les textes produits liée à la variété des locuteurs. De plus, faire faire de l'apprentissage sur corpus pourrait amener le générateur à apprendre des « fautes » ou tout au moins des tournures non standard. Pour nous, l'intérêt de l'utilisation des corpus en génération d'expressions référentielles tient essentiellement en deux points :

– Dans un premier temps, l'analyse de corpus permet d'évaluer la fréquence d'apparition d'un phénomène. Ceci évite par exemple de focaliser les études sur la génération de phénomène rares, et d'« oublier » la génération de phénomènes fréquents. Par ailleurs, la génération de textes reposant sur des tâches de choix entre plusieurs possibilités, la fréquence supérieure d'une structure par rapport à une autre peut être un argument pour en favoriser la génération.
– L'autre point sur lequel l'étude de corpus est fondamentale en génération est l'observation. Seule une étude de corpus permet de déterminer les contextes d'apparition des phénomènes linguistiques, ainsi que leurs structures, leurs fonctions, et leur contenu.

3.1.2 La difficulté de l'étude de corpus

Pour Maria Wolters, l'étude de corpus ne peut pas se faire avec une vision simplifiée de la communication dans laquelle un locuteur transmet un message à l'interlocuteur [Wolters, 2002]. On doit tenir compte de facteurs extérieurs à la communication qui influencent la forme du texte produit par le locuteur. Ces facteurs peuvent être des conventions liées au domaine et au genre du texte, des idiosyncrasies, c'est à dire des utilisations de la langue propres à l'individu ayant produit le texte, ou alors des ambiguïtés de plusieurs types.

Pour faire une étude de corpus, il faut alors distinguer trois types de contenu dans un corpus :

– le contenu primaire, qui correspond au contenu sémantique du message produit,
– le contenu secondaire, qui correspond à ce que l'interlocuteur infère à propos du locuteur sur la base du message qui a été produit,
– le contenu tertiaire, qui est le niveau où l'interlocuteur relie les contenus primaires et secondaires à ses propres connaissances.

Lorsqu'on analyse un corpus, on se retrouve en situation d'interlocuteur, et on doit alors faire particulièrement attention à ne pas biaiser l'étude lorsqu'on construit les contenus secondaires ou tertiaires, surtout dans le cas d'énoncés ambigus. La distinction de ces types de contenus est cruciale en matière d'annotation des phénomènes que nous étudions dans cette thèse. Nous verrons dans toute la deuxième partie de notre thèse que l'analyse des liens coréférentiels est basée sur les contenus secondaire et tertiaire du corpus. En effet, lorsqu'on annote un corpus au niveau référentiel, on doit être capable d'analyser les inférences provoquées par le contenu primaire du discours, ainsi que le rôle de nos connaissances dans la production de ces inférences, afin de déterminer et d'identifier les liens coréférentiels. Par ailleurs, c'est l'interprétation des contenus secondaire et tertiaire qui peut poser des problèmes d'accord entre les annotateurs (cf. section 3.2), puisqu'ils peuvent produire des inférences différentes pour comprendre le texte, ou tout simplement analyser différemment les inférences qu'ils produisent pour résoudre les relations coréférentielles.

3.1.3 Les buts de l'étude de corpus

On peut vouloir faire deux types de choses lorsqu'on étudie un corpus : compter des phénomènes ou tester des hypothèses [Wolters, 2002].

3.1.3.1 Compter

La tâche de comptage des occurrences de phénomènes est la tâche la plus basique de l'étude de corpus. Les comptages sont utilisés dans deux types de travaux : les études descriptives, qui ont pour but une description précise de la langue étudiée, et les études statistiques, qui ont pour but l'élaboration de modèles de langue quantitatifs ou d'applications de TAL basées sur les probabilités d'apparition des phénomènes. Dans ces cadres, les comptages servent à plusieurs choses :

- montrer qu'une construction existe, et dans quelle mesure elle est utilisée ;
- mesurer combien de fois une construction ou un mot apparaît dans un texte ;
- montrer dans quels contextes un phénomène apparaît.

Nous ne nous attarderons pas dans ce chapitre sur les techniques statistiques utilisées dans les comptages sur des corpus. En revanche, nous souhaitons insister sur un point technique important souligné dans [Wolters, 2002] : la difficulté de l'analyse de corpus réside dans le fait qu'il est nécessaire pour réaliser des comptages d'avoir une définition très précise de ce que l'on cherche. La définition doit ensuite pouvoir être convertie en quelque chose de mesurable et quantifiable, elle doit être opérationnelle. Il faut donc donner des exemples concrets de ce qu'on inclut dans le compte ou de ce qu'on en exclut. Il est aussi bon de tester sur un échantillon de texte l'« opérationnalité » de la définition du phénomène linguistique que l'on recherche. Ainsi, la rédaction de manuels d'annotation donnant des définitions claires et des tests simples permettant le classement des phénomènes est nécessaire (cf. les annexes A et B de cette thèse, où l'on peut consulter les manuels rédigés pour l'annotation de notre corpus). Par ailleurs, un bon indice de l'opérationnalité des définitions des phénomènes est l'accord entre les annotateurs. Plus les définitions sont vagues, et plus la difficulté de classement des phénomènes est grande, ce qui conduit inévitablement à des désaccords entre les annotateurs (cf. section 3.2.2.2).

3.1.3.2 Tester des hypothèses

Souvent, les chercheurs utilisent des corpus pour vérifier des hypothèses, des convictions ou des intuitions. Pour confirmer des théories, de simples comptages ne sont pas toujours suffisants. Il faut parfois planifier une série de tests d'hypothèses. Le plus souvent, les tests se font sous forme d'expériences dédiées, dans des conditions précisément décrites. Il faut alors, dans ces circonstances, prendre garde à ce que les conditions expérimentales n'influent pas sur les résultats de l'expérience.

3.2 Annotation de corpus

Nous avons vu que la plupart des linguistes en TAL travaillent aujourd'hui sur des corpus qui sont non seulement dans des formats électroniques, mais qui sont en plus annotés,

c'est-à-dire qu'ils contiennent des informations linguistiques sur certains phénomènes linguistiques apparaissant dans les textes qui les composent. L'accès à ce type de ressources n'est pas facile : tout d'abord, pour des raisons juridiques de propriété intellectuelle, les annotations n'étant pas toujours publiques, et les textes annotés étant encore sous droits pour la plupart. Ensuite pour des raisons de quantité de corpus disponible : la constitution de ce type de ressources est longue et difficile, ce qui explique leur rareté. Pour finir, il est difficile de trouver un corpus dont les phénomènes annotés sont exactement ceux sur lesquels on travaille, et qui sont annotés comme on le souhaiterait. Il est donc la plupart du temps nécessaire d'annoter soi-même le corpus dont on dispose, afin de mener les études comme on l'entend.

Nous présentons dans cette section les techniques mises en œuvre pour élaborer une annotation de corpus : une série de prétraitements nécessaires à l'annotation ; la définition d'un schéma d'annotation, c'est à dire l'élaboration « théorique » de l'annotation ; et les contraintes techniques pour pouvoir la réaliser. Nous terminerons par la description des problèmes spécifiques à l'annotation référentielle des corpus.

3.2.1 Prétraitements

3.2.1.1 Constitution du matériau primaire

La première phase d'annotation d'un corpus est la phase de nettoyage et d'homogénéisation du (fichier contenant le) corpus [Habert et al., 1997]. Si cette phase semble scientifiquement peu enthousiasmante au premier abord, elle est fondamentale pour la réalisation d'une bonne annotation. En effet, les données primaires qui vont constituer le corpus proviennent de sources différentes, qui ne sont pas toujours fiables (reconnaissance optique ou vocale, transcription manuelle de discours oraux avec des traitements de textes divers, ...). Cette homogénéisation se fait en deux temps : un repérage de la structure du texte et des éléments marqués par une typographie particulière, et une phase de correction de l'orthographe. La TEI (Text Encoding Initiative, un projet de normalisation des ressources textuelles, cf. section 3.3.2) donne des recommandations précises au sujet de cette phase de traitement du corpus.

Identification de la structuration du texte Cette structuration doit être repérée et encodée à trois niveaux selon la TEI [Veronis, 1998]. Le premier niveau permet de baliser la structure globale du texte en chapitres, sous-chapitres, jusqu'au niveau du paragraphe, des titres, et des sous-titres. Le second niveau est le niveau d'encodage des sous-paragraphes, des énumérations, des éléments présentant une typographie spéciale, des passages de dialogues. Enfin, le troisième niveau est celui du repérage des abréviations, des nombres, des mots étrangers, des noms propres.

Vérification de l'orthographe Un autre type de nettoyage est la correction des fautes d'orthographe et de typographie : en effet, la provenance des textes étant, comme nous l'avons dit, variée, et on ne peut pas être assuré de la correction orthographique et typographique des textes. Si l'annotation est (semi-)automatique, il est nécessaire que

les ordinateurs puissent reconnaître les mots ou les ponctuations, afin qu'ils les analysent correctement. Cette étape n'est pas toujours simple car il peut arriver qu'on doive aussi prendre soin de distinguer les erreurs volontaires des scripteurs (transcription de la mauvaise prononciation d'un mot par exemple) des autres erreurs. Le projet AVIATOR [Blackwell, 1993] avait pour objectif le développement de filtres permettant de nettoyer du texte brut, et a permis de montrer que cette distinction entre erreurs volontaires et involontaires ne va pas de soi. Cette vision des choses n'est pas partagée par les personnes ayant décidé de se conformer à la TEI. La TEI ne préconise pas la correction des fautes d'orthographe, et refuse la notion de texte brut, qui est trop vague. Pour être conforme à la TEI, les fautes doivent être signalées par des balises <sic> et <corr>, qui permettent de noter la forme standard du mot, et de renvoyer les outils d'analyse automatique à la forme contenue dans la balise, et non à la forme du mot présente dans le texte.

La préparation du corpus de façon à ce qu'il devienne possible de le traiter pour les normes TEI doit se dérouler de la manière décrite dans le schéma 3.2.1.1.

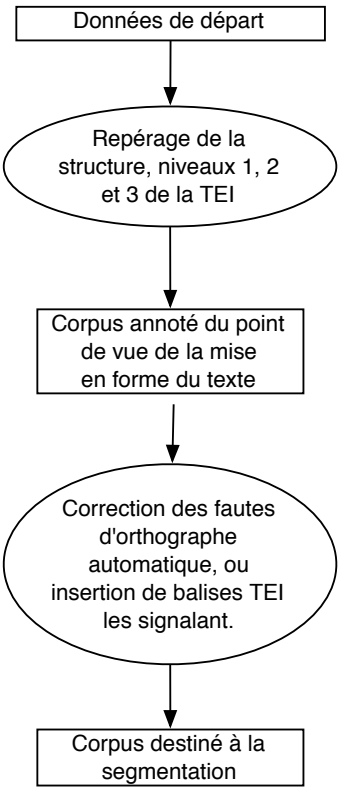

FIG. 3.1 – Préparation des données primaires

3.2.1.2 Segmentation

La segmentation est le processus qui permet de découper la suite de caractères qui composent le texte en mots et en phrases, et de repérer les signes de ponctuation dans le texte. Le problème de la segmentation est fondamental pour le repérage automatique

des unités que l'on souhaite étudier. En effet, si on veut qu'un programme puisse analyser correctement un texte, il est nécessaire de lui donner assez d'indications pour qu'il repère les mots simples, les mots composés, les débuts et les fins de phrases.

Le repérage des mots est une opération délicate pour l'analyse automatique de corpus. En effet, de nombreux caractères peuvent être à la fois des séparateurs de mots ou servir à relier les mots entre eux. Ainsi, dans une expression comme *va-et-vient* les tirets servent à indiquer qu'il s'agit d'un seul mot, ce qui n'est pas le cas pour *viendra-t-il ?*. De même, les blancs qui servent à indiquer le début et la fin d'un mot n'ont pas cette valeur dans le mot *pomme de terre* par exemple.

Il peut aussi être difficile de découper le texte en phrases : le point et la majuscule ne sont pas toujours des indications fiables dans la mesure où ils peuvent être respectivement la marque d'une abréviation ou d'un nom propre.

Pour repérer les mots et les phrases dans un texte, on utilise des segmenteurs, qui utilisent à la fois des règles permettant de s'appuyer sur le contexte pour délimiter les unités et des dictionnaires contenant des listes de mots simples et composés de la langue du corpus. Le système INTEX [Silberztein, 1993] est un exemple de ce type de fonctionnement.

Lorsque le texte est segmenté, la première phase de traitement du corpus servant directement à l'analyse du phénomène linguistique étudié peut commencer : il s'agit de la phase d'annotation du corpus.

3.2.2 Annoter le corpus

Il existe des règles d'annotation des corpus qu'il est bon de connaître et de respecter. Nous traduisons en introduction à cette section les sept maximes de Leech pour une bonne annotation ([Leech, 1993] cité par [McEnery et Wilson]), qui nous semblent être simples et résumer parfaitement le problème de l'annotation.

1. Il doit être possible d'enlever l'annotation d'un corpus pour le re-transformer en corpus brut.
2. On doit pouvoir extraire les annotations pour elles-mêmes.
3. Le schéma d'annotation doit être basé sur des directives disponibles pour l'utilisateur du corpus.
4. Il doit être spécifié clairement comment et par qui l'annotation a été réalisée.
5. L'utilisateur de corpus doit être averti que l'annotation de corpus n'est pas infaillible, seulement un outil potentiellement utile.
6. Les schémas d'annotation doivent être basés dans la mesure du possible sur des principes consensuels et théoriquement neutres.
7. Aucun schéma d'annotation ne peut être considéré comme un standard, les standards émergeant d'un consensus.

La plupart de ces maximes parlent d'elles-mêmes : les maximes 1 et 2 sont techniques et se justifient par l'idée qu'un corpus doit pouvoir être utilisé de la façon la plus flexible possible. Cependant, la maxime n°1 est discutable dans le sens où on ne sait pas exactement ce qu'est un corpus brut ; faut-il pouvoir récupérer le matériau tel qu'il est sorti de la

transcription ou des phases de reconnaissance vocale ou optique, ou faut-il le récupérer au format TEI niveau 1, 2, ou 3 ? ; les maximes n°3 et 4 se fondent sur l'idée que l'utilisateur du corpus annoté doit être complètement renseigné sur le contenu de l'annotation, et sur les définitions et décisions prises par l'annotateur. Nous revenons de façon plus approfondie sur les maximes n° 5, 6 et 7 dans les sections qui suivent.

3.2.2.1 Développement et définition d'un schéma d'annotation

La première tâche dans l'annotation d'un corpus est la création d'un schéma d'annotation, c'est à dire la définition des étiquettes nécessaires à l'annotation du corpus, le schéma d'annotation étant précisément l'ensemble des étiquettes utilisées pendant l'annotation. Cette tâche est une tâche ardue, dans la mesure où l'annotation du corpus doit pouvoir être indépendante de toute théorie linguistique (cf. maxime n°6 de Leech). De manière générale, l'étude de corpus doit permettre de valider des théories linguistiques ou d'en élaborer. [Salmon-Alt, 2001] décrit parfaitement le problème de la façon suivante : *la validation d'une théorie sur des données linguistiques présuppose un corpus annoté, et l'annotation du corpus doit reposer sur des principes cohérents, donc eux-mêmes issus d'une réflexion théorique.* Le schéma d'annotation est le reflet des principes sur lesquels repose l'annotation de corpus. Il contient toutes les relations et les données que le chercheur veut voir apparaître dans l'annotation du corpus. Ce sont en fait toutes les informations nécessaires à l'étude du phénomène visé, ou tous les éléments de la théorie qu'on cherche à valider. Il est spécifié dans un guide ou manuel de l'annotateur (cf maxime n°3).

3.2.2.2 L'accord inter-annotateurs

En général, une annotation est réalisée par plusieurs personnes. En effet, il est nécessaire de vérifier que les différents locuteurs d'une langue ont les mêmes avis ou la même perception des phénomènes linguistiques étudiés. Par ailleurs, cela permet d'avoir une idée de la pertinence du phénomène linguistique à annoter. Faire réaliser l'annotation du corpus par une seule personne fait prendre le risque de biaiser l'étude en cas d'incompréhension d'une définition, ou en cas d'erreur d'annotation.

Des mesures statistiques d'accord entre les annotateurs existent, et en particulier l'indice κ (kappa), qui mesure l'accord entre les annotateurs [Carletta, 1996]. Des mesures spécifiques sont en effet indispensables pour ce type particulier de tâches. Il s'agit de mesurer un accord sur une classification, il faut donc un outil qui ne permette pas seulement de mesurer que chaque catégorie de phénomène apparaît en proportions identiques d'un annotateur à l'autre, mais aussi que les segments linguistiques apparaissant à l'intérieur de ces catégories sont aussi les mêmes. La formule permettant de calculer cet indice repose sur les probabilités que deux annotateurs s'accordent sur une annotation. Nous reportons le lecteur à [Wolters, 2002] et [Carletta, 1996] pour les informations mathématiques concernant cet indice.

L'indice κ a deux fonctions particulières : il permet, s'il est bon, de vérifier la qualité d'un schéma d'annotation lors d'une annotation partielle par plusieurs annotateurs dans un premier temps. Dans un deuxième temps, il permet de vérifier la qualité d'un annotateur,

et sa compréhension du manuel d'annotation.

Pour terminer, ajoutons que malgré l'annotation par plusieurs personnes et les mesures statistiques permettant d'en valider la qualité, les maximes de Leech n°5 et 7 restent vraies. Il est difficile malgré tout cela de considérer que l'annotation puisse être sans erreurs et correspondre à un standard, étant donné la difficulté de la tâche et les désaccords qui persistent sur certains phénomènes linguistiques.

3.3 De la nécessité d'utiliser des ressources standardisées

Le travail d'annotation de corpus à quelque niveau que ce soit est extrêmement coûteux en temps, en énergie et en ressources informatiques. Les corpus standardisés présentent alors l'intérêt de pouvoir être ré-utilisés par d'autres chercheurs. Ainsi, le travail n'est pas fait pour un chercheur utilisant une application informatique spécifique mais pour une communauté et un ensemble d'applications utilisant les mêmes formats de données, les applications d'annotation automatiques ou semi-automatiques étant de plus en plus nombreuses. La standardisation doit donc se faire au niveau des formats de fichiers, et des contenus des fichiers. Dans cette section, nous présentons les deux types de standards définis aujourd'hui : le standard pour les formats de fichiers, les langages de balisage SGML et XML, et le standard permettant de coder le contenu des fichiers : la TEI.

3.3.1 Les langages de balisage SGML et XML

Un langage de balisage est un *système formel avec lequel l'information et le codage sont ajoutés à la version électronique d'un document dans le but de représenter son sens et donc d'en contrôler son traitement* [Bonhomme, 2000]. Le langage de balisage SGML est le langage de balisage standard. Les données au format SGML sont des données représentées sous forme d'arbre. Chaque segment de texte est délimité par des balises ouvrante et fermante qui donnent une indication sur ce segment de texte et que l'on reconnaît parce qu'elles sont entre le caractère « < » et le caractère « > ». L'information contenue dans la balise peut être soit un attribut (elle peut indiquer que le segment est un paragraphe, une phrase...), soit un couple attribut-valeur (par exemple, dans une annotation morphosyntaxique, l'attribut « catégorie » aura pour valeur un nom de partie du discours (i.e. nom, verbe, etc...)). Une balise prendra la forme suivante :

```
<information> SEGMENT DE TEXTE </information>
```

Le langage SGML est un langage extrêmement riche, et par conséquent, très lourd. Une nouvelle norme a été mise en place par le W3C[5], XML (Extended Markup Language), qui est un peu plus générale et moins lourde. Aujourd'hui, l'échange de documents par Internet se fait essentiellement dans le format XML.

Un document au format XML ou SGML doit être accompagné de ce qu'on appelle une DTD (Définition de Type de Document ou Document Type Definition), qui décrit

[5]W3C signifie World Wide Web Consortium. Il s'agit de l'organisation qui règlemente les aspects techniques de l'utilisation d'Internet. www.w3c.org

l'organisation de l'information contenue dans le document. Elle regroupe les types d'éléments contenus dans le document, l'organisation des éléments du document, les attributs que peuvent avoir les éléments, et les valeurs que peuvent prendre ces attributs. On peut la considérer comme une sorte de lexique des balises contenues dans le document. A la différence des documents SGML, la DTD est optionnelle dans les documents XML, bien qu'ils en respectent toujours une.

Nous le verrons dans le chapitre 5 de cette thèse où nous décrivons les traitements effectués sur le corpus, mais l'utilisation de standards a facilité notre propre étude. Nous disposions d'un corpus assez ancien mais au format SGML accompagné d'une DTD, ce qui nous a permis de le convertir en XML facilement, et de pouvoir utiliser les outils informatiques permettant de traiter des fichiers au format XML.

3.3.2 La TEI (Text Encoding Initiative)

La TEI est un projet international mis en place à la fin des années quatre-vingts dans le but de créer un environnement dans lequel les documents pourraient être encodés de façon à ce que leurs propriétés soient transcrites et que leur transcription puisse être échangée et survivre aux évolutions technologiques [Mueller]. Elle est utilisée par un comité de normalisation, au sein de l'organisme international ISO qui produit des normes dans tous les domaines de la vie courante ou scientifique. Ce comité de normalisation est le sous-comité TC 37|SC4 et il s'occupe des aspects de normalisation des données linguistiques[6].

La TEI travaille dans un format SGML, puisque ce format est le plus indépendant des applications et le plus standard. Il existe aussi une version XML de la TEI.

La TEI est en fait un format de représentation générique des ressources textuelles. Elle a permis de fournir une base commune pour la normalisation des documents, mais reste flexible. Les utilisateurs ont en effet la possibilité de choisir leur schéma de codage parmi les différents attributs qu'elle propose. Un document normalisé selon la TEI comporte cependant au moins deux éléments : l'en-tête et le texte qui constitue en lui même la ressource linguistique.

L'en-tête TEI *peut être vue comme une page de titre qui serait attachée à la version imprimée de la ressource* [Bonhomme, 2000]. Elle contient :

- une description du fichier - `<fileDesc>`
- une description du codage - `<encodingDesc>`
- une description du profil textuel, c'est à dire des informations liées au type de la ressource textuelle, de son contenu... - `<profileDesc>`
- un historique des révisions du texte - `<revisionDesc>`

L'en-tête TEI permet donc de documenter les ressources comme l'auraient fait des bibliothécaires ou des documentalistes dans le fichier d'un centre de ressources. Son contenu a d'ailleurs été élaboré par des professionnels de la documentation (documentalistes, archivistes et bibliothécaires). Le corpus que nous avons annoté et étudié comportait aussi une en-tête TEI, ce qui a facilité son

[6] www.tc37sc4.org

3.4 L'annotation de la référence, la coréférence et l'anaphore

Nous avons vu dans les sections précédentes comment standardiser, normaliser un document du point de vue de son format. Il est aussi important lorsqu'on annote un corpus de pouvoir avoir des normes d'annotation pour ne pas annoter le même phénomène dans plusieurs corpus en utilisant des balises ou des contenus différents. Pour notre thèse, nous avons annoté les phénomènes référentiels, ce qui en soi ne constituait pas une première. Un certain nombre de corpus ont en effet déjà été annotés à ce niveau linguistique, et des projets de normalisation de l'annotation de la référence ont publié des résultats. Nous présentons dans la section qui suit les schémas d'annotation de la coréférence proposés dans le cadre de deux grands projets internationaux : MUC [Chinchor et Hirschmann, 1997] et MATE [Mengel et al., 2000, Davies et al., 1998, Davies et Poesio, 1998].

3.4.1 Le projet MUC

Le projet MUC avait pour but l'évaluation de systèmes de compréhension automatique de textes, pour lesquels l'une des tâches les plus complexes est la résolution des coréférences. Une des sous-tâches de ce projet était donc d'élaborer un schéma d'annotation pour la coréférence [Chinchor et Hirschmann, 1997]. Dans ce projet, le but était d'identifier les expressions référant au même objet, et d'exprimer le lien entre ces expressions (les relations impliquant un verbe ayant été exclues). Par ailleurs, la coréférence événementielle et l'anaphore associative ne faisaient pas partie des phénomènes annotés.

Le schéma devait donc exprimer clairement les liens coréférentiels, et le type de lien coréférentiel entretenu par les deux expressions.

Voici un exemple de l'annotation produite dans le cadre du projet MUC :

```
In <COREF ID="11" TYPE="IDENT" REF="12" MIN="quarter"> the third quarter
</COREF>, <COREF ID="13" TYPE="IDENT" REF="10" MIN="company"> the company, which
is 61%-owned by Murphy Oil Corp. of Arkansas, </COREF> had
<COREF ID="100" MIN="loss"> a net loss of
<COREF ID="17" TYPE="IDENT" REF="100"> 46.9 million dollars</COREF>,
or <COREF ID="16" TYPE="IDENT" REF="17" MIN="91 cents"> 91 cents a share
</COREF> </COREF>.
```

Les balises encadrant les expressions coréférentielles sont les balises <coref> et </coref>. Elles signalent le début et la fin des expressions coréférentielles. Les attributs contenus dans ces balises sont les suivants :

L'attribut TYPE ne peut avoir dans le cadre du projet qu'une seule valeur, appelée **IDENT** pour « identity ». Ceci signifie que la relation anaphorique entretenue par les deux syntagmes annotés est une relation d'identité, et que les expressions réfèrent exactement au même objet ; ceci est explicable puisque ce projet ne tient compte que de la coréférence stricte, mais laisse une porte ouverte à l'annotation d'autres relations anaphoriques non coréférentielles.

L'attribut ID est un numéro identifiant uniquement l'expression référentielle dans le texte.

L'attribut REF indique le numéro de la chaîne de référence dans laquelle l'expression référentielle est incluse. C'est cet attribut qui indique la coréférence entre deux expressions : si deux segments portent la même valeur dans leur attribut REF, c'est qu'ils réfèrent au même objet.

L'attribut MIN indique la chaîne minimale à inclure dans la balise COREF pour le système d'évaluation. Cette chaîne doit pouvoir permettre d'identifier le référent dans la réponse du système (il semble que ce soit en fait la tête du syntagme lors de la première mention du référent).

Les critiques que l'on peut faire à ce schéma sont les suivantes [Salmon-Alt, 2001] : tout d'abord, il est vrai qu'il est possible d'étendre le schéma à d'autres types d'anaphores que l'anaphore strictement coréférentielle, mais aucune proposition n'est formulée, ce qui peut poser problème dans l'évaluation du système ou dans le cadre d'une étude plus spécifiquement centrée sur les problèmes de référence. Ensuite, se pose le problème des antécédents qui ne sont pas introduits verbalement dans le cadre des études sur le dialogue oral. Malgré tout, ce schéma constitue une première base pour l'étude et le repérage des expressions coréférentielles.

3.4.2 Les recommandations MATE

L'objectif de MATE (Multilevel Annotation Tools Engineering) est de définir des standards d'annotation de la référence pour permettre la réutilisation des corpus annotés, ainsi que des applications d'extraction et d'acquisition de connaissances [Mengel et al., 2000, Davies et al., 1998, Davies et Poesio, 1998]. Le projet est centré sur l'annotation de dialogues oraux, et ne se préoccupe donc pas uniquement de la référence.

La partie du projet concernant la coréférénce utilise deux balises différentes pour exprimer les liens entre deux segments linguistiques. La première balise, `<coref :de>` sert à délimiter les expressions référentielles à l'intérieur du texte (*de* signifie *discourse entity*). Elle contient un attribut `id` qui correspond à un identifiant unique pour le référent. La seconde balise, qui elle se situe à l'extérieur du texte, est la balise `<coref :link>` qui sert à exprimer le lien coréférentiel. Elle contient un attribut `type` qui exprime le type de la relation anaphorique (identité, partie de...) et encadre une troisième balise, `<coref :anchor>` qui sert à identifier l'expression référentielle qui sert d'antécédent, grâce à un attribut `href` qui a pour valeur la valeur de l'attribut `id` de la balise `<coref :de>` contenue dans le texte. Nous reprenons l'exemple suivant à [Salmon-Alt, 2001] :

maintenant je vais faire `<coref :de id="de1">` la maison `</coref :de>` faut mettre `<coref :de id="de2">` un toit `</coref :de>` dessus.

```
<coref :link href="de2" type="part">
   <coref :anchor href="de1"/>
```

```
</coref :link>
```

Dans cet exemple, on peut lire que le lien anaphorique entre l'entité de discours « de2 » et son antécédent « de1 » est de type « part ».

Les recommandations d'annotation de MATE incluent aussi la prise en compte de phénomènes de référence au contexte, ce qui est fondamental dans l'annotation de dialogues oraux. La proposition prise en compte dans le projet est celle de [Bruneseaux et Romary, 1997] qui propose en fait d'ancrer les référents dans une balise `<coref :universe>` qui représente la liste des référents du contexte mentionnés dans le discours.

Pour terminer, notons que [Salmon-Alt, 2001] fait une proposition de schéma d'annotation de la référence plus générique et probablement plus complète, en synthétisant les recommandations de MATE et MUC ainsi que le résultat de ses recherches personnelles.

3.5 Conclusion

Le but de ce chapitre n'était pas directement de présenter les normes et les standards, mais plutôt de présenter les motivations et la méthodologie sous-jacentes à l'analyse de corpus (et particulièrement l'annotation des phénomènes référentiels) à deux niveaux : celui de la standardisation des ressources et celui des contenus d'une annotation.

Standardisation des ressources Utiliser des ressources normalisées devient une nécessité dans la recherche scientifique. Nous avons vu que cela permet les échanges entre chercheurs et le passage d'une application informatique à une autre. La réflexion tourne donc aujourd'hui autour de la normalisation de corpus grâce aux technologies XML et à la TEI. Il semble en particulier nécessaire de se reporter autant que possible aux recommandations de la TEI, de MUC et de MATE avant de commencer une annotation de corpus au niveau référentiel. Nous avons suivi cette démarche, mais nous montrerons au chapitre 5 que cela n'a pas toujours été possible. Ceci signifie simplement qu'un travail important reste à faire pour que les outils d'annotation puissent permettre de se conformer aux normes et aux standards d'annotation.

Contenus de l'annotation La consultation des manuels d'annotation de grands projets concernant l'annotation de la référence permet aussi de voir ce que doit ou peut contenir une annotation de la référence. Si les contenus des annotations sont comparables, alors seulement les études de corpus pourront être faites de façon comparative.

Chapitre 4

Synthèse

Après l'état de l'art que nous venons de réaliser sur les trois domaines concernant notre thèse, nous réalisons ici une synthèse des conclusions auxquelles il nous amène. Nous abordons tout d'abord plus précisément la problématique de l'utilisation des déterminants dans une perspective de génération et la nécessité d'utiliser les corpus pour compléter les analyses théoriques. Nous revenons ensuite sur notre choix de l'algorithme de Gardent et Striegnitz qui nous semble plus adapté pour le type de génération que nous souhaitons réaliser.

4.1 Nécessité de définir des contraintes sur l'utilisation des déterminants

Nous avons conclu le chapitre 1 de cette thèse en dégageant une série de contraintes sur l'utilisation du défini et du démonstratif. Ces contraintes doivent être vérifiées sur les référents des descriptions (première mention et reprise, unicité du référent, saillance du référent) ou s'appliquer à la forme que prennent les expressions référentielles (reprises majoritairement directes et présence de modifieurs dans les reprises). Cependant, ces analyses ne donnent pas de tableau complet de l'utilisation du défini et du démonstratif. Les études comparatives portent essentiellement sur les principes d'interprétation [Corblin, 1987, Kleiber, 1986, Kleiber, 1988] mais très peu sur les contextes d'apparition des reprises, surtout quand les deux formes sont possibles (cf. le paradoxe de la reprise immédiate). Les études de corpus sont en revanche peu comparatives, mais portent effectivement plus sur les utilisations des déterminants que sur les principes d'identification des référents. Les contraintes que nous avons dégagées pour l'utilisation des déterminants nous semblent nettement insuffisantes pour réaliser un algorithme de génération d'expressions référentielles. Aussi, pour y parvenir, nous proposons deux pistes : l'analyse de corpus et la recherche de contraintes formalisables pour réaliser une application informatique.

4.1.1 Dépasser l'analyse par introspection

L'analyse par introspection a permis de dégager des principes fondamentaux pour l'interprétation des déterminants. Cependant, le petit nombre d'exemples sur lesquels les

linguistes travaillent ne permet pas d'exhiber des contextes d'utilisation des déterminants. Le paradoxe de la reprise immédiate en est à notre avis l'illustration la plus marquante. En effet, Kleiber [Kleiber, 1986, Kleiber, 1988] montre que le prédicat joue un rôle dans le choix du déterminant, mais, pour nous, les exemples ne sont pas convaincants. Il semble alors que le seul moyen de vérifier la théorie et de dégager une réelle contrainte sur l'utilisation du défini et du démonstratif en reprise immédiate soit de regarder ce que font spontanément les locuteurs en corpus. Même si les résultats ne sont pas très nets, il est possible qu'une tendance majoritaire d'utilisations se dessine, et bien que cela ne soit pas forcément satisfaisant pour une théorie explicative de la référence, nous pourrons alors dégager une contrainte pour la génération, en nous appuyant sur une « utilisation majoritaire ».

4.1.2 Trouver des contraintes formalisables

Un autre problème posé par les analyses existantes est qu'elles dégagent parfois des contraintes très difficilement formalisables. Ainsi, une rupture discursive [Kleiber, 1988] ou la valeur d'un adjectif dans l'argumentation [Theissen, 2001] sont des éléments non seulement difficiles à repérer, mais en plus difficiles à formaliser pour être intégrés à une application informatique. Ceci ne signifie pas que ces éléments ne sont pas des éléments intervenant réellement dans l'utilisation des déterminants, mais simplement que leur systématicité est difficile à tester, et leur réalité difficile à formaliser. Il est donc nécessaire d'identifier d'autres contraintes qui soient plus formalisables.

4.2 Nécessité de l'inférence en génération d'expressions référentielles

Nous avons expliqué au chapitre 2 les raisons qui motivent notre choix de l'algorithme de Gardent et Striegnitz dans notre travail. D'autres motivations apparaissent après l'étude de l'utilisation des déterminants.

La philosophie de l'algorithme de Gardent et Striegnitz repose sur l'importance des connaissances du monde et des inférences que font les locuteurs et les interlocuteurs dans le traitement des expressions référentielles. La simple caractéristique du démonstratif qui lui permet de reclasser les référents dans d'autres catégories est un argument en faveur de l'utilisation de l'inférence et des connaissances du monde. En effet, il semble difficile d'envisager un autre moyen que l'inférence pour résoudre ou générer des anaphores avec la reclassification.

Par ailleurs, les notions d'ancres nous paraissent être une bonne formalisation de la théorie de Corblin sur le défini qui ne renvoie pas à une mention antérieure mais se résoud dans un contexte ou *domaine d'interprétation*. Elle est aussi assez générale pour s'appliquer à toutes les utilisations d'un déterminant et non pas à un phénomène anaphorique particulier. Elle permet de résoudre les premières mentions, les coréférences et les anaphores associatives utilisant le défini.

4.3 Ce que doit apporter une nouvelle étude de corpus

Au vu des points que nous venons de développer, une nouvelle étude de corpus ne sera utile que si elle comporte les caractéristiques suivantes :

Adéquation avec les standards d'annotation Tout d'abord, il est nécessaire selon nous de nous conformer aux divers standards d'annotation décrits au chapitre 3 pour les raisons énoncées dans ce même chapitre. Nous souhaitons donc que notre corpus soit au format XML et que les annotations soient le plus possible en accord avec les recommandations internationales. Nous verrons que la mise en pratique de ces idées n'est pas toujours possible, et dans notre cas, nous n'avons pu les suivre totalement dans la mesure où notre outil d'annotation, s'il produit des balises XML, ne produit pas des annotations conformes à MATE ou MUC. Nous le verrons plus en détail au chapitre 5, mais l'outil d'annotation que nous avons utilisé produit une annotation externe au fichier, et la forme des balises qu'il produit n'est pas exactement la même que celle des balises préconisées dans les recommandations MATE et MUC pour des raisons techniques.

Quantification des phénomènes La seconde caractéristique que nous souhaitons attribuer à notre étude de corpus est une quantification des phénomènes, afin de dégager des tendances d'utilisation des déterminants. Il n'est pour nous pas question de faire des statistiques très fines sur nos résultats dans la mesure où pour des raisons de temps, nous n'avons pu faire annoter le corpus par une autre personne. Par ailleurs, notre corpus ne contient que des articles de presse, il n'a pas été constitué dans le but d'une étude des phénomènes référentiels, mais dans le but d'entraîner un étiqueteur morphosyntaxique : il s'agit donc d'une simple collection de textes, écrits par plusieurs auteurs sur des sujets variés. Pour cette raison, il ne nous semble pas utile de nous livrer à une véritable étude statistique des phénomènes, mais de simplement les quantifier, de les identifier, de les classer, et d'en étudier les fréquences relatives à l'intérieur du corpus.

Contexte d'apparition des phénomènes Nous essaierons aussi dans notre étude de corpus de tenir compte du contexte d'apparition des phénomènes de référence. Ceci sera une tâche des plus complexes dans la mesure où cela est très difficile à annoter. Cette étude ne pourra donc être réalisée qu'*a posteriori*, une fois que les phénomènes auront été classés et identifiés comme tels.

Sources de l'inférence Nous avons vu que les phénomènes référentiels reposent sur des mécanismes d'inférences. Une des questions fondamentales devient alors : à partir de quoi construit-on une inférence ? Nous allons donc annoter cet élément, afin de pouvoir donner à notre algorithme les bases des connaissances nécessaires pour construire les inférences.

Annotation de caractéristiques formalisables Enfin, les caractéristiques des expressions référentielles que nous annoterons seront dans la mesure du possible des caractéristiques formalisables. Ainsi, lorsque nous parlons de saillance du référent, il nous sera difficile de produire une annotation du type « antécédent saillant », d'autant plus que cette

notion est relative. En revanche, on sait qu'un élément en position de sujet grammatical a de fortes chances d'être saillant, et ceci est possible à annoter facilement. Même si cette solution n'est pas complètement satisfaisante, elle permet d'appréhender un élément concret et difficilement discutable du contexte d'apparition des déterminants.

Dans la deuxième partie de notre thèse, nous présenterons des analyses de corpus successives. Nous présenterons dans le premier temps le travail que nous avons réalisé sur le corpus dont nous disposions pour notre étude. Ce premier travail a consisté en une série de prétraitements, et une première annotation, dont nous donnons et discutons les résultats. Les trois chapitres suivants présentent les résultats d'études approfondies des phénomènes extraits de la première analyse du corpus : la première étude porte sur l'anaphore associative et se situe dans la continuité des travaux de Gardent et Striegnitz sur le sujet. La seconde étude concerne les utilisations coréférentielles des descriptions définies et démonstratives, du point de vue de l'apport d'information. Enfin, nous présentons une étude concernant le choix du déterminant.

Deuxième partie

Génération de descriptions définies et démonstratives coréférentielles

Chapitre 5

Etude de corpus

5.1 Introduction

Le but de l'analyse de corpus est le repérage dans des énoncés attestés de divers phénomènes pour en faire une étude empirique extensive. Il est alors nécessaire d'annoter le corpus et d'extraire des énoncés pertinents et des résultats chiffrés concernant les phénomènes étudiés. L'annotation est un travail extrêmement coûteux en temps, ce qui nous a poussée à nous donner deux contraintes fortes :

- l'automatisation maximale du traitement du corpus,
- l'utilisation d'outils standard afin que cette annotation puisse être réutilisable, aussi bien par nous-même que par d'autres membres de la communauté scientifique.

Dans ce chapitre, nous présentons le travail d'annotation du corpus de la façon suivante : nous commençons par présenter brièvement les outils que nous avons utilisés. Nous présentons ensuite les différentes étapes du traitement du corpus, puis le schéma d'annotation et, pour finir, les résultats de l'annotation.

5.2 Corpus et outils

5.2.1 Le corpus PAROLE

Le corpus que nous avons annoté est une sous-partie du corpus PAROLE[7] et comprend 65 000 mots annotés au niveau morphosyntaxique. Il est composé d'une série d'articles du journal *Le Monde* datant de septembre 1987. Ces articles appartiennent à toutes les rubriques du journal (Politique nationale et internationale, Economie, Sport, Culture et Loisirs). Le corpus est balisé mot à mot suite à l'annotation pour le projet PAROLE. Ceci signifie que dans le fichier qui contient le texte, on trouve des indications morphosyntaxiques concernant chaque mot du corpus. Ces indications sont contenues dans ce qu'on appelle des *balises* ou *étiquettes* qui encadrent chaque mot de la manière suivante :

```
<Balise> MOT </Balise>
```

[7]Corpus fourni par l'ATILF dans le cadre du contrat de plan Etat-Région Lorrain sur l'ingénierie des langues intégrant la collaboration de l'ATILF - UMR 7118 CNRS-Nancy 2 - et du LORIA - UMR 7503, CNRS, INRIA, INPL, Nancy 1, Nancy 2.

Chaque balise comporte plusieurs informations, en nombre différent selon la catégorie du mot annoté. Ces étiquettes sont héritées du schéma d'annotation Multext/Multitag pour l'action GRACE [Lecomte, 1997, Beaumont et al., 1998].

Chaque étiquette contient un certain nombre de champs, qui correspondent à des indications sur la description morphosyntaxique du mot annoté. Ces indications sont abrégées par une lettre. Nous ne détaillons pas le schéma complet des étiquettes pour chaque catégorie syntaxique entrant dans la composition d'un syntagme nominal, mais nous montrons ici comment différencier un syntagme démonstratif d'un défini.

Pour un syntagme nominal comme *le lapin*, les balises du déterminant et du nom sont les suivantes :

```
<w msd Da-ms-d> le </w> <w msd Ncms> lapin </w>
```

Les déterminants sont annotés sur 7 positions, indiquant respectivement la catégorie, le type, la personne, le genre, le nombre, le possesseur, la quantification. Les champs peuvent être vides s'ils ne sont pas pertinents. Ici, on doit lire que le mot (w) a la description morphosyntaxique (msd) suivante : c'est un déterminant (D), de type article (a), qu'il n'indique pas de personne (-), qu'il est masculin (m), singulier (s), qu'il n'y a pas d'indication du possesseur (-), et qu'il est défini (d).

Les noms sont annotés sur quatre positions : ici, *lapin* est un mot (w) qui a la description morphosyntaxique (msd) suivante : c'est un nom (N), commun (c), masculin (m), singulier (s).

On différenciera l'annotation du syntagme *ce lapin* de la manière suivante :

```
<w msd Dd-ms- -> ce </w> <w msd Ncms> lapin </w>
```

La balise pour le mot lapin ne change pas, en revanche, on doit lire la description morphosyntaxique du déterminant *ce* de la façon suivante : déterminant (D), démonstratif (d), personne non pertinente, masculin (m), singulier (s), pas de possesseur (-), pas de quantification (-).

La précision de l'annotation morphosyntaxique nous a été particulièrement utile dans l'utilisation des outils informatiques présentés dans la section suivante.

5.2.2 Outils utilisés

Nous décrivons ici brièvement les outils que nous avons utilisés dans notre travail sur corpus, puis le détail des traitements effectués pour passer du format de départ du corpus à un format permettant son annotation au niveau référéntiel[8].

5.2.2.1 G-search tool

G-search tool [Corley et al., 2001a, Corley et al., 2001b] est un outil qui permet d'identifier des structures syntaxiques dans un corpus annoté au niveau morphosyntaxique (i.e. dans lequel on a associé une partie du discours à chaque mot). Il permet de retrouver

[8]Cette partie du travail a été réalisée en collaboration avec Eric Kow, qui est l'auteur de l'intégralité des scripts conçus pour le traitement du corpus.

ces structures grâce à une grammaire définie par l'utilisateur (cf. section suivante). Les éléments terminaux de la grammaire sont des expressions régulières sur les éléments du corpus (mots, lemmes, étiquettes morphosyntaxiques). Le corpus dont nous disposions ayant été annoté au niveau morphosyntaxique de façon très fine, nous avons pu identifier les groupes nominaux nous intéressant et les isoler de façon à les repérer facilement dans une fenêtre de notre outil d'annotation, MMAX.

5.2.2.2 MMAX

L'outil d'annotation que nous avons utilisé, MMAX, a été conçu spécifiquement pour annoter manuellement les corpus au niveau référentiel, et plus précisément les relations de coréférence, d'anaphore associative et les corpus multimodaux [Müller et Strube, 2001a, Müller et Strube, 2001b]. Dans une fenêtre de l'application, on peut lire le texte à annoter. Par un simple système de sélection à la souris et de clic, on insère des balises XML qui permettent d'identifier le type de relation anaphorique et l'antécédent du syntagme considéré. L'intérêt d'utiliser ce type d'outil est d'une part qu'il facilite la tâche de balisage en XML, et d'autre part qu'il permet de transformer le corpus dans un format électronique standard, XML, qui permet la réutilisation du corpus.

5.2.2.3 Transformations XSL et tables HTML

La sortie de l'outil d'annotation MMAX est un fichier XML (eXtended Markup Language, cf. chapitre 3). Pour extraire les résultats, nous avons utilisé des feuilles de style XSL, qui transforment un document XML en fichier HTML. L'intérêt de ce type de traitement est qu'il permet de réaliser un comptage des différents phénomènes, ainsi qu'une présentation des résultats triés sous forme de tableaux.

5.3 Déroulement des traitements informatiques

Dans cette section, nous présentons les traitements réalisés sur le corpus afin de le préparer pour l'annotation. Nous présentons le schéma général du traitement (schématisé en figure 5.3.1), puis nous reprenons les étapes une à une : tout d'abord la première étape de préparation du corpus (section 5.3.2), puis l'écriture de filtres permettant d'utiliser G-search sur le corpus (section 5.3.3), l'écriture de la grammaire et le passage de G-search à MMAX (sections 5.3.4 et 5.3.5). Enfin, nous décrivons la phase d'annotation (section 5.3.6) et la façon dont les résultats ont été extraits suite à l'annotation (section 5.3.7)

5.3.1 Schéma général

De façon générale, notre but était de partir du format de départ du corpus PAROLE (format Multext), et d'arriver à un format XML pour pouvoir l'annoter avec MMAX et extraire les résultats sous forme de tableaux HTML (figure 5.3.1). Nous ne présentons ici que le traitement qui nous a permis de repérer les groupes nominaux à annoter. Un autre traitement a été réalisé pour séparer les titres des articles du reste du texte, grâce à des balises déja contenues dans le fichier Multext, et conservées tout au long du traitement.

Nous ne le détaillerons pas ici, pour des raisons de clarté. Précisons seulement que ce traitement a été réalisé pour des raisons ergonomiques, afin de pouvoir distinguer clairement les titres du texte des articles au moment de l'annotation.

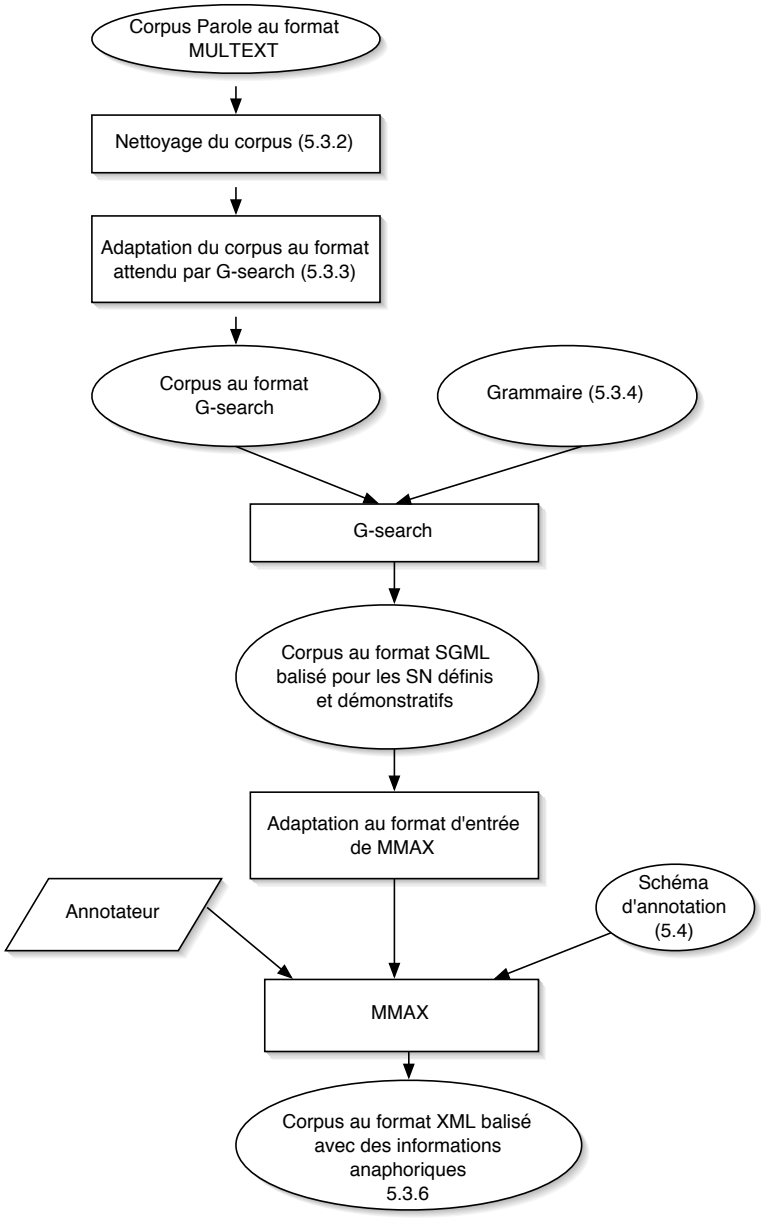

FIG. 5.1 – Traitement du corpus

5.3.2 Préparation du corpus

Le corpus contient quelques erreurs d'annotation dans les balises et des accents qui empêchent G-search de fonctionner. La première étape a donc consisté en un remplacement des caractères accentués par des caractères non accentués, une correction de fautes de

frappe dans les balises, et la suppression de certaines balises inutiles pour la suite du traitement.

Ensuite, le corpus a été divisé en sous-fichiers pour accélérer le traitement par les diverses applications. La difficulté de ce découpage était la suivante : pour identifier les relations anaphoriques, il ne fallait pas que les articles soient coupés. Nous avons donc utilisé les balises du format Multext indiquant le début et la fin des articles pour faire ce découpage automatiquement grâce à un script Perl. A ce stade du traitement, le fichier est toujours au format décrit en section 5.2.1, dont un exemple est reproduit figure 5.2.

```
<teicorpus.2> <tei.2>
<idno type="LeMonde"> 28937 </idno>  <head>
<w msd="Ncms===">CYCLISME</w>
<w msd="F======">:</w>
<w msd="Ncmp===">championnats</w>
<w lemma_1.2="de" msd_1.2="Sp=====" lemma_2.2="le" msd_2.2="Da-ms=-d">du</w>
<w msd="Ncms===">monde</w>
<w msd="Sp=====">sur</w>
<w msd="Ncfs===">route</w>
<w msd="Da-ms-i">un</w>
<w msd="Ncms===">triple</w>
<w msd="Sp=====">pour</w>
<w msd="Npfs===">Jeannie</w>
<w msd="Npms===">Longo</w>
</head>
<!- snip ->
</tei.2></teicorpus.2>
```

FIG. 5.2 – Format MULTEXT

5.3.3 Ecriture d'un filtre pour adapter le format du corpus au format de G-search

Il faut, pour pouvoir extraire les syntagmes nominaux étudiés, transformer le corpus au format que G-search demande en entrée (UIF - Uniform Input Format), ce que nous avons fait en réalisant un filtre. Ce filtre a aussi permis de traiter les formes contractées de préposition et de déterminants (du, des...), en faisant apparaître la trace du déterminant de la manière suivante dans le texte :

```
du -> du TRACE
```

Ceci a été réalisé en changeant les balises de la manière décrite en figure 5.3.

<w lemma-1.2=”de” msd-1.2=”Sp=====” lemma-2.2=”le” msd-2.2=”Da-ms=-d”>du</w>	⟶	du Sp===== de TRACE Da-ms=-d le

FIG. 5.3 – Transformation des formes contractées

Le format de sortie du corpus à ce stade du traitement est alors celui reproduit en figure 5.4.

```
CYCLISME                Ncms===
*+ :                    F======
championnats            Ncmp===
du                  Sp=====  de
TRACE                       Da-ms=-d  le
monde                   Ncms===
sur                     Sp=====
route                   Ncfs===
un                      Da-ms-i
triple                  Ncms===
pour                    Sp=====
Jeannie                 Npfs===
Longo                   Npms===
```

FIG. 5.4 – Format d'entrée de G-search

5.3.4 Ecriture de la grammaire

G-search réclame une grammaire en entrée afin de pouvoir repérer les structures recherchées. Nous avons donc écrit la grammaire reproduite en figure 5.5, qui décrit sommairement la grammaire du français, et qui permet de différencier les groupes nominaux en fonction de leur déterminant.

La première partie de la grammaire permet le découpage en syntagme. Les abréviations utilisées sont les abréviations anglaises. VP (*verbal phrase*) est mis pour syntagme verbal, NP (*noun phrase*) pour syntagme nominal, AP (*adjectival phrase*) pour syntagme adjectival, PP (*prepositionnal phrase*) pour syntagme prépositionnel. Les divers types de syntagmes nominaux (définis, démonstratifs ou autres) sont différenciés respectivement par les préfixes Def, Dem et Other. Les niveaux intermédiaires des constituants (correspondant aux niveaux N' ou N" dans les représentations syntaxiques arborescentes) sont suivis du chiffre 1. L'étoile de Kleene (*) signifie que l'élément peut être trouvé dans le constituant de 0 à une infinité de fois, et le symbole + signifie que l'élément doit figurer au moins une fois dans le constituant.

La deuxième partie de la grammaire donne les indications nécessaires à G-search pour retrouver les différents terminaux grâce à leurs balises, en utilisant des expressions régulières. Cette partie de la grammaire est celle qui permet de distinguer les syntagmes nominaux définis et démonstratifs des autres syntagmes.

Nous ne détaillons pas ici les diverses étapes menant à la sortie de G-search après l'application de la grammaire. Après le traitement par G-search, le corpus a le format reproduit en figure 5.6 : il s'agit d'un format SGML (format très proche de XML). Les balises notées entre [[[[ ]]]] signalent la présence d'un groupe nominal défini.

```
File:       GrammarFrenchDefNP
 Purpose:    Simple Gsearch grammar for French (NPs)
             without embedded NPs
 Time-stamp: <2002-13-04 13:41:32 kowey>

VP -> V1
VP -> V1 NP
VP -> V1 PP
V1 -> ADV* VERB ADV*

NP -> defNP
NP -> demNP
NP -> otherNP
defNP -> defDET N1+
demNP -> demDET N1+
otherNP -> otherDET N1+
NP -> N1+ PP*
NP -> NP CONJ NP

N1 -> AP* NOUN+ AP*
N1 -> N1 CONJ N1

AP -> ADV* ADJ
AP -> AP CONJ AP

PP -> PREP NP

 ------------------------------------ terminals

defterm "msd"          Saves writing

VERB -> <"V.*">
NOUN -> <"N.*">
PREP -> <"S.*">

ADV -> <"R.*">
ADJ -> <"A.*">

defDET -> <"Da.*d">
demDET -> <"Dd.*">
otherDET -> <"D[^ads].*[^d]">

CONJ -> <"Cc.*">
```

FIG. 5.5 – Grammaire pour G-search

```
<div><head>"file=subcorpus_0.xml sentence=s1"</head><p>
<phr type="s" id="s1"><w in_headline="1">CYCLISME</w>
<c in_headline="1">:</c> <w in_headline="1">championnats</w>
<w in_headline="1">du</w><phr type="FOO">

[[[[<phr type="defNP"><w in_headline="1">TRACE</w>
<phr type="N1"><w in_headline="1">monde</w>
</phr></phr>]]]]

</phr> <w in_headline="1">sur</w>
<w in_headline="1">route</w> <w in_headline="1">un</w>
<w in_headline="1">triple</w> <w in_headline="1">pour</w>
<w in_headline="1">Jeannie</w> <w in_headline="1">Longo</w></phr></p></div>
```

FIG. 5.6 – Balises SGML - sortie de G-search

5.3.5 Passage de G-search à MMAX

Nous avons ensuite préparé le corpus au format d'entrée de MMAX. Pour cela, nous avons dû contourner deux problèmes : celui de la répétition des analyses par G-search et celui de la numérotation des mots et des groupes nominaux, indispensable pour l'utilisation de MMAX.

5.3.5.1 Répétition des analyses

G-search répète la phrase à chaque fois qu'il trouve une structure du type recherché ; ainsi, si une phrase contient deux groupes nominaux définis, il la duplique, si elle en contient trois, il la répète trois fois, si elle en contient n, il la répète n fois. Un script a donc été conçu de manière à fusionner les phrases sorties de G-search, de façon à ce qu'une phrase contienne autant de balises que de groupes nominaux.

5.3.5.2 Numérotation des mots et des groupes nominaux

MMAX a besoin d'assigner des numéros à chaque mot et à chaque groupe nominal à annoter, ce que G-search ne fait pas. Un premier script a donc été réalisé pour résoudre le problème de la numérotation des éléments. Le corpus est alors dans un format intermédiaire (figure 5.7). On y voit la numérotation des mots ainsi que celle des différents constituants syntaxiques. Le syntagme nominal défini est identifié par `markable 2 DefNP 4 5`, ce qui signifie que ce constituant est le deuxième du fichier, que c'est un SN défini, composé des mots n°4 et 5.

Il nous a ensuite fallu adapter le format intermédiaire au format requis en entrée pour MMAX, c'est-à-dire deux fichiers au format XML.

Le premier fichier contient le texte balisé mot à mot (figure 5.8) et est appelé *wordfile*. Le second fichier (figure 5.9) contient des balises pour les syntagmes nominaux à annoter, de façon à ce qu'ils soient mis en évidence dans les fenêtres MMAX. Ces balises sont stockées dans le fichier appelé *markable*. On retrouve ici l'information déjà décrite à l'étape précédente : le constituant SN défini numéroté 0 ici est composé des mots n°4 et 5.

```
word 0 CYCLISME
word 1 :
word 2 championnats
word 3 du
word 4 TRACE
word 5 monde
word 6 sur
word 7 route
word 8 un
word 9 triple
word 10 pour
word 11 Jeannie
word 12 Longo
markable 0 s 0 12
markable 1 N1 5 5
markable 2 defNP 4 5
```

FIG. 5.7 – Format de sortie intermédiaire

```
<word id="word_0">CYCLISME</word>
<word id="word_1">:</word>
<word id="word_2">championnats</word>
<word id="word_3">du</word>
<word id="word_4">TRACE</word>
<word id="word_5">monde</word>
<word id="word_6">sur</word>
<word id="word_7">route</word>
<word id="word_8">un</word>
<word id="word_9">triple</word>
<word id="word_10">pour</word>
<word id="word_11">Jeannie</word>
<word id="word_12">Longo</word>
```

FIG. 5.8 – Balises mot à mot pour MMAX

```
<markable id="markable_0" span="word_4..word_5" tag="defNP"/>
```

FIG. 5.9 – Balises pour les SN définis

5.3.6 Annotation avec MMAX

L'annotateur clique sur la proposition qui lui semble pertinente, et l'information s'insère automatiquement dans le corpus, à l'intérieur des balises markable. Dans la copie d'écran (figure 5.10), nous pouvons voir une reproduction de la situation : dans la fenêtre en arrière plan, l'annotateur a cliqué sur le SN anaphorique (*le même choeur*), a sélectionné l'antécédent (*scander*), et a le choix pour annoter la relation entre les deux éléments. Les choix sont sélectionnés dans la fenêtre du premier plan : ici, nous voyons que l'annotateur a choisi la relation *bridging* (anaphore associative), que l'anaphore a un modifieur, n'est pas contenue dans un modifieur, et le type de l'anaphore associative est *thêta*[9]. L'annotation ainsi réalisée est intégrée au fichier *markable*, qui est un fichier XML.

L'annotation est donc une annotation externe au texte du corpus, réalisée par un seul annotateur.

Les propositions de balises qui apparaissent dans la fenêtre de premier plan proviennent du schéma d'annotation stocké dans le fichier *schemefile*. Le *schemefile* est un fichier qui permet de décrire le schéma d'annotation, et de créer les balises souhaitées dans l'annotation. Nous le présentons en Annexe C de notre thèse. Le schéma d'annotation regroupe l'ensemble des phénomènes que nous voulons annoter pour les étudier. Nous le décrivons dans la section suivante du point de vue théorique.

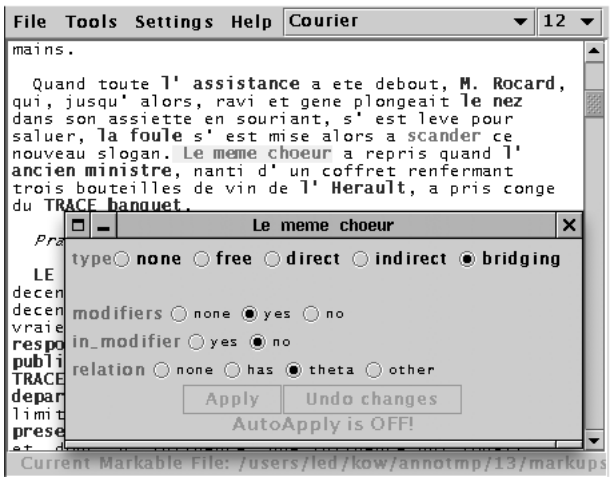

FIG. 5.10 – Fenêtres de MMAX pendant l'annotation

L'insertion dans le fichier *markable* des informations sur la coréférence prend le format reproduit en figure 5.11.

5.3.7 Feuilles de style XSL et sortie HTML

La rédaction de feuilles de style XSL et de scripts permettant de les appliquer aux fichiers XML sortis de MMAX nous a permis de réaliser des comptages de phénomènes catégorie par catégorie, et d'extraire des résultats triés ou non, en les présentant dans

[9] cf. section sur le schéma d'annotation

```
<markable id="markable_33" span="word_301..word_302"
type="free" member="" pointer="" in_modifier="no"  modifiers="no"  rela-
tion="cont_inf" />
<markable id="markable_32" span="word_288..word_289" type="coref" member=""
pointer="markable_55" in_modifier="no"  ante_status="tete"  modifiers="no"
source_repetition="wkl"  informational_status="IRA"
relation="ant_nprop"  ling_mean_add="none"  ante_funct="circ" />
```

FIG. 5.11 – Balises MMAX

des tableaux HTML (Figure 5.12). Ces tableaux ont ensuite été ouverts sous Excel (un tableur), ce qui nous a permis de les trier, de les classer ou de les filtrer.

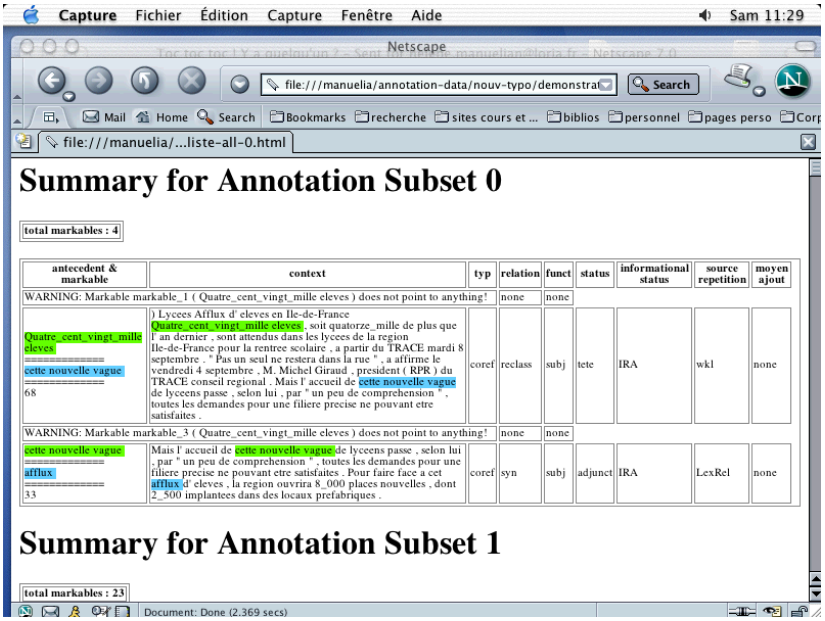

FIG. 5.12 – Tableaux HTML

La série de traitements décrits ici a été appliquée deux fois à partir de l'étape d'application de la grammaire de façon à créer deux corpus, le premier permettant d'annoter les groupe nominaux définis, le second d'annoter les démonstratifs. Pour faciliter le travail de l'annotateur et les extractions de résultats, il était nécessaire de séparer le traitement des deux phénomènes.

5.3.8 Conclusion

Nous nous étions fixé, au départ de l'annotation, deux contraintes que nous avons respectées : l'automatisation maximale du traitement du corpus, ainsi que l'utilisation de formats informatiques standard. Nous n'avons pu cependant respecter toutes les contraintes régissant habituellement l'annotation de corpus et ce, pour des raisons techniques :

Tout d'abord, pour des raisons de temps, nous n'avons pas pu réaliser de double anno-

tation du corpus. En dehors des anaphores associatives qui ont été vérifiées à deux suite à l'annotation, nous n'avons pas eu de deuxième annotateur. Ceci peut sembler un handicap à première vue, parce qu'il est important de pouvoir mesurer l'accord entre annotateurs pour vérifier la validité de l'annotation. Cependant, il est reconnu concernant l'annotation des relations anaphoriques ou coréférentielles que l'accord est très difficile à obtenir, surtout sur l'identification des antécédents [Poesio, 2002]. Nous considérons donc notre annotation comme indicative sur les relations existant entre les anaphores et les antécédents, sans toutefois la considérer comme totalement valide.

Aussi, en tenant compte de ces éléments, notre étude de corpus ne contiendra pas de statistiques précises sur le fait que les chiffres soient significatifs ; nous les considérerons comme des indicateurs sur les préférences des locuteurs.

Par ailleurs, nous aurions souhaité nous conformer complètement aux recommandations MATE pour l'annotation de la coréférence [Mengel et al., 2000, Bruneseaux et Romary, 1997], mais MMAX ne permet pas d'utiliser les balises selon ce schéma. Notre préférence s'est donc portée sur une annotation facilitée par l'application MMAX, se rapprochant au maximum des recommendations MATE, plutôt que sur une annotation complètement manuelle se conformant aux standards d'annotation.

5.4 Schéma d'annotation

Notre annotation se situe au niveau référentiel : notre étude portant sur des cas particuliers de reprises coréférentielles et d'anaphores associatives, nous avons annoté tous les syntagmes nominaux définis et démonstratifs (environ 10 000). Nous avons en effet choisi d'annoter les utilisations des déterminants en première mention, de façon à avoir un panorama complet de leurs utilisations. En revanche, nous n'avons pas annoté les indéfinis dans la mesure où ils ne sont, par définition, pas anaphoriques, sauf dans des cas très particuliers de coréférence événementielle [Danlos, 1999, Danlos et Gaiffe, 2000].

Le schéma d'annotation utilisé est le même pour les définis et les démonstratifs, les études théoriques et empiriques montrant que les deux déterminants peuvent être employés de la même façon, mais dans des contextes différents. Il nous a semblé important, pour que la comparaison soit possible, d'avoir des schémas d'annotation strictement identiques pour les deux déterminants. Le schéma d'annotation que nous avons utilisé a été élaboré en deux temps : dans un premier temps, nous avons suivi la classification des emplois des déterminants décrits dans le chapitre 1. Dans un deuxième temps, nous avons ajouté ou affiné des catégories, suite aux problèmes rencontrés durant l'annotation.

Les catégories utilisées dans le schéma d'annotation sont les suivantes :

5.4.1 Première passe d'annotation

Dans cette section, nous présentons les premières catégories que nous avons créées pour l'annotation.

5.4.1.1 Utilisation en première mention :

Le référent du groupe nominal est mentionné pour la première fois dans le texte.

(51) *(Le locuteur est assis dans un fauteuil) Ce fauteuil est confortable.*

(52) *Le soleil brille.*

5.4.1.2 Utilisation coréférentielle :

Le référent du groupe nominal a déjà été mentionné. La reprise peut être directe ou non, c'est-à-dire que le nom-tête du syntagme coréférentiel peut ou non être le même que le nom-tête de l'antécédent.

Reprise directe : La tête nominale de l'antécédent est identique à celle de l'anaphore.

(53) *Un chat entra dans la pièce. Ce chat avait l'air de s'être battu.*

(54) *Un chat et un chien entrèrent dans la pièce. Le chat avait l'air de s'être battu.*

Reprise indirecte : Il s'agit des cas où la tête nominale de l'antécédent est différente de celle de l'anaphore. Ce type de reprise peut prendre des formes variées qui sont les suivantes :

Reprise par hyperonymie : La reprise contient un nom plus générique que l'antécédent, et on peut dire : « Un *nom antécédent* est une sorte de *nom-anaphore* » :

(55) *Un chat entra dans la pièce. Cet animal semblait affamé.*

(Un *chat* est une sorte d'*animal*.)

(56) *Un chat entra dans la pièce. L'animal semblait affamé.*

(Un *chat* est une sorte d'*animal*.)

Reprise par hyponymie : La reprise comporte un nom plus spécifique que l'antécédent. On peut dire : « *Un nom-anaphore* est une sorte de *nom-antécédent*. »

(57) *Un chat entra dans la pièce. Ce siamois semblait affamé.*

(58) *Un chat entra dans la pièce. Le siamois semblait affamé.*

Reprise par synonymie : La tête de l'antécédent n'est ni plus ni moins spécifique que la tête de l'anaphore, mais elle est différente.

(59) *Le policier entra. Ce flic n'avait pas l'air aimable.*

(60) *Un policier entra. Le flic n'avait pas l'air aimable.*

Nous ajoutons à cette définition le fait que parfois, deux groupes nominaux ont le même sens si l'on tient compte des modifieurs. Nous en avons tenu compte dans les cas où l'un des noms est un nom prédicatif qui sous-catégorise des compléments[10].

(c-1) *« Nous sommes dans* ***l'ignorance la plus totale de ce qui se passe et de ce que veulent les Chargeurs****. (...) », constate un cadre.* ***Cette pénurie d'informations****, les ouvriers des usines quasi désertes de Tourcoing et de Cambrai n'en ont pas trop souffert.*

Reprise par le nom de rôle thématique (relation thêta) : Ce nom désigne les reprises par le nom de rôle thématique qui sont une forme de reprise utilisant la grille thématique d'un verbe [Clark, 1977, Manuélian, 2002]. Un syntagme anaphorique est classé dans cette catégorie lorsqu'il dénote le rôle du référent dans l'événement décrit précédemment.

(61) ***Jack*** *a vendu du vin à Bill.* ***Ce vendeur*** *est compétent.*

(62) ***Jack*** *a été assassiné hier.* ***La victime*** *se rendait au cinéma quand le malfaiteur a surgi.*

Reclassification : Cette catégorie comprend les reprises où l'antécédent et l'anaphore n'entretiennent pas de relation lexicale identifiée (qui peut aller jusqu'à un rapport métaphorique).

Dans ces utilisations, la tête nominale n'est pas impliquée directement dans l'identification du référent. Il n'y a pas de relation lexicale entre les têtes nominales des deux syntagmes. La reclassification permet au locuteur d'émettre un jugement ou d'apporter une information nouvelle sur le référent ou d'utiliser une figure de style :

(63) ***Un homme*** *entra dans le café.* ***Cet imbécile*** *commença à provoquer les habitués.*

(64) ***Jean*** *est venu.* ***Ce professeur de philosophie*** *est toujours ponctuel. (D'après Corblin, 1987)*

(65) *Tom regarde* ***les nuages****.* ***Cette écume*** *le fait rêver. (Cosse, 2001)*

(66) ***Un homme*** *entra dans le café.* ***L'imbécile*** *commença à provoquer les habitués.*

(67) ***Jean*** *est venu.* ***Le professeur de philosophie de ma sœur*** *est toujours ponctuel.*

Autres : La reprise n'entre pas dans les catégories décrites précédemment. Cette catégorie est nécessaire pour éviter aux annotateurs de devoir absolument classer des cas difficiles qui ne leur semblent pas correspondre aux définitions des autres catégories.

[10]Rappelons que les exemples extraits de notre corpus sont numérotés indépedemment des exemples construits, nous les préfixons avec un « c ».

5.4.1.3 Utilisation associative :

Cette catégorie (cf. chapitres 1 et 6) concerne essentiellement le défini. Cependant, on trouve aussi des mentions d'utilisation associative du démonstratif dans [Apothéloz et Reichler-Béguelin, 1999, Gundel et al., 2000] et [Nissim, 2001]. Il s'agit de cas où l'anaphore ne coréfère pas avec l'antécédent, mais où l'anaphore est associée à l'antécédent par les connaissances du monde. Le référent de l'anaphore n'est cependant identifiable que grâce à l'antécédent. Nous avons identifié trois grandes catégories d'anaphores associatives :

La relation has : Elle correspond à la *référence indirecte par association* de [Clark, 1977].[11] L'anaphore est associée à l'antécédent comme étant une de ses composantes :

(68) *Nous arrivâmes dans* ***un village***. ***L'église*** *était en ruines.*

(69) *Nous arrivâmes dans* ***un village***. ***Cette église****, tout de même, quelle horreur ! [Charolles, 1990])*

La relation « thêta » : Elle correspond à la *référence indirecte par personnification* de Clark. Il s'agit de cas où l'antécédent dénote un événement et le syntagme anaphorique dénote un participant à cet événement [Clark, 1977, Manuélian, 2002].

(70) *Le chef de l'Etat* ***a été assassiné*** *hier.* ***Le meurtrier*** *est toujours en fuite.*

La relation « autres » : Nous avons une catégorie supplémentaire pour tous les cas n'entrant pas dans les catégories décrites précédemment.

5.4.2 Catégories supplémentaires

Les classes décrites ci-dessus ont constitué une première base à notre annotation. Nous avons ensuite, dans une deuxième phase d'annotation, affiné les catégories « autres » (trop importante) et « reclassification » (trop hétérogène), et ajouté les catégories suivantes :

Relation métalinguistique : L'antécédent (nominal ou non) est repris par un nom désignant sa nature linguistique (exemples c-2 et c-3), ou par un élément mentionnel du type « ce dernier ». Ces reprises avaient été classées au départ dans la catégorie « autres ».

(c-2) ***Cinq Etats - l' Andhra-Pradesh, le Karnataka, le Maharashtra, le Madhya-Pradesh et le Rajasthan*** *- sont atteints pour la troisième année consécutive ; huit autres pour la seconde année ; enfin, huit nouveaux Etats sont venus s'ajouter en 1987 à* ***cette liste****.*

(c-3) ***Que faisait-il hier soir dans cet endroit sinistre ? La question*** *est dans toutes les têtes.*

[11] Nous utilisons des termes anglais parce que l'annotation est réalisée en parallèle avec une annotation du corpus NEGRA à Sarrebrück. Has est ici mis pour « avoir ».

Antécédent = nom propre : Nous avons classé à part les reprises de noms propres (au départ dispersées dans les catégories « hyperonymie », « autres » et « reclassification ») étant donné l'impossibilité d'établir une relation lexicale entre un nom propre et un nom commun. Cependant les reprises mentionnelles ayant pour antécédent des noms propres sont classées dans la catégorie « métalinguistique ». En effet, la forme de l'antécédent (nom propre ou non commun) n'a ici pas d'influence sur la relation entre la reprise et son antécédent.

(71) ***Jean*** *s'est cassé une jambe.* ***Cet homme*** *est malchanceux.*

(72) ***Jean*** *s'est cassé une jambe.* ***Le voisin de Marie*** *est malchanceux.*

Antécédent non nominal : Nous avons classé à part les cas où l'antécédent n'est pas un nom (classés dans « autres » au départ) pour les mêmes raisons qui nous ont poussée à établir une catégorie pour les noms propres. Nous avons cependant accepté la présence d'antécédents non nominaux dans les catégories « relation thêta » et « relation métalinguistique ».

(73) ***Jean s'est cassé une jambe***$_1$. ***Cette mésaventure***$_1$ *a provoqué des retards en série.*

Attributs et appositions : Nous avons défini une catégorie à part pour les cas où l'anaphore est liée à l'antécédent par une copule ou une virgule, considérant que la relation de coréférence est dans cette configuration exprimée explicitement, et utilise un autre mécanisme que la coréférence classique.

(74) ***Aggie*** *est* ***la voisine de ma sœur****.*

(75) ***Stephen*** *est* ***cet homme délicieux que tu as rencontré hier****.*

(76) ***Jarvis Cocker, le chanteur de Pulp,*** *se prend pour une pop-star.*

(77) ***Les frères Gallagher, ces primates dégénérés,*** *ont encore tout détruit dans leur hôtel hier soir.*

5.5 Résultats et discussion

Les résultats de l'annotation sont présentés dans le tableau 5.13. Nous étudierons une à une les grandes catégories d'emploi des syntagmes nominaux, et les mettrons en relation avec les travaux antérieurs réalisés sur le sujet [Poesio et Vieira, 1998, Vieira et al., 2002].

5.5.1 Première mention

Dans notre annotation, les emplois du défini sont majoritaires en première mention (78% des emplois). Ces résultats sont bien plus importants que ceux de [Poesio et Vieira, 1998] qui n'en trouvent qu'environ 50% dans leur première expérience et 43 à 46% dans la seconde. Ceci est peut-être lié à la différence d'utilisation du défini entre l'anglais et le français : en français, de nombreux noms sont déterminés avec le défini alors qu'ils ne prennent pas de déterminant en anglais :

définis	8863	100%
première mention	**6894**	**77,78%**
association	**392**	**4,42%**
association has	335	3,78%
association theta	12	0,14%
association other	45	0,51%
coref	**1577**	**17,79%**
directs	**602**	**6,79%**
indirects	**975**	**11%**
hypo	44	0,5%
hyper	127	1,43%
syn	131	1,48%
theta	26	0,29%
reclass	127	1,43%
other	62	0,7%
a = nom propre	228	2,57%
a non nominal	20	0,23%
metaling	66	0,74%
Att / app	72	0,81%

démonstratifs	557	100%
première mention	**98**	**17,59%**
association	**1**	**0,18%**
association has	1	0,18%
association theta	0	0%
association other	0	0%
coref	**458**	**82,22%**
directs	**92**	**16,51%**
indirects	**366**	**65,71%**
hypo	12	2,15%
hyper	62	11,13%
syn	29	5,21%
theta	18	3,23%
reclass	90	16,16%
other	6	1,01%
a = nom propre	42	7,54%
a non nominal	63	11,31%
metaling	23	4,13%
Att / app	10	1,79%

FIG. 5.13 – Résultats de l'annotation des SN

– Les noms de pays : La France vs. *France*
– Les emplois génériques : Les jeudis vs. *On thursdays*

Cette différence peut être aussi liée à la différence de définition des premières mentions. Dans notre manuel d'annotation, nous avons simplement défini la première mention comme étant l'emploi d'un syntagme défini référant à un objet non encore mentionné dans le contexte. Les catégories de [Poesio et Vieira, 1998] sont plus restrictives que cela (en apparence au moins), car ils ont défini deux types de première mention : les *unfamiliar* et les *larger situation uses* (cf chapitre 1). Il est possible que la précision des définitions ait poussé certains annotateurs à classer en anaphore associative certains emplois que nous avons classés en première mention.

Enfin, on peut expliquer ce résultat par un choix fait par l'annotateur : il arrive fréquemment dans ce corpus que les textes soient très longs, et que certains référents soient mentionnés plusieurs fois dans un texte, mais de façon très espacée. Dans ce cas, si le défini est utilisé dès la première mention du référent, (syntagmes du type *le Parti socialiste*), l'annotateur a ré-annoté le syntagme en première mention au lieu de chercher un antécédent très loin dans le texte.

En ce qui concerne les démonstratifs, [Salmon-Alt et Vieira, 2002] identifient environ 1% d'utilisations en première mention du démonstratif, alors que nous en trouvons près de 18%. Cette différence peut s'expliquer de deux manières :

Tout d'abord, les antécédents non nominaux des démonstratifs sont parfois très difficiles à identifier et à délimiter dans les textes. Les annotateurs ont pu annoter en première mention les cas difficiles à traiter comme des cas de coréférence.

Ensuite, le type de texte étudié par [Salmon-Alt et Vieira, 2002] est très différent du type de texte que nous avons annoté. Les textes étudiés par [Salmon-Alt et Vieira, 2002] sont des réponses des commissaires européens aux députés de l'Union européenne. Ils sont plus brefs que les textes du corpus PAROLE (qui est constitué d'articles du *Monde*) et leur style est moins littéraire, ce qui peut expliquer que l'emploi déictique du démonstratif soit moins souvent détourné que dans *Le Monde*, où nous trouvons en début d'article des exemples comme le suivant, sans référence à une photographie illustrant l'article :

(c-4) *Inde : les conséquences dramatiques de la sécheresse. La mousson, enfin... Son visage inondé de bonheur est tourné vers le ciel. Seul dans son champ noyé par les flots, il lève une main au-dessus de sa tête comme pour remercier les éléments et les dieux. La photo de* ***ce paysan des environs de Delhi*** *s'étalait récemment en première page de l'Hindustan Times.*

5.5.2 Utilisation Associative

L'anaphore associative représente environ 5% des emplois du défini dans notre corpus contre 6 à 11% chez [Poesio et Vieira, 1998]. [Salmon-Alt et Vieira, 2002] ne mentionnent pas ce type d'emploi pour le démonstratif, ce qui empêche toute comparaison. Nos résultats pour le défini ne sont pas très éloignés de ceux de [Poesio et Vieira, 1998], qui montrent que ce phénomène est difficile à identifier, puisque le résultat passe du simple au double en fonction des annotateurs. Nos résultats pour le démonstratif autorisent à penser qu'effectivement, ce phénomène existe, mais n'est pas significatif numériquement. Le défini reste donc le déterminant privilégié pour l'anaphore associative.

5.5.3 Utilisation coréférentielle

Généralités Les résultats du tableau 5.13 montrent que la coréférence représente environ un cas de défini sur cinq (environ 18%) tandis qu'elle représente la majorité des emplois du démonstratif est l'emploi coréférentiel (environ 82%). Les proportions sont donc presque exactement inversées entre emplois coréférentiels et emplois en première mention, ce qui confirme les données connues sur les déterminants, puisque la littérature a tendance à considérer le démonstratif comme le déterminant typique pour la coréférence, puisqu'il permet parfois de la forcer entre des groupes nominaux n'entretenant pas de relation lexicale (reclassifications).

Comparaison entre reprises directes et indirectes Par ailleurs, la fonction reclassifiante du démonstratif est parfaitement illustrée ici : la proportion de démonstratifs utilisés dans des reprises directes est bien plus faible (une pour quatre) que la proportion de définis utilisés dans des reprises directes (moins d'une pour deux).

Ces résultats ne confirment pas les résultats de [Salmon-Alt et Vieira, 2002], qui ont pour les définis comme pour les démonstratifs une majorité de coréférences directes. Nous pouvons toujours expliquer cette différence par la différence de type de texte, mais une étude des cas d'anaphores non classées par les annotateurs pourrait nous permettre de

mieux analyser cette différence (ces cas représentent environ 10% des définis et 8% des démonstratifs dans leur corpus).

Reprises indirectes Des résultats intéressants émergent de l'étude des reprises indirectes. Nous pouvons constater que la majorité des reprises indirectes avec le défini ont en fait pour antécédent un nom propre, ce qui n'est pas le cas des démonstratifs. La catégorie de reprises indirectes majoritaire pour les démonstratifs est la catégorie *reclassification*. Ceci tend à confirmer la fonction reclassifiante du démonstratif. Par ailleurs, nous constatons que pour la reprise d'un antécédent non nominal, le démonstratif est le moyen le plus utilisé (deuxième emploi en coréférence indirecte pour le démonstratif, alors qu'il est très minoritaire pour le défini). Enfin, pour la coréférence avec un nom commun, nous pouvons affirmer que la reprise par hyperonymie est la plus importante dans les deux cas, et que la synonymie est aussi très employée.

Nous n'avons pas de point de comparaison avec d'autres études sur les reprises indirectes, ces catégories n'étant pas détaillées dans les principaux travaux sur le sujet (i.e. [Salmon-Alt et Vieira, 2002, Vieira et al., 2002, Poesio et Vieira, 1998]).

5.5.4 Synthèse

Confirmation des données théoriques Nos résultats appuient les données théoriques connues sur l'utilisation du défini et du démonstratif (essentiellement [Corblin, 1987, Kleiber, 1986] et [Kleiber, 1988]). Nous montrons en effet que le démonstratif est employé essentiellement en coréférence, tandis que le défini introduit majoritairement des référents nouveaux. Nous montrons aussi que le démonstratif permet effectivement la reclassification des référents, tandis que le défini reprend en majorité des noms propres, ce qui est en accord avec les mécanismes d'identification des référents attribués aux syntagmes nominaux définis.

Confirmation des données empiriques Notre étude de corpus va aussi dans le même sens que les grandes études empiriques menées sur le sujet (principalement [Salmon-Alt et Vieira, 2002, Vieira et al., 2002, Poesio et Vieira, 1998]). L'accord porte essentiellement sur les quatre grandes catégories d'utilisation des syntagmes nominaux définis et démonstratifs (première mention, association, reprise directe et reprise indirecte). Les différences ne sont selon nous pas très significatives si l'on considère que l'une des études porte sur l'anglais, et l'autre sur un type de texte très différent de celui que nous étudions.

Intérêt pour la génération de textes Ces études constituent des données préliminaires pour la génération de textes. En effet, elles apportent la confirmation des données théoriques sur les moyens linguistiques utilisés pour les syntagmes coréférentiels. Ainsi, nous savons qu'une reprise indirecte sera sans doute meilleure avec un démonstratif si l'antécédent est un nom commun. Par ailleurs, on sait aussi que la reclassification est facilitée par l'emploi du démonstratif, et que la reprise par hyperonymie est la plus couramment employée.

Problèmes en suspens Si ces données sont déjà importantes pour la génération de textes, une telle étude est insuffisante pour la conception d'un algorithme de génération d'expressions référentielles. Les données manquantes sont les suivantes :

- Pour la génération d'anaphores associatives : nous n'avons spécifié que deux grandes relations possibles entre l'antécédent et l'anaphore (« has » et « thêta »). Il est clair depuis les travaux de [Clark, 1977] que ces anaphores nécessitent des inférences, et notre classification ne nous donne pas d'information sur la source de ces inférences, ni sur l'ensemble des relations sémantiques possibles unissant l'antécédent et l'anaphore. Nous devons donc répondre à ces questions pour pouvoir générer des anaphores associatives. C'est ce que nous ferons dans le chapitre 6 de cette partie.

- Pour la génération de descriptions coréférentielles définies ou démonstratives : les emplois reclassifiants et les reprises de noms propres laissent supposer que l'anaphore nécessite des processus inférentiels aussi bien pour sa résolution que pour sa génération, il nous faut donc en déterminer les sources. Au cours de notre étude, nous avons par ailleurs observé que dans de nombreux cas, les reprises définies et démonstratives n'ont pas le même contenu sémantique que leur antécédent parce qu'elles sont utilisées pour ajouter de l'information sur leur référent. Nous étudierons dans le chapitre 7 la façon dont l'information nouvelle et l'information connue sur le référent sont utilisées dans les reprises, afin d'affiner les algorithmes de génération existants pour la détermination du contenu des expressions référentielles, en identifiant les sources d'inférences permettant de réaliser des descriptions ajoutant de l'information ou répétant de l'information.

- Pour le choix du déterminant : nous n'avons pas de données sur ce qui privilégie le choix de l'un ou l'autre des déterminants. Nous savons certes que l'un sera privilégié pour certains types de reprise, mais l'autre n'étant jamais formellement exclu, nous devons déterminer des critères de choix plus précis. Par ailleurs, l'étude de corpus présentée dans ce chapitre ne donne d'indication que sur la tête des syntagmes nominaux (antécédents ou coréférentiels), alors que le problème des modifieurs reste posé. Pour finir, il nous manque des paramètres plus formels tels que la distance entre antécédent et anaphore ou la fonction syntaxique de l'antécédent. Nous donnerons au chapitre 8 des contraintes issues d'une nouvelle étude de corpus pour le choix du déterminant, afin de parvenir à un algorithme.

Chapitre 6

Les relations associatives

Les anaphores associatives sont étudiées depuis les travaux de [Clark, 1977]. Il s'agit de cas d'anaphores où l'antécédent et l'anaphore n'entretiennent pas de relation de coréférence, mais où la construction d'un lien entre les deux syntagmes est nécessaire à l'interprétation du deuxième syntagme, comme dans l'exemple suivant :

(78) ***Cette maison*** *n'est plus habitable.* ***Le toit*** *s'est effondré.*

Ici, le syntagme défini *le toit* est interprété de façon non ambiguë comme le toit de la maison mentionnée dans la première phrase. Ce lien est réalisé grâce à nos connaissances du monde qui nous disent qu'une maison a un toit, et que si on mentionne un toit dans un texte, il s'agit sans doute possible du toit d'une maison mentionnée précédemment dans le contexte.

Diverses typologies ont été élaborées pour définir une sémantique des liens entretenus dans ces cas particuliers d'anaphores. Nous présentons celles qui sont les plus importantes dans la première partie de ce chapitre. La seconde partie du chapitre s'attachera à montrer en quoi les classifications présentées sont insuffisantes pour le traitement automatique des langues, et quelles contraintes doit se donner une classification des anaphores associatives pour être utile à la génération. Nous présentons ensuite notre classification, et l'annotation de corpus consécutive à cette nouvelle classification. Nous exposons pour terminer les résultats de l'analyse de corpus et discutons leur pertinence pour la génération.

6.1 Etudes théoriques

La première typologie que nous présentons est celle de [Clark, 1977], qui est à l'origine des nombreux travaux sur l'anaphore associative qui ont été produits en linguistique. Les anaphores associatives sont très souvent réduites à la notion de méronymie (relation partie - tout). Nous présenterons donc ensuite une typologie de la méronymie proposée par [Winston et al., 1987]. Pour finir, nous présenterons les travaux de [Kleiber, 1997, Kleiber, 2001], qui a essayé d'affiner la typologie de Clark, en la basant sur des tests formels.

6.1.1 Typologie de Clark

Le phénomène que [Clark, 1977] appelle *bridging* est de façon générale le phénomène qui permet de relier un antécédent et une anaphore grâce à une série d'implicatures. Clark distingue plusieurs types d'implicatures qui sont les suivantes :

6.1.1.1 Appartenance à un ensemble

On peut faire référence à un objet après avoir évoqué l'ensemble auquel il appartient. La façon dont l'ensemble est mentionné peut varier.

(79) *J'ai rencontré* ***deux personnes*** *hier.* ***La femme*** *m'a raconté une drôle d'histoire.*

(80) *J'ai rencontré* ***un couple*** *hier.* ***L'homme*** *était réellement stupide.*

Dans l'exemple 79, on mentionne un groupe de deux personnes, puis une femme spécifique. On interprète le syntagme nominal *la femme* comme un membre de l'ensemble de deux personnes mentionné dans la première phrase grâce à la relation d'hyponymie entrenue par les noms *personne* et *femme*. Dans l'exemple 80, on interprète le syntagme *l'homme* comme référant à l'homme faisant partie du couple parce que nos connaissances du monde nous disent qu'on désigne généralement par le mot *couple* un ensemble composé d'un homme et d'une femme.

6.1.1.2 Référence indirecte par association

Cette catégorie de *bridging* regroupe les cas où le référent du syntagme anaphorique n'est pas encore mentionné directement, mais sa présence est prédictible car il est plus ou moins étroitement associé au référent du syntagme antécédent. Clark distingue trois niveaux de proximité des référents dans ce cas d'anaphore associative[12].

Parties nécessaires Ici, l'anaphore et l'antécédent sont associés très étroitement. La présence de l'antécédent implique nécessairement la présence de l'anaphore, qui est vue comme une partie de l'antécédent (qui peut référer à un objet, un événement ou une situation).

(81) *Je suis entré dans* ***la pièce.*** ***Le plafond*** *était très haut.*

(82) ***La course*** *a été très longue.* ***Le sprint final*** *fut plein de suspense.*

(83) *J'aime beaucoup* ***ce tableau.*** ***Les couleurs*** *sont saisissantes.*

Parties probables Ici, les liens entre l'antécédent et l'anaphore sont un peu plus lâches. La présence de l'antécédent n'implique pas nécessairement celle de l'anaphore, mais l'implique très probablement.

(84) *Je suis entré dans* ***la pièce.*** ***Les fenêtres*** *donnaient sur la mer.*

(85) *Je suis* ***allée en courses*** *hier.* ***La marche*** *m'a fait du bien.*

(86) *Je me suis levée à* ***huit heures*** *hier.* ***L'obscurité*** *m'a surprise.*

[12]Tous les exemples donnés ici ne sont pas exactement ceux donnés par Clark, mais dans la plupart des cas, nous avons essayé d'en faire des traductions fidèles.

Parties induites Dans ce dernier cas d'anaphore associative, l'anaphore est résolue par le besoin de retrouver la présence d'un antécédent.

(87) *Je suis entré dans* ***la pièce. Les chandeliers*** *étincelaient.*

(88) *Je suis* ***allée en courses*** *hier.* ***La montée*** *m'a épuisée.*

(89) *Je me suis levée à* ***huit heures*** *hier.* ***La précipitation*** *était de circonstance.*

6.1.1.3 Référence indirecte par caractérisation

La référence indirecte par caractérisation procède des mêmes mécanismes que la référence indirecte par association, mais concerne plus spécifiquement les antécédents événementiels. L'anaphore se fait sur l'un des participants de l'événement mentionné dans la première phrase. Ici encore, on distingue trois types de *brigding*.

Rôles nécessaires L'événement décrit dans la première phrase fait appel obligatoirement à des participants, même s'ils ne sont pas mentionnés :

(90) *Jean* ***a été assassiné*** *hier.* ***Le meurtrier*** *court toujours.*

Rôles optionnels Ici, le participant mentionné dans l'anaphore n'est pas obligatoire dans l'événement décrit précédemment, mais peut très probablement en faire partie.

(91) *Jean* ***a été assassiné*** *hier.* ***Le couteau*** *a été retrouvé à côté du corps.*

Raisons, causes, conséquences, concurrences Dans les cas décrits ici, l'anaphore porte soit sur des événements ayant provoqué une situation (cause ou raison), soit sur des événements étant la conséquence d'une situation ou une implication possible du premier événement.

(92) *John est tombé. Il voulait effrayer Marie.* [Raison]

(93) *John est tombé. Il a glissé sur un rocher.* [Cause]

(94) *John est tombé. Il s'est cassé un bras.* [Conséquence]

(95) *John a fait la fête hier. Il sera encore saoûl ce soir.* [Concurrence]

6.1.2 Typologie des méronymies (Winston et al.)

Dans la littérature sur l'anaphore associative, celle se concentrant sur la relation d'une partie à un tout - ou référence indirecte par association pour Clark - est la plus abondante. Les travaux de [Winston et al., 1987] sont l'illustration de cette affirmation. Leur typologie des relations partie-tout distingue les relations méronymiques des relations non méronymiques. La méronymie est définie comme étant la relation présentant au moins une des trois caractéristiques suivantes :

- fonctionnalité : la partie remplit une fonction par rapport au tout.
- homéomère : les parties sont similaires entre elles et par rapport au tout auquel elles appartiennent.

– séparabilité : les parties peuvent physiquement être déconnectées du tout.

Nous reproduisons en figure 6.1 simplement le tableau résumant la classification, avec les exemples donnés dans l'article :

Relation	Exemple	Fonctionnel	Homéomère	Séparable
Composant/Objet	anse - tasse	+	-	+
Membre/Collection	arbre - forêt	-	-	+
Portion/Masse	part - gâteau	-	+	+
Matière/Objet	cuir - valise	-	-	-
Trait/Activité	achat - paiement	+	-	-
Sous-lieu/Lieu	oasis - désert	-	+	-

FIG. 6.1 – Relations méronymiques chez Winston et al.

[Winston et al., 1987] mentionnent aussi des relations partie-tout qu'ils considèrent comme n'étant pas méronymiques. Nous ne détaillerons pas ici la définition des relations, mais nous nous contentons de les énumérer, leurs noms étant relativement explicites. Ces relations sont les suivantes :

– inclusion topologique (localisations) : prisonnier - cellule
– inclusion de classe : roses - fleurs
– attributs : bâtiment - hauteur
– attachements : doigts - mains
– possession : millionnaire - argent

6.1.3 Typologie de Kleiber

Kleiber, [Kleiber, 1997, Kleiber, 2001] propose une typologie basée à la fois sur les théories de Clark et de Winston, mais montre qu'elles ne sont pas satisfaisantes. Il propose donc une nouvelle classification en cinq grandes catégories, auxquelles sont associées des batteries de tests.

Anaphores associatives entre membre et collection Cette relation est directement inspirée de [Winston et al., 1987, Clark, 1977], et les exemples donnés sont identiques. Il s'agit de la relation entre un élément et un ensemble, illustrée par la relation entre *couple* et *mari*.

Anaphores méronymiques L'anaphore est une partie de l'antécédent, la relation est illustrée par le couple *un arbre - le tronc*.

Les tests donnés par Kleiber sont les suivants :

X est un méronyme de Y si :

- la phrase générique *Un X est une partie de Y* est possible. (Un tronc est une partie d'arbre.)

- la phrase *C'est un X de Y* est possible. (C'est un tronc d'arbre.)

Anaphores locatives le référent de l'anaphore est « situé » dans le référent de l'antécédent et la relation est illustrée par la relation entre *un village - l'église*.

Une anaphore associative est locative si les phrases suivantes sont possibles :

- Un X a un Y pour ...

- Dans un X il y a un Y.

- Un X inclut un Y.

et si la phrase suivante est impossible :

Un Y est une partie de X.

Kleiber distingue ensuite les anaphores associatives locatives canoniques des anaphores associatives locatives facultatives de la manière suivante (nous citons) :

Il y a relation canonique si :

Les éléments X et Y peuvent figurer dans la structure générique « Dans un X il y a un Y ».

Les éléments X et Y ne peuvent pas s'insérer dans la structure générique « Un X est une partie d'un Y ».

La relation est facultative si :

La phrase générique « Dans un X il y a un Y » et sa négation sont également fausses.

Les éléments X et Y ne peuvent pas s'insérer dans la structure générique « Un X est une partie d'un Y ».

Anaphores fonctionnelles L'exemple prototypique d'anaphore fonctionnelle est la relation entre *automobile* et *conducteur*. Pour Kleiber, l'impossibilité de certains noms à être utilisés dans des anaphores associatives est liée au fait qu'ils ne sont pas fonctionnels, comme dans les phrases suivantes :

(96) ***La voiture*** *dérapa et s'écrasa contre un platane.* ***Le conducteur*** *fut éjecté.*

(97) *?* ***La voiture*** *dérapa et s'écrasa contre un platane.* ***L'automobiliste*** *fut éjecté.*

Sont considérés comme des noms fonctionnels des noms comme *auteur, metteur en scène* et *conducteur* alors que des noms comme *automobiliste, écrivain* et *cinéaste* ne le sont pas. Pour distinguer les premiers des seconds, Kleiber cherche alors les propriétés des noms fonctionnels. Les propriétés dégagées sont les suivantes :

Si Ni est un nom fonctionnel et Nj le nom anaphorique, les formes suivantes seront possibles :
Le Ni de Nj (Le conducteur d'automobile)
Le Ni d'un Nj (Le conducteur d'une automobile)
Son Ni (à Nj) (Son conducteur - à l'automobile)

Les noms non fonctionnels n'entrent pas dans la catégorie des noms relationnels, c'est-à-dire des noms qui sont en fait des prédicats binaires (ou concepts fonctionnels à deux places). En effet, dans l'expression *X est le conducteur de Y*, le nom conducteur met en relation X et Y. En revanche, la forme *X est l'automobiliste de Y* n'est pas acceptable parce que le nom *automobiliste* ne peut pas mettre en relation X et Y. Les noms non fonctionnels ne prennent pas X comme un argument mais dénotent une propriété de X.

Anaphores actancielles Cette catégorie est la même que celle que Clark appelle *référence indirecte par caractérisation*. L'anaphore est un actant de l'événement décrit précédemment, un exemple serait le lien entre *couper* et *le couteau* ou *assassiner* et *le meurtrier*.

6.2 Problèmes posés par les données théoriques

Si les relations décrites dans la littérature ont été utiles à une première analyse de corpus, elles ne sont pas suffisantes pour des applications informatiques. Notre critique porte sur deux points : après une analyse de corpus, on constate que de nombreux cas ne sont pas pris en compte par ces typologies. Par ailleurs, aucune de ces typologies ne permet d'identifier les sources des inférences nécessaires à la résolution ou à la production de l'anaphore associative, c'est à dire l'élément de la base de connaissances qui contient les informations nécessaires pour bâtir les inférences permettant de résoudre ou générer l'anaphore.

6.2.1 Exemples non couverts par les typologies existantes

Au cours de la première analyse de corpus, nous avons relevé un certain nombre d'anaphores associatives non couvertes par les diverses typologies existantes. Quatre exemples emblématiques sont reproduits ci-dessous. Nous avons donné, en gras dans les couples antécédent/anaphore, la relation entretenue par les deux noms :

- Opération [**suivie par**] Convalescence
- Athlétisme [**de**] Fédération Nationale
- Question [**à**] Réponse
- Enquête [**basée sur**] Témoignages

Ces exemples contrastent avec les typologies existantes qui finalement ne tiennent compte que de trois grands types de relation sémantique entre l'antécédent et l'anaphore : l'inclusion, la possession, et la relation actancielle. Une typologie destinée à la génération automatique de textes doit donc tenir compte des relations standard décrites dans la littérature autant que des relations citées ici.

6.2.2 Sources de l'inférence

Le second problème posé par les typologies existantes est le suivant : si on veut générer du texte à base de moteurs d'inférence, et permettre la génération d'anaphores associatives, on a besoin de connaître les sources possibles de l'inférence. Il apparaît clairement que l'inférence peut se faire à partir de sources variées, comme les relations lexicales (méronymie), les connaissances lexicographiques, la grille thématique d'un événement, ou à partir d'un savoir encyclopédique plus large (nous détaillons et illustrons ceci dans la section suivante).

6.3 Nouvelle classification

Nous avons vu aux chapitres 3 et 5 que la priorité lorsqu'on annote des corpus est de fournir des définitions opérationnelles des phénomènes à rechercher. Notre étude de corpus nous a donc amenée à établir notre classification sur une combinaison de paramètres utilisables en traitement automatique des langues et facilement identifiables par des tests lors de l'annotation :

- la relation sémantique entretenue entre l'antécédent et l'anaphore
- le type ontologique des référents de l'antécédent et de l'anaphore
- la source épistémique à partir de laquelle on fait les inférences pour établir le lien entre l'antécédent et l'anaphore.

Cette classification est résumée dans le tableau 6.2, où les abbréviations ont les significations suivantes : I signifie « individu », E, « événement », Features, « attribut », T, « temps », Loc, « lieu ».

Catégorie	Relation sémantiques	Types de l'antécédent et de l'anaphore	Source
Ensemble / élément	$\in, \subset$	$\langle I, \mathcal{P}(I)\rangle$	Hyponymie
Thématiques	Rôles Thématiques agent, patient etc.	$\langle I, E\rangle$ (or $\langle E, I\rangle$)	Grille thématique de l'événement
Définitionelles			
Indiv./attribut	donnée par la déf.	$\langle I$, *Features*$\rangle$	Déf. lexico.
Associé/indiv.	donnée par la déf.	$\langle I, I\rangle, \langle I, E\rangle$	Déf. lexico.
Méronymie	partie de	$\langle I, I\rangle, \langle E, E\rangle$ $\langle I, \mathcal{P}(I)\rangle$	Méronymie
Coparticipants	donnée par la déf. de l'antécédent et de l'anaphore	$\langle I, I\rangle, \langle I, E\rangle$	Déf. lexico.
Non lexicales			
Circonstancielles	spatiales ou temporelles	$\langle I, Loc\rangle, \langle I, T\rangle$	Structure du discours
Connaissances encyclopédiques	donné par les CE	Indifférent	Connaissances encyclopédiques

Fig. 6.2 – Typologie proposée

6.3.1 Relation ensemble / élément

Cette classe couvre les cas où l'anaphore est un élément ou un sous-ensemble d'un ensemble. La relation est une relation d'inclusion entre deux objets de type INDIVIDU. Une des sources de l'inférence peut être la relation lexicale d'hyponymie.

- *séminaires - le dernier séminaire*
- *La CGT et FO - FO*

- L'armée - le tiers

(c-5) *Sinon, elle pourrait se compliquer car les deux ressortissants allemands sont bel et bien entre les mains* **des proches des deux détenus en RFA** *et en premier* **du frère de ceux-ci**.

6.3.2 Relation thématique

L'anaphore peut être reliée à l'antécédent par une relation thématique définie dans la grille thématique de l'événement antécédent. Ainsi dans l'exemple suivant, l'anaphore *les victimes* est reliée par la relation thématique *patient* de l'événement dénoté par le nom *attaque*.

(c-6) *Les enquêteurs ont par ailleurs identifié les auteurs de* ***l'attaque*** *grâce aux témoignages* ***des victimes***.

6.3.3 Relations définitionnelles

Dans ces cas, la relation implicite entre l'antécédent et l'anaphore est donnée dans la définition lexicographique de l'un, de l'autre ou des deux. Ainsi, une *convalescence* peut être définie comme la période suivant une *opération*, et la relation est alors une relation de succession temporelle. Cette catégorie couvre de nombreux cas identifiés dans la littérature qui peuvent être différenciés par des critères ontologiques. Nous distinguons trois types de relations définitionnelles.

6.3.3.1 Relation méronymique

Cette relation est dans la plupart des cas identifiée par
[Winston et al., 1987] et peut être exprimée par les phrases *« X est une partie de Y »* et *« Y a un X »*. Cette relation impose que les deux éléments impliqués aient le même type ontologique. Une illustration est présente dans l'exemple suivant, où l'on peut dire que *le gaz propulseur* est une partie des *bombes à aérosols*.

(c-7) *Huit associations suisses de consommateurs et de protection de l'environnement ont lancé, vendredi 4 septembre, une campagne de boycottage* ***des bombes à aérosols***. ***Le gaz propulseur*** *est, en effet, fait de chlorofluorocarbones dont on pense qu'ils détruisent l'ozone de la haute atmosphère.*

De plus on considère comme méronymique la relation INDIVIDU/FONCTION, qui donne le métier ou la fonction d'un objet de type INDIVIDU par rapport à un autre objet de ce type. Un exemple de ce type de relation est la relation entre un *club de football* et *le président*.

La relation sémantique impliquée est donc une relation d'inclusion.

6.3.3.2 Relation individu/attribut

Cette relation est tout d'abord caractérisée par une relation d'implication. Cette relation implique un lien entre un objet de type INDIVIDU et un objet vu comme un trait

pouvant prendre une valeur dans un ensemble fini de valeurs (une personne - le regard, comme dans l'exemple suivant).

(c-8) ***Pierre-Marie Valentin*** *n'a décidément rien de ces hommes à poigne, cœur sec et mise austère qui peuplent la galerie des grands redresseurs d'entreprise. Il garde même dans l'allure et* ***le regard****, curieusement juvéniles, quelque chose de ces adolescents aux tempes argentées qu'incarne Jacques Perrin à l'écran.*

6.3.3.3 Relation individu/associé

Cette relation est aussi une relation d'implication entre deux individus, deux événements ou entre un individu et un événement. Elle se trouve par exemple dans le couple *enseignant - classe*, ou dans le couple *pouvoir - opposition* de l'exemple suivant.

(c-9) *Autre maladresse* ***du pouvoir*** *: les interventions dans le classement des clubs de football, censées, selon* ***l'opposition****, assurer la promotion du bulletin « non » dans les régions de clubs passés, sur intervention ministérielle, en première division.*

6.3.4 Relations entre coparticipants

Il existe des relations où le lien entre l'antécédent et l'anaphore est donné par les deux définitions. Par exemple, les noms *otage* et *ravisseur* sont liés par la relation *enlèvement* qui peut être reconstruite par les définitions du [Trésor de la langue française informatisé] des noms *otage* et *ravisseur* :

Ravisseur : Celui, celle qui enlève une personne de force.
Otage : Personne dont on s'est emparé et qui est utilisée comme moyen de pression, de chantage.

(c-10) *La libération* ***des deux otages ouest-allemands, MM. Rudolf Cordes et Alfred Schmidt****, pourrait être imminente. Un message des* ***ravisseurs****, accompagné de la photo d'un des deux captifs, M. Schmidt, méthode usuelle d'authentification utilisée à Beyrouth, annonce en effet qu'elle interviendrait pour l'un d'eux dans les dix jours « si le gouvernement allemand tient ses engagements ».*

6.3.5 Relations non lexicales

Les relations non lexicales nécessitent des connaissances extralinguistiques importantes.

6.3.5.1 Relation circonstancielle

Cette relation est donnée par la structure du discours, et relie un référent de type INDIVIDU à un LIEU ou un MOMENT. On trouve ce type de lien dans le couple *Besançon - la région.*

(c-11) *Dominique Gutknecht n'a pu être retrouvé, malgré l'importance des moyens mis en œuvre depuis près d'une semaine à* ***Besançon*** *et dans* ***la région****.*

6.3.5.2 Relation passant par les connaissances encyclopédiques

Cette relation n'est établie que grâce à nos connaissances encyclopédiques. Elles relient des éléments de n'importe quel type entre eux. C'est le cas par exemple du couple *licenciement - procédure*.

(c-12) *C'était une loi d'équilibre qui permettait aux différents partenaires - dirigeants, syndicats, collectivités - de discuter de l'avenir de l'entreprise en cas de* ***licenciement économique****. Il faudra revenir à quelque chose d'analogue, en assouplissant* ***les procédures*** *pour les PME.*

6.4 Etude de corpus - Résultats et discussion

6.4.1 Etude de corpus

Nous avons utilisé la classification décrite ci-dessus comme schéma d'annotation pour le corpus.

Classe	Nb	Prop
Ensemble / élément	45	11,5%
Thématique	12	3,1%
Définitionelle	262	66,8%
Indiv./Attribut	14	3,6%
Associé/Indiv.	112	28,6%
Meronymique	136	34,7%
Coparticipants	37	9,4%
Non lexicale	36	9,2%
Circonstantielle	18	4,6%
Encyclopédique	18	4,6%

FIG. 6.3 – Relations associatives dans le corpus

6.4.2 Remarques générales

Nous présentons les résultats de l'annotation dans le tableau 6.3. Cette annotation nous permet de montrer que les sources des inférences sont nombreuses, mais identifiables. Nous distinguerons trois types de sources d'inférence : celles pour lesquelles il existe déjà des outils de traitement automatique des langues, celles dont la source est le dictionnaire, et celles qui sont plus problématiques, car elles font appel au savoir encyclopédique ou à la structure du discours.

6.4.3 Relations calculables à partir d'outils de TAL

La majorité des relations associatives trouvées dans le corpus sont récupérables à partir d'outils de TAL existants (49% du total). On trouve trois grandes catégories de relations

de ce type :

Méronymie L'étude de corpus confirme l'importance de la relation de méronymie dans les diverses réalisations d'anaphores associatives. Cette relation est encodée dans Wordnet [Fellbaum, 1998], ce qui est intéressant puisque la base de connaissances nécessaire à la construction d'inférences est disponible. Environ un tiers des relations méronymiques trouvées dans le corpus sont déja présentes dans Wordnet. Ceci peut paraître insuffisant, mais en étudiant de plus près ces relations, nous nous sommes aperçue que les relations manquantes étaient peu nombreuses, mais très souvent répétées. Il s'agit essentiellement de relations entre villes et parties de villes, pays et partie de pays, ou de parties d'entreprises. Il semble alors réalisable d'étendre Wordnet aux domaines appropriés, en y intégrant ces concepts.

Relations thématiques Ces relations pourtant très fréquemment mentionnées dans la littérature ne sont que très peu représentées (3%). Cependant, il s'agit de relations productives et complètement formalisées dans Framenet [Baker et al., 1998]. Framenet est un outil qui, étant donné un verbe, lui associe une grille thématique ou un script, et spécifie les rôles compatibles avec l'événement qu'il décrit. Nous avons trouvé environ les trois quarts des relations thématiques du corpus dans Framenet. Ce résultat est encourageant, bien qu'il faille encore trouver le moyen d'interroger Framenet pour calculer la relation.

Relations d'ensemble Ces relations représentent 11% des relations associatives et sont relativement faciles à calculer. En effet, elles sont réalisées soit par une relation d'hyponymie, soit par une reprise directe avec un quantifieur numérique.

6.4.4 Relations impliquant des connaissances lexicographiques

Ces relations sont nombreuses (42% des relations associatives du corpus). Elles regroupent trois types de relations, qui ont toutes pour source la définition du dictionnaire. Nous avons pu vérifier la réalité des relations dans toutes les définitions des termes du corpus en consultant [Le Petit Robert] et le [Trésor de la langue française informatisé].

Relations individu/attribut Cette relation (4% des cas) trouve son origine dans la définition de l'attribut.

Relations individu/associé Cette relation - relativement fréquente dans le corpus (28% des cas) - trouve sa source soit dans la définition de l'antécédent, soit dans la définition de l'anaphore. Elle est donc plus complexe à rechercher, mais au plus deux définitions doivent être consultées.

Relation entre coparticipants Dans le cas de ces relations manifestement plus difficiles à identifier (9% selon les annotateurs), plusieurs définitions de dictionnaire doivent être consultées. Il faut donc consulter les définitions de l'antécédent et de l'anaphore et

trouver le terme permettant de les relier. Il sera peut-être nécessaire de les relier avec plus d'un terme.

6.4.5 Relations impliquant des connaissances du monde

Ces relations, qu'il s'agisse des relations circonstancielles ou des relations liées à des connaissances plus larges, nécessitent des représentations structurées des connaissances. Bien que leurs proportions soient relativement faibles, elles sont non négligeables (9%).

6.5 Application à l'algorithme de Gardent et Striegnitz

Nous avons désormais une typologie pour les anaphores associatives qui nous a permis de classer la totalité des cas identifiés dans notre corpus. Nous sommes donc parvenue aux deux buts que nous nous étions fixés au début de ce chapitre.

Nous avons dans un premier temps défini des catégories d'anaphores associatives couvrant de manière exhaustive notre corpus. Les définitions des catégories se sont révélées assez précises pour que deux annotateurs puissent arriver à un accord sur la classification d'une anaphore associative. Cette typologie est actuellement testée sur un corpus en allemand, le corpus NEGRA. Une comparaison des résultats pourrait être une perspective intéressante à ce travail.

Par ailleurs, le deuxième objectif que nous avons atteint est l'identification de trois éléments importants pour la génération de textes :

- la nature des référents impliqués
- la nature de la relation sémantique entre l'antécédent et l'anaphore
- la source permettant d'inférer la relation entre antécédent et anaphore

Pour finir, il est nécessaire de lier ces travaux avec un algorithme de génération adapté à la génération d'anaphores associatives. Dans cette section, nous montrons comment les recherches menées sur corpus peuvent permettre d'étendre l'algorithme proposé par [Gardent et Striegnitz, 2003]. Notre apport porte sur deux points : l'identification et la structuration des connaissances servant de base à l'inférence, et l'identification des relations entretenues par l'ancre et la cible.

6.5.1 Identification des bases de connaissances

Pour pouvoir générer des anaphores associatives (et des expressions référentielles de manière générale), il est nécessaire d'avoir un modèle structuré du contexte et de formaliser les notions de familiarité et d'unicité en fonction de ce modèle (cf. chapitre 2). Le modèle du contexte de [Gardent et Striegnitz, 2003] est constitué de la manière suivante :

- Le modèle du discours (discours déjà produit - DM)
- Les connaissances du monde (savoir partagé par le locuteur et l'interlocuteur - WKL)
- Le modèle du locuteur (savoir additionnel du locuteur - SM)

Cette structuration du contexte leur permet de construire un ensemble de connaissances du monde, dans lesquelles doivent être mentionnées les relations entre les objets qui permettent la génération d'anaphores associatives. D'ailleurs, [Gardent et Striegnitz, 2003] ne

prennent en compte pour les relations associatives que les relations méronymique. Pour mémoire, voici l'exemple cité dans la première partie de notre thèse où les conditions de familiarité et d'unicité sont satisfaites :

```
DM = (restaurant (r))
WKL = (∀x (restaurant(x)) → ∃y(cook(y) ET part-of(x,y)))
SM = (cook(c), part-of(c,r))
Cible = c
Description = cook(c)
```

Nous voyons dans cet exemple que la relation encodée dans la base de connaissance est la relation part-of. Cette relation a été encodée à la main ou inférée à partir d'un outil de type Wordnet [Fellbaum, 1998]. Suite à notre étude des anaphores associatives, nous avons défini une série de classes d'anaphores associatives, en fonction du type des référents, de leurs relations sémantiques, et de la source permettant d'identifier les relations sémantiques.

Nous proposons donc de diviser les connaissances partagées du locuteur en plusieurs parties, ce qui nous permet de varier les relations possibles, et les sources à partir desquelles ces relations sont calculées :

```
WKL = {Connaissances lexicographiques ⋃ Connaissances lexicales
(Framenet ⋃ Wordnet) ⋃ Connaissances du monde (Connaissances générales
⋃ connaissances du domaine)}
```

Afin que l'algorithme fonctionne, toutes ces bases de connaissances doivent être formulées en logique du premier ordre, comme le sont les exemples que nous avons donné précédemment.

Les connaissances sont donc divisées en trois sous-parties que nous détaillons maintenant :

6.5.1.1 Les connaissances lexicographiques

Il s'agit des connaissances issues des définitions de dictionnaires. Ainsi, pour un exemple comme l'exemple c-13, si l'on dispose d'une représentation sémantique des définitions de dictionnaire, on peut retrouver la relation entretenue par le nom *otage* et le nom *ravisseurs* :

(c-13) *La libération* ***des deux otages ouest-allemands, MM. Rudolf Cordes et Alfred Schmidt****, pourrait être imminente. Un message des* ***ravisseurs****, accompagné de la photo d'un des deux captifs, M. Schmidt, méthode usuelle d'authentification utilisée à Beyrouth, annonce en effet qu'elle interviendrait pour l'un d'eux dans les dix jours « si le gouvernement allemand tient ses engagements ».*

Les définitions trouvées dans le [Trésor de la langue française informatisé] sont les suivantes :

Ravisseur : Celui, celle qui enlève une personne de force.

Otage : Personne dont on s'est emparé et qui est utilisée comme moyen de pression, de chantage.

En reprenant les deux définitions, si elles sont formalisées en logique du premier ordre, on peut envisager, à partir de leur contenu, de construire des inférences permettant de relier les deux référents.

Ce type de travail est réalisable à condition de pouvoir analyser les définitions des dictionnaires afin d'extraire une représentation sémantique de leur contenu. Ce type de travail est actuellement en cours à l'ATILF, à partir des travaux de [Deschizeaux et Reb, 1995].

6.5.1.2 Les connaissances lexicales

Les connaissances lexicales impliquées dans les anaphores associatives sont de deux types : il s'agit des connaissances intégrant les relations méronymiques, qu'on retrouve dans un outil comme Wordnet [Fellbaum, 1998] et les relations thématiques qui sont encodées dans Framenet [Baker et al., 1998]. Dans cette section, nous montrons des exemples d'utilisation de ces deux outils.

Identification de la méronymie avec Wordnet Dans l'exemple suivant, on trouve une relation d'anaphore associative entre *pavillon* et *fenêtre*. Ce type d'anaphore associative est basée sur la méronymie, et est encodée dans Wordnet. La figure 6.4 reproduit une partie des résultats de la recherche des méronymes de l'entrée lexicale *maison* dans Wordnet (nous avons en effet dans un premier temps dû rechercher un terme plus générique que *pavillon*). Les éléments HOUSE (pavillon) et WINDOW (fenêtre) étant ceux que nous souhaitions montrer parmi tous les résultats ont été mis en majuscule. Dans la réelle sortie de Wordnet, seules les relations de méronymie sont exprimées en majuscules par les termes HAS-PART (partie de) ou HAS MEMBER (dans le cas de relations d'ensemble).

(c-14) *Tous les autres membres de la délégation présents cette nuit-là dans* ***le pavillon****, soit neuf personnes au total, sont capturés par le commando, à l'exception d'un seul, Touviah Sokolovsky, qui a eu la présence d'esprit de sauter en pyjama par* ***la fenêtre****.*

Identification des relations thématiques avec Framenet Si on considère l'exemple c-15, on constate une anaphore associative entre *attaque* et *victime*. La figure 6.5 montre l'entrée du nom *attaque* dans Framenet. On y retrouve clairement mentionné que l'un des participants à l'événement d'*attaque* est *la victime*. Ici encore, pour des raisons de lisibilité, nous avons mis le terme VICTIM en majuscule, alors qu'il n'est pas mis en relief dans la sortie de Wordnet habituelle. Il est donc possible d'utiliser ce type de ressources, si les données sont formalisées en logique du premier ordre, pour générer une anaphore associative impliquant ce type de relation.

(c-15) *Les enquêteurs ont par ailleurs identifié les auteurs de* ***l'attaque*** *grâce aux témoignages* ***des victimes****.*

6.5.1.3 Les connaissances du monde

Connaissances du domaine Les connaissances du domaine sont des connaissances liées à l'application pour laquelle est utilisé le générateur. On doit les exprimer manuel-

```
Results for "Meronyms (parts of this), inherited" search of noun "house"

Sense 1
HOUSE - (a dwelling that serves as living quarters for one or more fami-
lies; "he has a house on Cape Cod"; "she felt she had to get out of the house")
          => dwelling, home, domicile, abode, habitation, dwelling house - (hou-
sing that someone is living in; "he built a modest dwel-
ling near the pond"; "they raise money to provide homes for the homeless")
HAS PART: bathroom, bath - (a room (as in a residence) containing a bath or sho-
wer and     HAS PART: toilet, can, commode, crapper, pot, potty, stool, throne -
(a plumbing fixture for defecation and urination)
    HAS PART: bedroom, sleeping room, chamber, bedchamber -
(a room used primarily for sleeping)
        => building, edifice - (a struc-
ture that has a roof and walls and stands more or less perma-
nently in one place; "there was a three-story building on the cor-
ner"; "it was an imposing edifice")
         (...)
    HAS PART: wall - (an architectural partition with a height and length grea-
ter than its thickness; used to divide or enclose an area or to support ano-
ther structure; "the south wall had a small window"; "the walls were cove-
red with pictures")
    HAS PART: arch, archway - (a passageway under an arch)
    HAS PART: capstone, copestone, coping stone, stretcher -
(a stone that forms the top of wall or building)
     HAS PART: door - (a swinging or sliding barrier that will close the en-
trance to a room or building; "he knocked on the door"; "he slam-
med the door as he left")
       (...)
       HAS PART: WINDOW - (a framework of wood or metal that contains a glass win-
dowpane and is built into a wall or roof to admit light or air)
        HAS PART: casing, case - (the enclosing frame around a door or window ope-
ning; "the casings had rotted away and had to be replaced")
        HAS PART: mullion - (a nonstructural vertical strip between the case-
ments or panes of a window (or the panels of a screen))
          HAS PART: pane, pane of glass, window glass -
(sheet glass cut in shapes for windows or doors)
             HAS PART: sash, window sash -
(a framework that holds the panes of a window in the window frame)
             HAS PART: sash fastener, sash lock, window lock - (a lock atta-
ched to the sashes of a double hung window that can fix both in the shut position)
             HAS PART: window frame - (the framework that supports a window)
              HAS PART: windowpane, window -
(a pane in a window; "the ball shattered the window")
         HAS PART: crawlspace, crawl space - (low space beneath a floor of a buil-
ding; gives workers access to wiring or plumbing)
       (...)
```

FIG. 6.4 – Liste des méronymes du nom « maison » dans Wordnet

```
Definition:
An Assailant physically attacks a VICTIM (which is usually but not always sen-
tient), causing or intending to cause the Victim physical damage. A Wea-
pon used by the Assailant may also be mentioned, in addi-
tion to the usual Place, Time, Purpose, Reason, etc. Sometimes a loca-
tion is used metonymically to stand for the Assailant or the Vic-
tim, and in such cases the Place FE will be annotated on a second FE layer.

As soon as he stepped out of the bar he was SET upon by four men in ski-masks.

Is he INVADING Iraq just to cover other shortcomings?

Then Jon-O's forces AMBUSHED them on the left flank from a line of low hills.

FEs:
Core:

Assailant [Asl] The person (or other self-
directed entity) that is attempting physical harm to the Victim.

The mysterious fighter ATTACKED the guardsmen with a sabre.

VICTIM [Vic] This FE is the being or entity that is injured by the Assai-
lant's attack.

The mysterious fighter ATTACKED the guardsmen with a sabre.

Weapon [Wep] An entity used by the Assailant to cause damage to the Victim.

The mysterious fighter ATTACKED the guardsmen with a sabre.
```

FIG. 6.5 – Grille thématique du verbe « attaquer » dans Wordnet

lement en logique du premier ordre, ce qui n'est pas forcément une tâche trop complexe dans la mesure où la somme de connaissances nécessaires est limitée.

Connaissances générales Les connaissances du monde posent plus de problèmes que les connaissances du domaine. On ne connaît pas d'autre moyen que celui de les construire manuellement, la difficulté étant qu'elles sont très nombreuses et qu'on n'a pas d'indication sur la façon de les représenter : on ne sait pas, par exemple, jusqu'à quel niveau de détail on doit les représenter. Cependant, elles sont nécessaires, parfois même pour pouvoir relier entre elles certaines connaissances liées au domaine spécifique de l'application.

6.5.1.4 Recherche des relations

Les relations entre les objets étant encodées dans les diverses sources de connaissances, nous pourrions désormais générer des anaphores associatives du même type que celles trouvées dans notre corpus (nous simplifions l'exemple) :

(98) ***Le pouvoir*** *serait intervenu dans le classement des clubs de football, afin, selon* ***l'opposition****, d'assurer la promotion du bulletin « non » dans les régions de clubs passés, sur intervention ministérielle, en première division.*

```
DM = (pouvoir (p))
WKL = (
Connaissances Lexicographiques : (∀x (pouvoir(x)) → ∃y(opposition(y) ET
associé(x,y)))
Connaissances lexicales  : ()
Connaissances du monde : ()
)
SM = (pouvoir(p), associé(p,o))
Cible = o
Description = opposition(o)
```

6.5.2 Organisation de la recherche des relations

A partir des données du corpus, on peut par ailleurs optimiser le processus de génération en ordonnançant la recherche des informations dans les différentes bases de connaissances. L'étude de corpus peut nous donner des indications sur les bases de connaissances à privilégier. Elle nous montre en effet la fréquence d'utilisation des sources de connaissances. Nous pouvons utiliser ces fréquences comme un moyen de décider de l'ordre dans lequel on effectue les recherches dans les bases de connaissances. Si nous reprenons les résultats présentés dans le tableau 6.3, en les classant en fonction de la base de connaissances dans laquelle apparaît la relation, nous obtenons les chiffres présentés dans le tableau 6.6.

Ce tableau nous indique alors que la première base de connaissances à utiliser est Wordnet (46,2% des relations y sont codées) puis la base lexicographique (41,6% des relations) ; les connaissances du monde sont à consulter seulement après (elles contiennent 9,2% des relations impliquées dans les anaphores associatives), et enfin, on doit interroger la base de connaissances encodée dans Framenet (3,1% des relations).

Classe	Nb	Prop
Wordnet	181	46,2%
Ensemble / élément	45	11,5%
Méronymique	136	34,7%
Framenet	12	3,1%
Thematique	12	3,1%
Dictionnaire	163	41,6%
Indiv./attribut	14	3,6%
Associé/indiv.	112	28,6%
Coparticipants	37	9,4%
Connaissances du monde	36	9,2%
Circonstantielle	18	4,6%
Encyclopédique	18	4,6%

FIG. 6.6 – Relations associatives classées par source de connaissances

Ceci peut se formaliser dans le point n°7 de l'algorithme de Gardent et Striegnitz de la façon dont nous le représentons en figure 6.7.

```
7. Si but-courant ∈ terms(DM) alors R ← DM sinon Ctxt ← DM + SM
8. Essayer de sélectionner une formule atomique p applicable telle que
Ctxt+WKL ⊨ p, en recherchant prioritairement la relation de la façon
suivante :
Ctxt + Wordnet ⊨ p
Ctxt + base lexicographique ⊨ p
Ctxt + connaissances du monde ⊨ p
Ctxt + Framenet ⊨ p
```

FIG. 6.7 – Proposition d'extension à l'algorithme de Gardent et Striegnitz

6.6 Conclusion

Après avoir établi que l'analyse de corpus permet d'identifier clairement des relations associatives et d'en retrouver les sources, nous avons pu montrer qu'à partir du moment où les bases de connaissances sur lesquelles se construisent les inférences nécessaires à la production de langage naturel sont formalisées, il est possible d'appliquer des données théoriques à un algorithme de génération existant.

Ces résultats sont positifs, dans la mesure où la plupart des relations associatives sont calculables à partir de données existantes et formalisées pour le traitement automatique. Cependant, un des problèmes à poser rapidement est celui de la structuration de dictionnaires ou de lexiques électroniques de façon à ce qu'ils soient utilisables comme base aux inférences nécessaires dans le cas des relations définitionnelles, en les représentant par exemple en logique du premier ordre.

Chapitre 7

Statut informationnel des descriptions coréférentielles

Les travaux décrits au chapitre 5 nous ont permis d'extraire un grand nombre de description définies et démonstratives coréférentielles classées sous le terme de « reclassification ». Ceci nous confirme l'idée que la coréférence est basée en partie sur des processus inférentiels dont il convient d'identifier les sources. Par ailleurs, notre étude nous a permis de constater que quelle que soit la relation entretenue entre l'antécédent et la reprise, il peut arriver que la reprise contienne plus d'information que son antécédent. Il nous a semblé alors nécessaire de faire une étude plus approfondie et systématique de ces observations.

Dans cette section, nous présenterons donc dans un premier temps la question de l'anaphore utilisée comme un support permettant de communiquer de l'information nouvelle sur son référent (section 7.1). Nous montrerons ensuite pourquoi ce problème est important en génération (section 7.2), puis nous présenterons une classification des reprises définies et démonstratives tenant compte non seulement de la capacité de l'anaphore à communiquer de l'information nouvelle, mais aussi de la source d'inférence utilisée pour produire l'information connue ou nouvelle dans l'anaphore (section 7.3). Nous ferons ensuite un bilan de l'étude de corpus que nous avons menée en utilisant cette classification (section 7.4) et nous conclurons ce chapitre par une extension de l'algorithme de [Gardent et Striegnitz, 2003] en intégrant les résultats de notre étude (section 7.5).

7.1 L'anaphore vue comme un support à de l'information nouvelle

Afin de cerner complètement le problème de la distribution de l'information nouvelle et de l'information donnée dans les reprises définies et démonstratives, considérons les reprises anaphoriques suivantes :

(99) ***Paul*** *a été agressé.* ***La victime*** *se rendait à son bureau lorsqu'un malfaiteur a surgi.*

(100) ***Paul*** *a été agressé en allant au cinéma.* ***Ce professeur de mathématiques*** *est un cinéphile.*

Ces deux phrases illustrent bien les données connues sur les emplois du défini et du démonstratif [Corblin, 1987, Kleiber, 1986, Kleiber, 1988]. Dans le premier texte, le syntagme « Paul » est repris par un groupe nominal défini dont le contenu sémantique permet d'identifier le référent sans ambiguïté dans la mesure où il est le seul à correspondre à la description dénotée par le syntagme (le nom « victime » étant un moyen de référer à une personne ayant subi une agression). Dans le second texte, le syntagme « Paul » est repris par un groupe nominal démonstratif dont le contenu sémantique illustre la propriété « reclassifiante » du démonstratif, qui permet la coréférence entre deux syntagmes sans qu'ils entretiennent une relation lexicale particulière (nous considérons que « Paul » et « la victime » entretiennent une relation lexicale par le biais du verbe dans le premier texte).

Ces deux exemples illustrent les données issues de la littérature sur les déterminants et montrent que deux syntagmes coréférents peuvent ne pas avoir le même contenu sémantique. Dans un cas cependant, l'anaphore n'apporte pas d'information nouvelle sur le référent (le fait que Paul soit une victime est inféré à partir du contexte), tandis que dans l'autre, l'anaphore sert de support à de l'information nouvelle (rien ne peut laisser deviner que Paul est professeur de mathématiques). On pourrait penser que la possibilité d'ajouter de l'information sur le référent est une propriété du syntagme démonstratif, mais si l'on considère les exemples suivants, on constate que ce n'est pas le cas :

(101) ***Paul*** *a été agressé.* ***Le professeur de mathématiques de ma sœur*** *se rendait à son bureau lorsque le malfaiteur a surgi.*

(102) ***Paul*** *a été agressé en allant au cinéma.* ***Cet homme*** *a toujours des ennuis.*

L'idée selon laquelle une reprise anaphorique peut apporter de l'information sur les référents n'est pas inexistante dans la littérature : on en trouve des mentions et des définitions dans les travaux de [Danlos, 1999] et [Danlos et Gaiffe, 2000] sur le discours particularisant et généralisant, ainsi que dans ceux de [Wiederspiel, 1994] et [Corblin, 1987] à propos de l'attribution de propriétés par le démonstratif ou de la possibilité de faire des métaphores. Cependant, cette idée ne sert jamais de base aux études sur la référence. Les travaux de Danlos sur la coréférence événementielle donnent des définitions, mais parlent finalement peu des reprises nominales en elles-mêmes. On ne trouve pas de comparaison en termes linguistiques des contextes d'apparition du phénomène, ni d'étude pour savoir quel est le déterminant approprié en fonction du contexte.

Par ailleurs, ces données ne sont pas utilisées en génération, alors que l'informativité de l'énoncé est un élément important en sémantique computationnelle, autant que la consistance [Blackburn, 2003]. L'énoncé généré doit être informatif, et l'anaphore peut permettre qu'il le devienne. Un des moyens linguistiques de rendre un énoncé informatif est donc la reprise anaphorique.

7.2 Information nouvelle ou donnée en génération

La thématique de l'apport d'information dans le cas d'expressions coréférentielles a été développée essentiellement par [Danlos, 1999] et [Danlos et Gaiffe, 2000] à propos de la coréférence événementielle. Pour Danlos, deux descriptions sont en relation de particularisation si la seconde apporte de l'information sur la première, et en relation de

généralisation si la seconde n'apporte rien de nouveau sur la première. Sur la coréférence entre groupe nominaux, la particularisation correspond à la reprise par un hyponyme de la tête nominale de l'antécédent, et la généralisation correspond à une reprise par hyperonymie. Nous ne réutiliserons pas la terminologie de Danlos dans la mesure où ce qu'elle appelle généralisation ne correspond pas exactement à ce que nous appellerons plus loin « répétition d'information ». En effet, dans les exemples illustrant le discours généralisant, non seulement les reprises n'apportent pas d'information sur l'antécédent, mais elles en donnent moins. Nous incluons dans « répétition d'information » tous les cas où la somme d'informations contenue dans la reprise est égale ou inférieure à celle contenue dans l'antécédent, c'est à dire que nous y incluons aussi les cas de reprise directe. Par ailleurs, cette terminologie est très marquée pour la coréférence événementielle et nous faisons le choix d'une terminologie plus spécifique à notre problème, destinée uniquement aux reprises coréférentielles nominales. Dans cette section, nous définirons l'apport d'information et la répétition d'information, et nous exposerons les problèmes que posent ces notions en génération d'expressions référentielles.

7.2.1 Apport d'information

On peut se demander s'il est vraiment nécessaire de générer les reprises coréférentielles apportant de l'information nouvelle. Notre réponse tient en deux arguments que nous développons ici.

Si l'on souhaite générer un texte dont le but communicatif est de transmettre les informations suivantes :

{*Mort(François Mitterrand), Président(François Mitterrand), 85ans(François Mitterrand)*}

on veut pouvoir générer le texte 103, qui utilise la reprise anaphorique pour apporter une information nouvelle à propos du référent. Le texte 103 est beaucoup plus concis et naturel que le texte 104, qui est le texte qui sera le plus facilement généré à partir des buts communicatifs énoncés.

(103) *François Mitterrand est mort. Le président de la République avait 85 ans.*

(104) *François Mitterrand est mort. François Mitterrand était président de la République française. François Mitterrand avait 85 ans.*

Economie de moyens L'avantage d'utiliser l'anaphore comme support à d'information nouvelle est le suivant : cela évite de produire un texte long peu naturel. Apporter de l'information nouvelle sur un référent est un moyen de faire des textes plus synthétiques. Par ailleurs, cette utilisation des groupes nominaux donne une solution alternative à l'apposition qui peut parfois être lourde et qu'on ne sait pas générer. Bien entendu, pour pouvoir permettre au générateur d'apporter l'information « Président(François Mitterrand) » dans le groupe nominal, nous avons besoin d'un module de planification du texte qui autorise ce type de tournure. Ceci n'entrant pas directement dans le sujet de notre thèse, nous supposons que le planificateur de document utilisé le permet.

Le problème de l'interlocuteur L'utilisation de l'anaphore comme support à de l'information nouvelle permet aussi une conception plus souple des connaissances de l'interlocuteur. On ne préjuge plus complètement de ses connaissances, puisque deux cas de figure sont possibles : soit il sait qui est François Mitterrand, auquel cas il résoud l'anaphore grâce à ses connaissances du monde, soit il ne sait pas, et accommode l'anaphore grâce à ses connaissances linguistiques sur l'utilisation du déterminant défini (et apprend une information sur le référent).

7.2.2 Répétition d'information

Comme l'illustre l'exemple 99, les reprises coréférentielles qui répètent de l'information n'utilisent pas toujours de l'information donnée explicitement dans le contexte ou dans l'antécédent. Ce type de reprise est donc basé sur des inférences dont on doit identifier les sources.

En effet, si l'on veut générer un texte comme le texte 99, vu dans cette section, où *Paul* est repris par *la victime*, il est nécessaire de savoir d'où proviennent les inférences et comment elles sont réalisées. Ici, les inférences proviennent du cotexte, et plus précisément du verbe de la première phrase.

7.2.3 Questions posées par la gestion de l'information nouvelle et de l'information donnée dans la génération des reprises coréférentielles

Notre problème serait donc de savoir comment les deux types de descriptions peuvent être produits, et particulièrement les descriptions ajoutant de l'information. Rappelons que les algorithmes existants ne génèrent que des reprises contenant de l'information donnée explicitement dans le discours antérieur [Dale et Reiter, 1995]. Nous orientons alors notre recherche dans deux directions : identifier les moyens linguistiques utilisés pour réaliser les deux types de reprises et identifier la source des inférences utilisables dans la production de ces reprises.

7.2.3.1 Les moyens linguistiques mis en œuvre pour apporter (ou non) la nouvelle information

Pour pouvoir générer les deux types d'anaphore identifiés ici, il est indispensable de connaître les moyens linguistiques utilisés pour produire ces anaphores (relations lexicales, déterminants, modifieurs...). Notre étude de corpus vise donc à identifier clairement tous les moyens à disposition dans la langue pour produire des descriptions répétant l'information donnée (désormais DRID) ou des descriptions ajoutant de l'information nouvelle (désormais DAIN). En effet, nous pouvons supposer que les deux types d'anaphore se réalisent différemment du point de vue linguistique. Par exemple, on peut imaginer qu'un moyen d'ajouter de l'information sur un référent est l'ajout de modifieurs, mais quels types de modifieurs sont utilisés ? Si l'ajout d'information se fait grâce à un nom, quel type de relation entretient le nom de la reprise avec le nom de l'antécédent ?

7.2.3.2 Localiser les inférences nécessaires à la construction d'anaphores

Les reprises indirectes font intervenir des inférences provenant de sources diverses (connaissances lexicales, connaissances du monde...). Pour produire des reprises indirectes, il faut donc identifier les sources possibles de l'inférence. Certaines sont faciles à identifier, à stocker et à représenter (les relations lexicales par exemple), tandis que d'autres le sont moins (relations encyclopédiques).

Les moyens d'étendre l'algorithme de Gardent et Striegnitz seront donc les suivants :
- localiser les sources de l'inférence, car les bases de connaissances utilisées dans l'algorithme sont structurées, mais il nous faut savoir où chercher les informations,
- trouver les moyens linguistiques utilisés et éventuellement les corréler avec les sources de l'inférence,
- enfin, introduire la distinction entre le défini et le démonstratif dans l'algorithme (par exemple en trouvant une contrainte nouvelle à opposer aux contraintes d'unicité et de familiarité).

7.3 Mise au point d'une classification tenant compte des données décrites précédemment

La première étude de corpus que nous avons présentée (chapitre 5) nous a permis de trier les données et de pouvoir les étudier de façon ordonnée. Nous y avions recensé la totalité des groupes nominaux définis et démonstratifs coréférentiels, et identifié la relation sémantique qu'ils entretenaient avec leur antécédent. Nous avons débouché sur la classification suivante, tenant compte des paramètres cités dans les sections précédentes de ce chapitre.

7.3.1 Descriptions répétant de l'information donnée (DRID)

Dans cette section, nous traitons des reprises sans ajout d'information, c'est-à-dire des groupes nominaux coréférentiels contenant autant ou moins d'informations que leur antécédent. Ce type de reprise sera désormais abrégé DRID (description répétant de l'information donnée). Nous montrons dans cette classification que l'information répétée peut provenir de différentes sources. L'information peut bien entendu provenir de l'antécédent où elle est explicitement mentionnée, ou bien d'inférences construites à partir de plusieurs bases de connaissances : des ressources linguistiques dans le cas de l'utilisation d'une relation lexicale, du modèle de discours quand l'information provient du cotexte, ou des connaissances encyclopédiques. Les exemples que nous donnons dans cette section proviennent tous du corpus PAROLE.

7.3.1.1 L'information utilisée dans la reprise provient de l'antécédent

Dans un premier temps, nous avons regroupé dans cette catégorie toutes les reprises coréférentielles réutilisant uniquement des informations explicitement données dans le syn-

tagme nominal antécédent. Dans les deux exemples suivants, la reprise est une reprise directe dans laquelle les modifieurs ne sont pas répétés ; elle ne fait donc que répéter une partie de l'information donnée antérieurement.

(c-16) *Celle-ci, (...) aurait en effet tissé un réseau de* ***liens ambigus dans la gendarmerie, la sûreté de l'Etat, les clubs de tir****. Le procès, au printemps dernier de deux membres d'une organisation néo-nazie, (...), avait permis de mettre* ***ces liens*** *en relief.*

(c-17) *Le gaz propulseur est, en effet, fait de chlorofluorocarbones dont on pense qu'ils détruisent* ***l'ozone de la haute atmosphère****.* ***L'ozone*** *protège la terre du rayonnement ultraviolet du soleil.*

7.3.1.2 L'information est inférée à partir de l'antécédent et du cotexte :

Cette catégorie regroupe les reprises dont certains éléments sont explicitement donnés dans l'antécédent, alors que d'autres doivent être inférés à partir du cotexte.

Dans l'exemple suivant, la reprise est directe et le nom tête est *comportement* dans les deux syntagmes. L'adjectif *nouveau*, en revanche, est inféré à partir du verbe *modifier*. On peut en effet dire que *modifier un comportement* revient à *adopter un nouveau comportement*.

(c-18) *Lors des périodes ayant précédé les trois dernières grandes échéances électorales, le patronat avait très sensiblement modifié* ***son comportement****. (...) La clé de* ***ce nouveau comportement*** *tient en deux chiffres : 79 % des patrons interrogés déclarent qu'ils sont satisfaits de la politique actuelle menée par Jacques Chirac.*

Dans l'exemple suivant, le syntagme *le gouvernement Aquino* mentionné à la fin du texte n'apporte pas de nouvelle information sur le référent, dans la mesure où même si la mention antérieure du gouvernement en question est simplement le syntagme *le gouvernement*, tous les éléments mis en gras dans le texte antérieur permettent de comprendre qu'il est question du gouvernement de la présidente Aquino.

(c-19) *Huit jours après le putsch du 28 août qui faillit bel et bien renverser* ***le gouvernement Aquino*** *(...) le colonel Honasan, chef de la rébellion, a déclaré que son groupe n'avait nullement l'intention de faire du mal à* ***la présidente*** *et à sa famille (...). Mais il a blâmé* ***Mme Aquino*** *pour avoir renoncé à l'idéal de la révolution de février 1986 (...) Alors que* ***le gouvernement*** *cherche à le présenter comme un « traître », « Gringo » Honasan affirme au contraire que ses hommes et lui-même incarnent les idéaux de février 1986. Par ce message, revendiquant une sorte de mission historique, il ne semble en rien prêt à capituler. (...) La société Sigma qui emploie un millier de « gardes » est connue pour appartenir en sous-main à M. Ponce Enrile, ancien ministre de la défense et « bête noire » du* ***gouvernement Aquino****.*

7.3.1.3 L'information est inférée à partir de l'antécédent grâce au savoir lexical :

Cette catégorie de reprises contient en fait toutes les reprises s'appuyant sur la relation lexicale entretenue par le nom de l'antécédent et celui de la reprise (hyperonymie et synonymie essentiellement). Dans l'exemple c-20, la reprise du terme *sécheresse* se fait par un hyperonyme, *phénomène*, qui par définition n'apporte aucune information, *la sécheresse* étant un type de *phénomène*. De la même façon, dans l'exemple c-21, le nom *crime* n'apporte pas d'information sur le nom *assassinat*, puisqu'il s'agit d'un type de *crime*.

(c-20) *D'année en année, l'Inde paie un tribut sans cesse plus lourd à* ***la sécheresse****, notamment en raison de l'aggravation de la déforestation, qui a détruit l'équilibre écologique.* ***Ce phénomène*** *a été accentué par des choix économiques erronés.*

(c-21) *Ce témoin affirme en effet qu'il aurait assisté, ce soir-là, à Nivelles, à* ***l'assassinat d'un couple venu en voiture faire le plein d'essence****. Et il aurait reconnu, parmi les auteurs* ***du crime****, Jean Bultot, trente-six ans, (...).*

7.3.1.4 L'information est inférée à partir de l'antécédent grâce au savoir lexical et à partir du cotexte :

Cette catégorie combine les éléments mis en avant dans les deux catégories précédemment citées. Ici, le lien avec l'antécédent est réalisé par une relation lexicale n'apportant pas d'information nouvelle, et toutes les informations contenues dans le groupe nominal sont inférables à partir de cette relation et d'autres éléments du cotexte. Dans l'exemple suivant, *bâtiment* est un hyperonyme de *Palais des concerts*, l'adjectif *confortable* est inféré de *somptueux*, et l'expression *flambant neuf* est inférée de *s'est dotée récemment*.

(c-22) *La municipalité s'est dotée récemment d'****un somptueux Palais des concerts****. C'est dans* ***ce bâtiment confortable et flambant neuf*** *qu'a eu lieu l'inauguration.*

Dans l'exemple suivant les *Etats victimes de la sécheresse* sont mentionnés à nouveau par le terme *zone*, qui est un hyperonyme du nom *Etat*. L'adjectif qualificatif *sinistrées* est inférable à partir de l'expression *victime de la sécheresse*.

(c-23) *En tout,* ***vingt et un Etats de l'Union*** *sur vingt-cinq sont victimes en tout ou partie de la sécheresse. (...) Un comité de crise a été constitué, présidé par le premier ministre, qui, depuis une semaine, multiplie les voyages dans* ***les zones sinistrées****.*

7.3.1.5 L'information est inférée grâce aux connaissances encyclopédiques à partir de l'antécédent et du contexte :

Dans cette catégorie d'anaphores, la reprise n'apporte pas d'information nouvelle, ne contient pas d'élément explicitement mentionné dans l'antécédent et le lien avec l'antécédent n'est pas inférable directement à partir de connaissances lexicales. On considère donc que les inférences qui permettent de lier les deux groupes nominaux proviennent des connaissances encyclopédiques.

Dans l'exemple suivant, on peut supposer que pour un interlocuteur quelconque, il est possible de comprendre que lors d'un voyage officiel, la visite d'un chef d'Etat au cimetière où sont enterrés ses parents est considérée comme une *partie privée de son voyage*.

(c-24) *« Les journalistes est-allemands ne feront pas de reportage sur* ***la visite de M. Honecker au cimetière de Neunkirchen, dans la Sarre, où sont enterrés ses parents****. Ainsi en a-t-il décidé, explique Otto Schwabe, rédacteur en chef de la revue Horizont, après que le chef d'Etat lui-même eut requis la « tranquillité » pour* ***cette partie « privée » de son voyage en République fédérale****.*

Dans l'exemple suivant, nos connaissances du monde nous permettent d'interpréter la reprise dans la mesure où l'on sait que les *camarades* de soldats sont aussi des *soldats* et que l'adjectif *rebelle* peut convenir à des soldats *qui se sont mutinés*.

(c-25) *Le commandement militaire est toujours dans l'incertitude : il ne sait pas, huit jours après le putsch, sur quelle unité il peut vraiment compter et jusqu'où ira la loyauté des soldats appelés éventuellement à combattre* ***leurs camarades qui se sont mutinés****. S'il y a une division au sein des militaires, ce n'est pas sur le contenu des demandes* ***des soldats rebelles*** *(...) mais sur les méthodes d'action pour atteindre ces objectifs.*

Dans l'exemple suivant, le lien avec le nom propre a été considéré comme provenant des connaissances du monde. Cet exemple est représentatif des cas où les connaissances de l'interlocuteur sont difficiles à prévoir et à représenter avec certitude. En effet, il n'est pas évident que l'interlocuteur sache qui est M. Perez De Cuellar. Pour la personne ayant annoté le corpus, le fait qu'il soit secrétaire général de l'ONU n'était pas une donnée nouvelle sur le référent, elle a donc considéré que le lien se faisait grâce à ses connaissances.

(c-26) *En revanche, interrogé sur le sens de son séjour dans la région du Golfe - probablement du 13 au 17 ou 18 septembre, -* ***M. Perez de Cuellar*** *a fait valoir qu'il n'y avait pas besoin de répondre à la résolution, puisque celle-ci était « obligatoire ». Certains diplomates à l'ONU se montrent d'ailleurs très sceptiques quant aux chances de succès du* ***secrétaire général****.*

7.3.2 Descriptions apportant de l'information nouvelle (DAIN)

Dans cette section, nous montrons qu'il est possible de faire des reprises en ajoutant de l'information sur le référent. Cette information étant nouvelle, elle provient du modèle du locuteur qui veut la communiquer et n'est pas connue de l'interlocuteur. Le problème est donc de savoir par quels moyens linguistiques on peut la réaliser. Nous avons quatre catégories ici : l'ajout d'information grâce à une relation lexicale, grâce à des modifieurs, une catégorie combinant les deux, et enfin, l'apport d'information par un groupe nominal sans relation avec l'antécédent (reclassification).

7.3.2.1 Relation lexicale spécifiante :

Dans les cas regroupés dans cette catégorie, le lien entre l'antécédent et l'anaphore est inféré d'une relation lexicale d'hyponymie, qui permet d'apporter une nouvelle information sur le référent ou plus vraisemblablement de préciser la nature d'un référent. Ainsi, dans

les exemples suivants, on peut dire qu'un *rapport* est un type spécifique de *document* et que *la mousson* est un type de *pluies torrentielles*.

(c-27) ***Ce document*** *souligne la gravité croissante des conséquences médicales de la consommation de tabac, responsable en France de plus de 10% des décès. Les auteurs de* ***ce rapport*** *formulent une série de propositions à bien des égards très dérangeantes.*

(c-28) *La nuit précédente,* ***des pluies torrentielles*** *s'étaient abattues sur la capitale, dont les rues étaient transformées en torrents boueux. De l'eau jusqu'aux genoux, les habitants sont sortis pour manifester leur joie.* ***La mousson****, enfin ...*

7.3.2.2 Relation lexicale spécifiante et modifieurs :

Cette catégorie contient les cas de reprises où la tête de la reprise est un hyponyme de l'antécédent auquel est adjoint une série de modifieurs. Dans l'exemple suivant, le syntagme *le personnel* est repris pas l'hyponyme *ouvrières* qui spécifie le travail du personnel et son sexe, et auquel on a adjoint une série de modifieurs décrivant des informations nouvelles et non inférables du reste du contexte.

(c-29) *Mais à Roubaix (...),* ***le personnel*** *a l'impression de seulement compter les points. La Lainière va peut-être supprimer des cars de ramassage ! Pour* ***ces ouvrières du bassin houiller dont quelques-unes ont déjà trois heures de transport par jour****, la nouvelle (...) a relégué au second plan les manœuvres boursières dont leur entreprise fait l'objet depuis deux mois.*

De même, dans l'exemple suivant, on considère une *manifestation* comme un type particulier de *lutte*, et les modifieurs nous apportent de l'information sur l'origine des manifestations, et leur date.

(c-30) *Pour cette journée « de fête », une exposition de photos retrace* ***les luttes des dernières années****. Sept personnes ont été tuées à la Victoria, depuis le début* ***des premières manifestations de l'opposition en 1983****.*

7.3.2.3 L'information nouvelle est dans les modifieurs :

Dans ce type de reprise, l'information non inférable provient des modifieurs. Deux types d'exemples se retrouvent dans cette catégorie. Dans le premier exemple cité ici, la reprise est directe (les deux syntagmes ont le même nom tête), seuls les modifieurs varient et ceux utilisés dans la reprise décrivent une situation non inférable. Dans le deuxième exemple, la relation entre les deux groupes nominaux provient de la relation lexicale entre le syntagme nominal apposé au nom propre *Ficofrance* et le syntagme anaphorique, la reprise contient des informations nouvelles.

(c-31) *L'aviation israélienne a effectué le samedi 5 septembre un raid sur* ***le camp de réfugiés palestiniens d'Ain-el-Heloue, dans les faubourgs de Saida, chef-lieu du Liban-sud****, ont rapporté les correspondants sur place. Les chasseurs-bombardiers israéliens ont effectué à partir de 10h15 locales plusieurs attaques en piqué sur* ***ce camp qui compte soixante-mille habitants****, (...).*

(c-32) *Parallèlement, il prendrait la présidence de* ***Ficofrance, la société financière de GMF***. ***Ce groupe familial, qui a réalisé en 1986 un chiffre d'affaires de 5 milliards de francs****, figure parmi les candidats les plus sérieux (...).*

L'exemple suivant est différent dans le sens où le nom propre est repris par le nom commun *la société*, ce qui était inférable à partir du texte antérieur à cet exemple, comme l'adjectif *suisse* est inférable à partir du nom de la société qui contient le nom propre *Genève*. En revanche, les informations contenues dans la proposition subordonnée relative ne sont pas inférables.

(c-33) *Sur ce montant,* ***Chaumet-Genève*** *estime que 115 millions sont dus par la société mère de Paris, 43 millions ne souffrent pas de contestation et 93 millions font l'objet de vérifications. Investment International, qui s'est porté acquéreur de la maison mère, offre de reprendre aussi* ***la société suisse dont la valeur de liquidation est estimée à 10 millions de francs suisses****.*

7.3.2.4 L'information nouvelle est dans tout le syntagme et ne passe pas par une relation lexicale :

Dans ce type de reprise, on ne peut établir de lien lexical entre l'antécédent et l'anaphore. La reprise entière porte des informations nouvelles, ou des jugements. On retrouve plusieurs phénomènes dans cette catégorie.

C'est dans cette catégorie que nous plaçons les métaphores, bien qu'il ne soit pas certain qu'elles apportent des informations nouvelles, elles ont un caractère subjectif qui peut parfois permettre au locuteur d'introduire un jugement :

(c-34) *Les huit journées de compétition ont été dominées par l'Allemagne de l'Est, qui, à l'heure du bilan, totalise* ***31 médailles dont 10 d'or****. Une large partie de* ***cette moisson*** *a été récoltée par les athlètes féminines (...).*

(c-35) *Il est probable que* ***les péronistes, invités avant les élections par le président Alfonsin à « participer au gouvernement »****, auront d'autres exigences. (...) Les résultats très surprenants de ces élections, loin de clarifier la situation, risquent de remettre en cause la longue marche de l'Argentine vers la démocratie parlementaire : les militaires relèvent la tête et* ***la nébuleuse péroniste****, porteuse de tous les dérapages et de toutes les fuites en avant, contamine de nouveau la société.*

L'information nouvelle peut être l'introduction du point de vue du locuteur :

(c-36) *A un moment, (...) je tombe sur un article intitulé « Pourquoi* ***les maris*** *prennent le large ». Je me dis : cherche pas, ils se débinent pendant que tu t'échines à faire des pompes et des flexions,* ***ces salauds-là****.*

(c-37) ***Le mec*** *roule des yeux. Beaucoup plus loin, à distance respectable* ***du cerbère****, sous la protection rapprochée d'un flamboyant, je griffonne les mots que je craignais d'oublier.*

L'information nouvelle peut consister en un changement de point de vue sur le référent :

(c-38) *Et si* ***Carl Lewis*** *était condamné à se battre sans cesse contre les chimères du sport moderne ?* ***Ce petit garçon qui avait une mauvaise croissance*** *est devenu adulte, un athlète prodigieusement doué.*

L'information nouvelle peut être une propriété objective du référent :

(c-39) ***Richard Vivien*** *a créé une surprise de taille en devenant samedi champion du monde de la catégorie, que l'on définissait il y a peu de temps comme étant celle des « purs ».* ***Ce Normand de vingt-trois ans, remarqué par Yves Hezard****, est peut-être un pur, mais il possède déjà beaucoup de métier (...).*

(c-40) *Est-ce la seule raison qui a conduit* ***Kodak*** *à se lancer dans l'arène, avec toute la puissance d'un groupe de 11 milliards de dollars de chiffre d'affaires mondial ? Il semble bien que pour* ***la firme de Rochester*** *ce lancement très coûteux (d'ici à la fin de l'année, elle dépensera en publicité autant que tous ses concurrents réunis) se situe avant tout dans le cadre de sa stratégie de diversification tous azimuts.*

7.4 Etude de corpus

Nous avons réannoté le corpus PAROLE selon la classification présentée en section 7.3. Le but de notre annotation est double :

- Dans un premier temps, nous souhaitons déterminer la fonction des différentes reprises et trouver la provenance des inférences impliquées dans les reprises définies et démonstratives.
- Dans un deuxième temps, notre objectif est d'identifier les contraintes qui permettent le choix entre le défini ou le démonstratif.

Pour plus de clarté, nous présentons les résultats de notre annotation de corpus dans deux chapitres séparés :

Dans ce chapitre, nous présentons les résultats qui nous permettront d'affiner l'algorithme de Gardent et Striegnitz sur la détermination du contenu de la reprise. Les résultats concernant les définis et les démonstratifs sont toujours présentés séparément pour une raison technique, qui nous a fait séparer l'annotation des groupes nominaux définis et celle des groupes nominaux démonstratifs dans des fichiers séparés. Cependant, nous ferons l'étude globale de ces résultats pour la raison suivante : l'emploi du déterminant est conditionné par le contenu sémantique de la description, par sa forme syntaxique et par le contexte, et non l'inverse. Aussi, il nous semble normal de faire d'abord l'étude du contenu sémantique des descriptions de façon globale, et d'établir ensuite des contraintes sur l'utilisation des déterminants en fonction du contenu sémantique et de la forme syntaxique de la description. Nous présenterons les résultats concernant la disctinction entre les définis et les démonstratifs dans le chapitre suivant.

7.4.1 Résultats généraux

Le tableau 7.1 présente les éléments suivants : dans le premier tableau, on peut lire la proportion de groupes nominaux retenus pour l'étude. On travaille essentiellement sur les descriptions coréférant à des syntagmes nominaux (sauf cas exceptionnels - cf Manuel

cat	Total Coref.	Ant =Nom	prop	Autres	prop
Définis	1505	1402	94%	103	6%
Démonstratifs	442	369	83,50%	73	16,51%
Total	1947	1771	90,96%	176	9,04%

cat	Total A = Nom	DRID	prop	DAIN	prop
Définis	1402	1314	93,72%	89	6,28%
Démonstratifs	369	299	81,03%	70	18,97%
Total	1771	1613	91,1%	159	8,9%

FIG. 7.1 – Résultats généraux de la deuxième annotation

d'annotation, Annexe B), ce qui correspond à la colonne « Ant = Nom » et on exclut les cas d'apposition ou de constructions attributives. En effet, nous considérons que la coréférence est marquée explicitement dans les appositions par la virgule, et dans les constructions attributives par la copule. Nous faisons donc l'hypothèse que la génération ou la résolution de la coréférence dans ces cas ne se font pas de la même façon. On travaille alors sur 94% des définis coréférentiels et 83% des démonstratifs coréférentiels, soit sur environ 91% des emplois coréférentiels de ces descriptions dans notre corpus. Les colonnes « Ant = Nom » et « Autres » présentent le nombre de cas inclus dans l'étude en valeur absolue, et les colonnes « prop » la proportion de cas retenus, la colonne « Total Coref » indiquant le nombre total de descriptions coréférentielles trouvées dans les deux corpus.

On constate dans ce tableau que les deux fonctions des reprises coréférentielles (ajouter de l'information et répéter de l'information) sont représentées dans notre corpus. Cependant, on note une grande disproportion dans les emplois : les descriptions répétant de l'information donnée (DRID) sont très majoritaires (91% des cas) par rapport à celles qui ajoutent de l'information nouvelle (DAIN), qui représentent 9% des cas.

7.4.2 Descriptions Répétant de l'Information Donnée

Le tableau 7.2 présente les résultats de la façon suivante : on trouve en abrégé le nom de chaque catégorie de DRID dans la première colonne. Pour chaque type de déterminant, on trouve les trois informations suivantes : le nombre total de cas trouvés, la proportion que ces cas représentent par rapport au nombre total de descriptions trouvées dans le corpus, et la proportion que ces cas représentent dans la catégorie des DRID. La dernière colonne donne ces informations pour la totalité des reprises trouvées dans le corpus.

Notons une fois encore que l'apport d'information est un jugement variable en fonction des interlocuteurs. Nous reprenons l'exemple cité dans la section 7.3, où nous ne pouvons pas être sûre que le fait que M. Perez De Cuellar soit le secrétaire général de l'ONU fasse partie des connaissances de n'importe quel interlocuteur.

(c-41) *En revanche, interrogé sur le sens de son séjour dans la région du Golfe - probablement du 13 au 17 ou 18 septembre, -* ***M. Perez de Cuellar*** *a fait valoir qu'il n'y avait pas besoin de répondre à la résolution, puisque celle-ci était « obligatoire ». Cer-*

	Définis			Démonstratifs			TOTAL		
Antéc. :	642	45,79	48,71	92	24,93	30,77	734	41,44	45,5
Ant + Ctxt :	131	9,34	9,96	24	6,50	8,03	155	8,75	9,61
Rel. Lex :	226	16,12	17,15	71	19,24	23,75	297	16,77	18,41
Rel. Lex + Ctxt :	57	4,06	4,34	25	6,78	8,36	82	4,63	5,08
Connaiss. Encycl. :	258	18,40	19,58	87	23,58	29,10	345	19,48	21,39
Total DRID :	1314	-	100	299	-	100	1613	-	100
Total coref. :	1402	100	-	369	100	-	1771	100	-

FIG. 7.2 – Résultats DRID

tains diplomates à l'ONU se montrent d'ailleurs très sceptiques quant aux chances de succès du ***secrétaire général****.*

Nous sommes donc prudente vis-à-vis de la catégorie DRID quand l'information donnée provient des connaissances encyclopédiques, qui pourrait se retrouver presque intégralement dans la catégorie d'ajout d'information par un syntagme sans relation lexicale avec un autre annotateur. En effet, surtout quand l'antécédent est un nom propre, il est difficile de préjuger des connaissances de l'interlocuteur. Ceci transfèrerait près de 20% des cas d'utilisations coréférentielles du défini et du démonstratif en DRID dans la catégorie des DAIN.

7.4.2.1 Répétition d'informations provenant explicitement de l'antécédent

Antécédent seul La source principale de l'information répétée est l'antécédent dans les deux cas. Ceci est valable pour 41% des reprises coréférentielles.

Antécédent et contexte Parfois le contexte est aussi source d'information en plus de l'antécédent, ce qui porte à près de 50% les reprises coréférentielles utilisant l'information contenue dans l'antécédent.

7.4.2.2 Répétition d'informations inférées

Information inférée des connaissances lexicales L'utilisation d'une relation lexicale pour répéter de l'information est relativement importante dans les reprises sans ajout d'information (environ 18%). Il est intéressant d'observer la distribution des relations lexicales impliquées dans le tableau 7.3. Nous constatons que de manière générale, l'hyperonymie et la synonymie sont les deux relations les plus employées. Nous reportons le lecteur au chapitre suivant pour un commentaire plus détaillé sur la comparaison entre le défini et le démonstratif.

Rôle du cotexte De manière générale, le cotexte est utilisé dans la production des anaphores dans 12% des cas, ce qui est relativement peu. Il est par ailleurs toujours utilisé en même temps qu'une autre source d'inférence.

	Démonstratif		Défini	
	Relation lex.	Rel. Lex et Ctxt	Relation lex.	Rel. Lex et Ctxt
hyperonymes	52	3	106	6
synonymes	19	4	118	4
thématiques	0	11	0	24
hyponymes	0	7	2	23

FIG. 7.3 – Relations lexicales impliquées dans la répétion d'information

Information inférée des connaissances du monde Les connaissances encyclopédiques sont utilisées comme source d'inférence dans environ 20% des cas. Ceci peut être vu comme un véritable problème, dans la mesure où les connaissances encyclopédiques sont les connaissances les plus difficile à modéliser. Cependant, si on sépare l'étude des reprises ayant pour antécédent un nom propre (tableau 7.4), de celles ayant un nom commun pour antécédent, pour les raisons énoncées au début de cette section, on constate que le problème est moins important. En effet, seules 10,7% des reprises de nom communs impliquent des connaissances du monde. On peut alors imaginer que les entités habituellement désignées par un nom propre sont plus faciles à représenter dans les connaissances du monde, les caractéristiques par lesquelles on peut les désigner pouvant être facilement listées (i.e. profession, genre, nationalité...).

	Défini		Démonstratif		Total	
	Nb.	Prop	Nb.	Prop.	Nb.	Prop.
Nom commun	105	8%	67	22,4%	172	10,67
Nom propre	153	11,6%	20	6,69%	173	10,72
Total DRID - C. Encycl.	258	19,6%	87	29,1%	345	21,39
Total DRID	1314	-	299	-	1613	-

FIG. 7.4 – Antécédents nominaux des reprises inférées des connaissances du monde

7.4.3 Descriptions Ajoutant de l'Information nouvelle (DAIN)

Le tableau 7.5 présente les résultats de l'annotation de corpus sur les descriptions ajoutant de l'information. On retrouve chacune des catégories décrites précédemment (Rel. Lex pour l'ajout d'information par une relation lexicale, l'ajout d'information par les modifieurs, l'ajout d'information par les modifieurs et une relation lexicale, et l'ajout d'information par un groupe nominal). Trois grandes colonnes présentent séparément les résultats pour les définis, les démonstratifs et leur total, chacune d'elle est séparée en trois, présentant tout d'abord le nombre d'occurrences trouvées pour chaque phénomène, la proportion qu'il représente parmi les syntagmes coréférentiels, et la proportion qu'il représente parmi les descriptions ajoutant de l'information nouvelle sur le référent.

	Définis			Démonstratifs			Total		
Rel. Lex :	7	0,50	7,87	2	0,54	2,86	9	0,51	5,7
modifieurs :	22	1,57	24,72	19	5,15	27,14	41	2,31	25,95
Rel. Lex. + mod :	2	0,14	2,25	3	0,81	4,29	5	0,28	3,16
Syntagme nominal :	52	3,71	58,43	46	12,47	65,71	98	5,53	62,03
Total DAIN :	88	-	100	70	-	100	158	-	100
total coref :	1402	100	-	369	100	-	1771	100	-

FIG. 7.5 – Résultats DAIN

7.4.3.1 Utilisation de modifieurs

L'utilisation des modifieurs est un moyen très courant pour ajouter l'information. Les reprises ajoutant de l'information au moyen des modifieurs représentent en effet environ un quart des DAIN. C'est effectivement un moyen qui semble commode, dans la mesure où le locuteur peut utiliser un nom permettant facilement de faire le lien avec l'antécédent, et ajouter de l'information en adjoignant des modifieurs à ce nom.

7.4.3.2 Utilisation d'hyponymes (avec ou sans modifieurs)

Ce moyen est globalement peu utilisé, il représente à peine 6% des cas de DAIN sans modifieurs, et en valeur absolue, il représente 9 cas. De la même façon, notre corpus ne compte que 5 cas de DAIN par hyponymie et modifieurs, ce qui nous conforte dans l'idée qu'il s'agit de cas marginaux.

7.4.3.3 Utilisation d'un syntagme sans lien avec l'antécédent et le cotexte

L'utilisation d'un syntagme sans lien avec l'antécédent ou le contexte pour ajouter de l'information est de très loin le moyen le plus utilisé. Il représente plus de 60% des cas de DAIN. Il semble donc que la plupart du temps, l'information nouvelle est impossible à utiliser avec une reprise directe, qui demanderait la construction de moins d'inférences.

7.4.4 Conclusions

Les résultats de notre étude de corpus montrent les éléments suivants :

Tout d'abord, les reprises coréférentielles peuvent être utilisées pour apporter de l'information nouvelle sur le référent. Ceci doit pouvoir être intégré dans un algorithme de détermination du contenu des expressions coréférentielles.

Ensuite, lorsque la reprise répète de l'information donnée, les sources de l'inférence permettant de faire le lien avec l'antécédent sont au nombre de cinq, et sont utilisées sans l'ordre de préférence suivant (cf. tableau 7.2) :

1. Antécédent,

2. Connaissances encyclopédiques,
3. Connaissances lexicales,
4. Antécédent et modèle de discours,
5. Connaissances lexicales et modèle de discours.

Ces résultats pourront être pris en compte dans un algorithme de détermination du contenu des expressions coréférentielles.

Enfin, nous avons des indications sur la réalisation de l'information nouvelle dans les DAIN. Elle sera réalisée préférentiellement par un syntagme nominal complet sans lien lexical avec l'antécédent. Les modifieurs sont eux aussi très utilisés. Ces résultats portant sur la réalisation de l'information nouvelle, ils ne pourront être intégrés à un algorithme déterminant le contenu des descriptions coréférentielles, mais pourront peut être, à plus long terme, être utilisés dans le module de réalisation d'un générateur et nous seront utiles lors du choix du déterminant (chapitre 8).

7.5 Une extension de l'algorithme de génération des descriptions définies

L'algorithme de Gardent et Striegnitz permet la génération de descriptions définies en construisant des inférences sur le contexte et sur les connaissances du monde. Notre étude a montré différents éléments qui ne sont pas pris en compte par cet algorithme :

- Les sources d'inférence pour la construction des descriptions coréférentielles et des anaphores associatives sont variées.
- Les descriptions coréférentielles peuvent apporter de l'information nouvelle sur le référent.

Pour rappel, nous redonnons l'algorithme de Gardent et Striegnitz en figure 7.6.

Notre apport à cet algorithme est le suivant : nous avons pu définir des préférences sur les diverses sources d'inférence utilisées dans les description définies et démonstratives n'ajoutant pas d'information. Nous avons montré que ces sources d'inférences ont le même ordre de préférence quel que soit le déterminant utilisé (cf. tableaux 7.2 et 7.5). Il s'agit donc probablement d'un ordre de préférence pour la construction du contenu sémantique du syntagme, sans lien particulier avec le déterminant. Nous choisissons par conséquent d'établir un ordre fixe dans la recherche d'une source d'inférence pour la production de la description coréférentielle basée sur la fréquence d'apparition des diverses sources possibles dans notre corpus.

Cet ordre de préférence est le suivant pour les DRID : la source de l'inférence est (1) l'antécédent lui même, (2) la base des connaissances encyclopédiques, (3) la base des connaissances lexicales, (4) l'union des informations contenues dans l'antécédent et dans le modèle de discours en général, (5) l'union des connaissances lexicales et du modèle de discours.

Ensuite, nous pouvons introduire dans l'algorithme la possibilité d'introduire de l'information nouvelle dans l'expression coréférentielle. Nous avons vu précédemment que

Entrée :
WKL (connaissances du monde) : ensemble de règles reliant les entités les unes aux autres
DM (modèle de discours) : ensemble de formules atomiques
SM (modèle du locuteur) : ensemble de formules atomiques
t : entité cible, t appartient aux termes de SM et de DM.

Initialisation :
1. buts ← pile avec l'élément t
2 N ← structure syntaxique initiale avec une place vide pour un nom

Check success :
3. Si buts est vide, alors retourner <uniquely identifying, N>
4. but-courant ← but en sommet de pile
5. Si IA(but-courant) $\not\subseteq$ PA(but-courant, L(N)), alors retourner <unfamiliar, N>
6. Si PA(but-courant, L(N)) = IA(but-courant) et $\forall$ a $\in$IA(t) : t est unique selon a étant donné L(N) alors empiler but en sommet de pile ; aller en 4.

Etendre la description :
7. Si but-courant $\in$ terms(DM) alors Ctxt ← DM sinon Ctxt ← DM + SM
8. Essayer de sélectionner une formule atomique p applicable tel que Ctxt+WKL $\models$ p
9. S'il n'existe pas de tel p alors retourner <non identifying, N>
10 pour chaque o $\in$ termes(p) - termes(L(N)) dépiler(o, buts)
11. N ← N' tel que L(N') = L(N) $\cup$ {p}
12. Aller en 4.

FIG. 7.6 – algorithme de Gardent et Striegnitz

nous ne pouvons pas, dans cet algorithme, contraindre la forme que prend l'information nouvelle. Aussi, nous intégrons seulement la possibilité de produire des DAIN.

Notre extension de l'algorithme se situe au point n°8 de l'algorithme de Gardent et Striegnitz : Le point n°7 de l'algorithme original dit que si le but courant est dans le modèle du discours, en d'autres mots, si le référent a été mentionné, on doit choisir une propriété donnée soit dans le modèle du discours soit dans les connaissances du monde. Si le référent n'a pas été mentionné, on doit choisir la propriété le décrivant dans l'ensemble constitué du modèle de discours, du modèle du locuteur, et des connaissances du monde. Dans ce dernier cas, on produit une anaphore associative.

Nous ajoutons dans l'entrée du générateur les données suivantes :

F(L) : la fonction de la description (apporter ou non de l'information, L étant le nom de la description). Elle peut prendre deux valeurs : DRID ou DAIN.
ϕ : contenu sémantique de la description de l'antécédent.
LEX : les ressources lexicales.

Nous proposons dans un premier temps de couper cette condition 7, en insérant une conditionnelle décrite dans les points 8' et 8". Cette conditionnelle permet de tester si le but communicatif est de donner de l'information nouvelle ou de l'information déja connue sur le référent. Si F(L) = DRID, on va rechercher p (la propriété distinguante) dans les bases de connaissances, en recherchant dans l'ordre spécifié dans l'algorithme.

Si F(L) = DAIN, la propriété p ne doit pas se trouver dans les connaissances du monde ou dans le modèle du discours, mais uniquement dans le modèle du locuteur. Nous ne pouvons pas insérer ici les paramètres de réalisation de l'information nouvelle, cet algorithme n'étant destiné qu'à la détermination du contenu du syntagme.

Le point 7' reprend la suite de la conditionnelle d'origine, permettant de générer une anaphore associative.

Plus formellement, ceci peut être reformulé comme nous le présentons en figure 7.5 :

Nous proposons une version intégrale de notre extension de l'algorithme, incluant l'extension pour les anaphores associatives, et celle que nous venons de proposer en Annexe D de notre thèse.

L'étude de corpus a montré qu'il était possible de permettre la génération de reprises qui apportent de l'information nouvelle sur le référent ou qui répètent de l'information connue. Cependant, des problèmes restent encore en suspens, et particulièrement celui du choix du déterminant. Dans le chapitre suivant, nous verrons comment l'apport d'information joue un rôle sur le choix entre le défini et le démonstratif.

7. Si but-courant $\in$ terms(DM) alors R $\leftarrow$ DM
8'. Si F(L) = DRID, alors choisir une formule atomique p applicable telle que :

$\phi \models$ p
ou
WKL $\models$ p
ou
LEX $\models$ p
ou
ϕ + Ctxt $\models$ p
ou
LEX + Ctxt $\models$ p

8". Sinon, F(L) = DAIN, alors sélectionner une formule atomique p applicable telle que Ctxt+WKL $\not\models$ p et SM $\models$ p
7'. Sinon Ctxt $\leftarrow$ DM + SM
8. Essayer de sélectionner une formule atomique p applicable telle que Ctxt+WKL $\models$ p

FIG. 7.7 – Proposition d'extension à l'algorithme de Gardent et Striegnitz

Chapitre 8

Choix du déterminant

Comme nous l'avons montré au chapitre 5, les données théoriques et empiriques connues et confirmées par notre première annotation ne sont pas suffisantes pour permettre le choix entre le défini et démonstratif en génération automatique. En effet, les résultats de notre première annotation ne donnent pas de critères de choix précis du déterminant basé uniquement sur la relation sémantique entretenue par les têtes de l'antécédent et l'anaphore. Il est donc nécessaire de poursuivre l'étude, en utilisant les données issues de notre seconde étude de corpus (présentée au chapitre 7).

8.1 Etude de corpus

8.1.1 Résultats généraux

Le tableau 8.1 reprend les éléments présentés au chapitre précédent et en apportent de nouveaux : dans le premier tableau, on peut lire la proportion de groupes nominaux retenus pour l'étude. On travaille essentiellement sur les descriptions coréférant à des syntagmes nominaux (94% des définis coréférentiels et 83% des démonstratifs coréférentiels). Les colonnes « Ant = Nom » et « Autres » présentent le nombre de cas inclus dans l'étude en valeur absolue, la colonne « prop » la proportion de cas retenus et la colonne « Total Coref » indique le nombre total de descriptions coréférentielles trouvées dans les deux corpus.

Le tableau 8.1 montre ensuite que le démonstratif est plus utilisé que le défini quand la reprise ajoute de l'information sur le référent (19% contre 6% des cas de coréférences retenus, deuxième tableau). Ceci signifie qu'il est probable que le démonstratif serve à attribuer des propriétés à un référent [Corblin, 1987], mais montre aussi que cette propriété, bien que minoritaire, est valable pour le défini, et ne constitue pas la fonction principale du démonstratif. Nous devons donc ne pas éliminer ces cas pour le défini, et tenter de trouver les contraintes permettant ou non de réaliser les deux fonctions identifiées pour les reprises nominales (i. e. répétition d'information ou ajout d'information).

Comme nous l'avons déjà noté au chapitre précédent, la notion d'apport d'information peut être difficile à manipuler parce qu'il s'agit d'un jugement qui peut varier selon les interlocuteurs. Les cas les plus flagrants sont les cas où l'antécédent est un nom propre et où l'information connue provient des connaissances encyclopédiques. Si un annotateur ne

cat	Total Coref.	Ant =Nom	prop	Autres	prop
Définis	1505	1402	94%	103	6%
Démonstratifs	442	369	83,50%	73	16,51%

cat	Total A = Nom	DRID	prop	DAIN	prop
Définis	1402	1314	93,72%	89	6,28%
Démonstratifs	369	299	81,03%	70	18,97%

cat	Total A = Nom	N prop	prop	N com	prop
Définis	1413	228	16,14%	1185	83,86%
Démonstratifs	369	42	11,38%	327	88,62%

cat	Total A = Nprop	DRID	prop	DAIN	prop
Définis	228	191	83,78%	37	16,23%
Démonstratifs	42	21	50%	21	50%

cat	Total A = Ncom	DRID	prop	DAIN	prop
Définis	1185	1127	95,10%	89	4,40%
Démonstratifs	327	278	85%	49	15%

FIG. 8.1 – Résultats généraux de la deuxième annotation

connaissait aucun des individus mentionnés dans le corpus, il pourrait faire passer près de 20% des cas de DRID du défini et 23% des de DRID du démonstratif dans la catégorie DAIN.

Pour cette raison, nous faisons le choix dès maintenant de distinguer les cas d'anaphores dont l'antécédent est un nom propre, des cas dont l'antécédent est un nom commun (cf. troisième tableau, où nous indiquons la proportion de cas dont l'antécédent est un nom commun (N com) et la proportion dont l'antécédent est un nom propre (N prop)). En effet, il nous semble que la variation de jugement entre nouvelle information et information donnée sera importante essentiellement dans les cas où l'antécédent est un nom propre. La proportion de cas d'anaphore dont l'antécédent est un nom propre n'est pas très différente que le déterminant soit défini ou démonstratif. La supériorité du démonstratif pour l'apport d'information est très nette quand l'antécédent est un nom propre (16,23% vs 50%, quatrième tableau) mais aussi quand l'antécédent est un nom commun (4,4% vs 15%, cinquième tableau).

Nous allons maintenant traiter plus précisément les deux types d'anaphores (i. e. les DRID et les DAIN), toujours de façon comparative entre le défini et le démonstratif. Nous commencerons par les reprises répétant de l'information, qui sont les emplois majoritaires des deux déterminants, puis nous étudierons plus précisément les reprises apportant de l'information.

8.1.2 Descriptions Répétant de l'Information Donnée

Le tableau 8.2 présente les résultats de la façon suivante : on trouve en abrégé le nom de chaque catégorie de DRID dans la première colonne. Pour chaque type de déterminant, on trouve les trois informations suivantes : le nombre total de cas trouvés, la proportion que ces cas représentent par rapport au nombre total de descriptions trouvées dans le corpus, et la proportion que ces cas représentent dans la catégorie des DRID. Du tableau 8.2, on peut faire de premières constatations très importantes pour la différenciation entre les utilisations du défini et du démonstratif :

Le défini est utilisé dans les reprises contenant une information provenant explicitement de l'antécédent dans 49% des cas. Dans 51% des cas, l'information répétée est donc inférée.

Pour le démonstratif, en revanche, l'information déjà connue est inférée dans 70% des cas. Ceci peut être vu une fois encore comme l'illustration du fait que le démonstratif permet de reclasser le référent.

Par ailleurs, ceci montre que la reclassification ne signifie pas forcément apport d'information systématique, contrairement à ce qu'on pourrait déduire de la définition de [Corblin, 1987]. L'exemple suivant illustre le fait que la reclassification n'apporte pas toujours d'information :

(c-42) *Soudain, le groupe s'arrête et escalade* ***la clôture****.* ***L'obstacle*** *franchi,* ***les huit hommes*** *parcourent rapidement une dizaine de mètres par la rue Connoly, qui passe sous le pavillon argentin, et débouchent aussitôt devant le pavillon 31 que les Israéliens partagent avec les délégations de Hongkong et de l'Uruguay.*

	DEFINIS			DEMONSTRATIFS		
nombre DRID-Ant. :	642	45,79	48,71	92	24,93	30,77
nombre DRID Ant + Ctxt :	131	9,34	9,96	24	6,50	8,03
nombre DRID Rel. Lex :	226	16,12	17,15	71	19,24	23,75
nombre DRID Rel. Lex + Ctxt :	57	4,06	4,34	25	6,78	8,36
nombre DRID Connaiss. Encycl. :	258	18,40	19,58	87	23,58	29,10
nombre DRID :	1314	93,72	100,00	299	81,03	100,00
total coref :	1402	100,00	-	369	100,00	-

FIG. 8.2 – Résultats DRID

8.1.2.1 Répétition d'informations provenant explicitement de l'antécédent

Antécédent seul L'une des sources principales de l'information répétée est l'antécédent dans les deux cas. Précisons que ceci est valable dans 49% des cas de DRID pour le défini, contre 30% pour le démonstratif. Le défini est donc privilégié dans les cas de reprises directes.

Antécédent et contexte Parfois le contexte est aussi source d'information en plus de l'antécédent, ce qui porte à près de 57% les utilisations du défini s'appuyant sur l'antécédent

et à 38% celles du démonstratif, ce qui renforce l'idée qu'on doit privilégier le défini pour faire des reprises directes.

8.1.2.2 Répétition d'informations inférée

Information inférée des connaissances lexicales L'utilisation d'une relation lexicale pour répéter de l'information est beaucoup plus importante pour le démonstratif que pour le défini. Nous reportons le lecteur au tableau 8.3 pour plus de précision, mais l'hyperonymie semble un moyen très utilisé dans les reprises par le démonstratif, ce qui va dans le sens du rôle de simplification de ce déterminant, identifié par Kleiber et Wiederspiel, puisque le terme utilisé dans la reprise est plus générique que celui utilisé dans l'antécédent. Cette théorie est appuyée par le fait que l'hyponymie n'est presque jamais utilisée. Elle doit alors s'appuyer sur des informations provenant du contexte.

	Démonstratif		Défini	
	Relation lex.	Rel. Lex et Ctxt	Relation lex.	Rel. Lex et Ctxt
hyperonymes	52	3	106	6
synonymes	19	4	118	4
thématiques	0	11	0	24
hyponymes	0	7	2	23

FIG. 8.3 – Relations lexicales impliquées dans la répétion d'information

Rôle du cotexte De manière générale, le cotexte semble pouvoir aider à la résolution des anaphores : 16% pour le démonstratif, 14% pour le défini. Les deux déterminants l'utilisent donc de la même manière.

Information inférée des connaissances du monde Pour finir, regardons les reprises qui n'apportent pas d'information, mais qui passent par les connaissances encyclopédiques pour la résolution de l'anaphore. Pour établir des contraintes, il nous semble fondamental de séparer l'étude des reprises ayant pour antécédent un nom propre, de celles ayant un nom commun pour antécédent, pour les raisons énoncées au début de cette section.

	Défini		Démonstratif	
	nombre	proportion	Nombre	proportion
Nom commun	105	41%	67	77%
Nom propre	153	59%	20	23%
Total	258	100%	87	100%

FIG. 8.4 – Antécédents nominaux des reprises inférées des connaissances du monde

Le tableau 8.4 montre clairement que le démonstratif a un pouvoir reclassifiant en utilisant les connaissances du monde : en effet, il reprend des noms communs dans 77%

des cas, contre 41% pour le défini. On constate aussi que les connaissances du monde servent beaucoup plus lorsqu'il s'agit de lier une reprise définie à un nom propre.

8.1.2.3 Fonction grammaticale de l'antécédent

Nous souhaitons nous pencher maintenant sur la fonction grammaticale des reprises définies et démonstratives pour la raison suivante afin de dire si ce paramètre est pertinent pour le choix du déterminant, dans la mesure où la fonction grammaticale est parfois utilisée pour évaluer la saillance du référent. Le tableau 8.5 présente ces résultats de la façon suivante : chaque catégorie de reprise répétant de l'information a été divisée en quatre, en fonction de la fonction grammaticale de l'antécédent (sujet, objet, objet indirect et circonstant). Verticalement, on trouve les résultats pour les démonstratifs et les définis, sous forme de trois colonnes : la première présente le nombre de cas trouvés pour chaque catégorie dans chaque fonction, la seconde présente la proportion que ces cas représentent sur la totalité des syntagmes coréférentiels, et la troisième, la proportion qu'ils représentent dans la catégorie des anaphores répétant de l'information donnée.

Le tableau 8.5 nous montre les éléments suivants : tout d'abord, il est vrai que le défini reprend - dans le cas de la répétition d'information donnée - majoritairement des éléments en fonction sujet (41% contre 28% pour le démonstratif). Les données qui nous intéressent maintenant sont les suivantes : la reclassification est essentiellement illustrée par la catégorie « répétition d'information inférée d'après les connaissances encyclopédiques ». Or nous constatons que s'il est vrai que la plus grande partie de cette catégorie reprend des antécédents sujets, le nombre de reprise d'éléments occupant d'autres fonctions n'est pas négligeable (99 syntagmes sujets contre 148 occupant d'autres fonctions). L'exemple c-43 illustre cette affirmation. L'antécédent *Barcelone* est complément circonstanciel, et l'anaphore *la ville* est complément de l'adjectif *bloquées*. Nous ne pouvons donc pas dire que la reclassification n'est possible avec le défini que si l'antécédent est sujet, même si elle est majoritaire. Concernant le démonstratif, nous constaterons simplement que la fonction de l'antécédent semble sans conséquence sur le type de la reprise : il est vrai que le défini reprend globalement moins d'objets directs et de circonstants, mais les proportions restent importantes.

(c-43) *Une brusque montée de l'hygrométrie (le taux d'humidité est passé de 60% à 90%) et une absence totale de vent ont provoqué à* ***Barcelone****, dans la nuit du 4 au 5 septembre, une serie d' intoxication, dont deux mortelles. Asphyxiées par les émanations de la circulation automobile et les fumées d'usines soudain bloquées au-dessus de* ***la ville****, quelque soixante personnes ont dû être transportées d'urgence à l'hopital.*

8.1.3 Descriptions Ajoutant de l'Information Nouvelle sur le référent

Le tableau 8.6 présente les résultats de l'annotation de corpus sur les descriptions ajoutant de l'information. On retrouve chacune des catégories décrites dans le chapitre précédent (chapitre 7) (Rel. Lex. pour l'ajout d'information par une relation lexicale, l'ajout d'information par les modifieurs, l'ajout d'information par les modifieurs et une relation lexicale, et l'ajout d'information par un groupe nominal). Deux grandes colonnes

	DEMONSTRATIFS			DEFINIS		
DRID sujet :	86	23,31	28,76	547	39,02	41,63
DRID Ant sujet :	33	8,94	11,04	262	18,69	19,94
DRID Ant + C sujet :	11	2,98	3,68	78	5,56	5,94
DRID Rel. Lex sujet :	18	4,88	6,02	95	6,78	7,23
DRID Lex Rel + C sujet :	5	1,36	1,67	13	0,93	0,99
DRID Connaiss. Encycl sujet :	19	5,15	6,35	99	7,06	7,53
DRID obj ind :	48	13,01	16,05	184	13,12	14,00
DRID Ant obj ind :	10	2,71	3,34	83	5,92	6,32
DRID Ant + C obj. ind. :	6	1,63	2,01	20	1,43	1,52
DRID Rel. Lex obj. ind. :	13	3,52	4,35	42	3,00	3,20
DRID Lex Rel + C obj. ind. :	3	0,81	1,00	6	0,43	0,46
DRID Connaiss. Encycl obj. ind. :	16	4,34	5,35	33	2,35	2,51
DRID objet :	100	27,10	33,44	311	22,18	23,67
DRID Ant objet :	30	8,13	10,03	158	11,27	12,02
DRID Ant + C objet :	6	1,63	2,01	41	2,92	3,12
DRID Rel. Lex objet :	24	6,50	8,03	47	3,35	3,58
DRID Lex Rel + C objet :	12	3,25	4,01	10	0,71	0,76
DRID-Connaiss. Encycl objet :	28	7,59	9,36	55	3,92	4,19
DRID circ :	49	13,28	16,39	226	16,12	17,20
DRID Ant circ :	17	4,61	5,69	117	8,35	8,90
DRID Ant + C circ :	1	0,27	0,33	11	0,78	0,84
DRID Rel. Lex circ :	14	3,79	4,68	36	2,57	2,74
DRID Lex Rel + C circ :	4	1,08	1,34	2	0,14	0,15
DRID Connaiss. Encycl circ :	13	3,52	4,35	60	4,28	4,57
TOTAL DRID	299	-	100,00	1314	-	100,00
total coref :	369	100,00	-	1402	100,00	-

FIG. 8.5 – fonction de l'antécédent pour les DRID

présentent séparément les résultats pour les définis et les démonstratifs, et chacune d'elle est séparée en trois, présentant tout d'abord le nombre d'occurrences trouvées pour chaque phénomène, la proportion qu'il représente parmi les syntagmes coréférentiels, et la proportion qu'il représente parmi les descriptions ajoutant de l'information nouvelle sur le référent.

	DEFINIS			DEMONSTRATIFS		
DAIN Rel. Lex :	7	0,50	7,87	2	0,54	2,86
DAIN modifieurs :	22	1,57	24,72	19	5,15	27,14
DAIN Rel. Lex. + mod :	2	0,14	2,25	3	0,81	4,29
DAIN syntagme nominal :	52	3,71	58,43	46	12,47	65,71
DAIN :	88	6,28	100,00	70	18,97	100,00
total coref :	1402	100,00	-	369	100,00	-

FIG. 8.6 – Résultats DAIN

8.1.3.1 Utilisation de modifieurs

L'utilisation des modifieurs est un moyen très courant pour ajouter l'information. C'est effectivement un moyen simple, dans la mesure où le locuteur peut utiliser un nom permettant facilement de faire le lien avec l'antécédent, et ajouter de l'information en adjoignant des modifieurs à ce nom. Nous n'avons pas poussé plus avant la recherche, mais il semble que les modifieurs n'auront pas la même nature grammaticale selon le déterminant utilisé (par exemple, on sait que certains types de modifieurs, comme les relatives restrictives, ou les génitifs sont la plupart du temps utilisés avec le défini). Il est intéressant de noter que quel que soit le déterminant utilisé, l'ajout d'information par les modifieurs représente environ un quart des utilisations de syntagmes nominaux coréférents ajoutant de l'information sur leur antécédent, ce qui est important mais loin d'être majoritaire.

8.1.3.2 Utilisation d'hyponymes (avec ou sans modifieurs)

Ce moyen est globalement peu utilisé. Notons simplement que le défini semble plus apte à ce type de reprise (environ 10%) que le démonstratif (environ 7%), la différence n'étant cependant pas très significative à nos yeux.

8.1.3.3 Utilisation d'un syntagme sans lien avec l'antécédent et le contexte

Les définis reprennent dans près de 60% des cas des noms avec lesquels ils n'ont pas de lien lexical. La proportion de démonstratifs dans ce cas est de 65%, ce qui ne permet pas, une fois encore, de distinguer réellement les utilisations du défini et du démonstratif. Cependant, comme pour les anaphores répétant de l'information donnée, nous souhaitons approfondir notre étude, en différenciant les anaphores ayant un nom propre comme antécédent, des anaphores ayant un nom commun comme antécédent.

8.1.3.4 Antécédents nom propres vs. noms communs

Nous retirons des tableaux suivants les lignes concernant les relations lexicales, parce qu'elles ne sont pas utilisables dans le cadre d'une comparaison nom commun / nom propre, étant admis qu'on ne peut pas établir de relation lexicale entre un nom propre et un nom commun.

	DEFINIS		DEMONSTRATIFS	
Catégorie	NB	Prop	NB	Prop
DAIN antec = n.p. :	37	42,05	21	30,00
DAIN-mod antec = n.p. :	2	2,27	1	1,43
DAIN SN antec = n.p. :	33	37,50	20	28,57
DAIN antec = n.c.	51	57,95	49	70,00
number DAIN-mod :	20	22,73	18	25,71
number DAIN SN :	19	21,59	26	37,14

FIG. 8.7 – DAIN Noms propres vs. noms communs

Nous constatons alors (tableau 8.7) que les résultats sont très tranchés. Alors que 42% des définis reprennent un nom propre, seulement 30% des démonstratifs le font. Ceci amène directement au fait que les reprises sans lien lexical avec l'antécédent sont très majoritaires pour le démonstratif quand l'antécédent est un nom commun (37% pour le démonstratif contre 20% pour le défini). En revanche, les deux déterminants reprennent énormément de noms propres de cette façon, le défini étant le plus utilisé (37% contre 28% pour le démonstratif). Pour terminer, constatons que lorsque l'antécédent est un nom propre et que l'anaphore ajoute de l'information sur le référent, il y a très peu d'utilisations des modifieurs seuls pour apporter de l'information. En général, l'apport d'information se fait par l'intermédiaire d'un syntagme entier. Les résultats vus précédemment à propos de l'utilisation des modifieurs pour apporter de l'information sont donc valables essentiellement pour les noms communs.

8.1.3.5 Fonction de l'antécédent

Comme pour les descriptions n'ajoutant pas d'information, nous nous intéressons maintenant à connaître la fonction grammaticales de l'antécédent quand la reprise donne de l'information nouvelle. Nous regarderons particulièrement les éléments suivants :

Tout d'abord, une fois encore, la fonction majoritairement reprise pour les deux déterminants est le sujet. Notons tout de même que ceci est plus flagrant pour le défini.

Si nous regardons de plus près les reprises d'antécédent sujet, nous nous apercevons que la reclassification est aussi importante avec le défini qu'avec le démonstratif. Il semble donc que le démonstratif n'est réellement pas un substitut du défini lorsque la reclassification doit se faire sur autre chose qu'un sujet, puisqu'il est tout autant employé pour les éléments sujets. En revanche, la reclassification apparaît beaucoup plus souvent avec le démonstratif pour toutes les autres fonctions.

	DEMONSTRATIFS			DEFINIS		
DAIN subj :	30	8,13	42,86	52	3,71	59,09
DAIN Rel. Lex subj :	0	0,00	0,00	3	0,21	3,41
DAIN modifieurs subj :	5	1,36	7,14	11	0,78	12,50
DAIN Rel. Lex. + mod subj :	3	0,81	4,29	2	0,14	2,27
DAIN SN subj :	22	5,96	31,43	35	2,50	39,77
DAIN obj :	21	5,69	30,00	17	1,21	19,32
DAIN Rel. Lex obj :	2	0,54	2,86	0	0,00	0,00
DAIN modifieurs obj :	7	1,90	10,00	7	0,50	7,95
DAIN Rel. Lex. + modifieurs obj :	0	0,00	0,00	0	0,00	0,00
DAIN SN obj :	12	3,25	17,14	9	0,64	10,23
DAIN obj ind :	6	1,63	8,57	9	0,64	10,23
DAIN Rel. Lex obj ind :	0	0,00	0,00	2	0,14	2,27
DAIN modifieurs obj ind :	4	1,08	5,71	2	0,14	2,27
DAIN Rel. Lex + mod obj ind :	0	0,00	0,00	0	0,00	0,00
DAIN SN obj ind :	2	0,54	2,86	4	0,29	4,55
DAIN circ :	9	2,44	12,86	8	0,57	9,09
DAIN Rel. Lex. circ :	0	0,00	0,00	2	0,14	2,27
DAIN modifieurs circ :	3	0,81	4,29	2	0,14	2,27
DAIN Rel. Lex. + mod circ :	0	0,00	0,00	0	0,00	0,00
DAIN SN circ :	6	1,63	8,57	3	0,21	3,41
TOTAL DAIN	70	-	100	88	-	100
total coref :	369	100	-	1402	100	-

FIG. 8.8 – Résultats fonctions DAIN

8.1.4 Synthèse

Des éléments importants et troublants par rapport aux études théoriques apparaissent dans ce corpus. Nous les résumons brièvement maintenant, avant de poser les questions qu'ils ouvrent

8.1.4.1 Description répétant l'information

Le démonstratif autorise plus d'inférences que le défini de façon générale, particulièrement lorsqu'elles ont lieu à partir des connaissances encyclopédiques, mais aussi lorsqu'elles proviennent des connaissances lexicales. Ceci illustre donc la capacité de reclassification du démonstratif, sans pour autant la retirer au défini, puisque celui-ci permet l'inférence dans près de la moitié des cas de reprise sans ajout d'information.

Nous devons ajouter que le défini n'est pas reclassifiant majoritairement dans les cas où l'antécédent est sujet. Par ailleurs, le fait que le démonstratif puisse être employé dans sa fonction reclassifiante aussi dans les cas où l'antécédent est sujet n'en fait pas une caractéristique distinctrice des déterminants. Le démonstratif est autant utilisé pour la reprise de sujets ou d'objets, la notion de saillance ou de proximité ne nous semble donc pas pertinente pour distinguer les déterminants.

8.1.4.2 Descriptions ajoutant de l'information

Les modifieurs sont un moyen courant avec les deux déterminants pour ajouter de l'information, et l'hyponymie sert peu. La reclassification est très courante avec le défini comme le démonstratif. La reclassification définie avec un nom commun comme antécédent est plus rare, mais est un phénomène existant. Ici encore la fonction de l'antécédent ne semble pas non plus pertinente pour le choix du déterminant.

8.1.4.3 Paramètres non pris en compte

Distance entre l'antécédent et la reprise La notion de distance entre l'antécédent et l'anaphore est un notion souvent invoquée pour expliquer le choix du déterminant. Nous n'avons pu ici utiliser ce paramètre pour plusieurs raisons. La première est liée à un problème de définition : doit-on mesurer la distance en nombre de phrases, de paragraphes, de clauses ou de mot ? Aucune théorie ne répond à la question, et ceci est explicable. Si le nombre de clauses peut paraître une bonne mesure, il est très certainement parasité par le nombre de mots contenus dans la clause. Par ailleurs, ceci ne rendrait pas compte de constructions détachées ou d'inversions de l'ordre canonique de la phrase. L'étude semble donc difficile à mener sur corpus, malgré les moyens dont nous disposons. Nous avons par ailleurs annoté comme antécédent le dernier syntagme plein utilisé pour la référence. Il est donc possible dans de nombreux cas que la chaîne de référence contienne des pronoms coréférents entre les deux syntagmes annotés, ce qui rend encore plus difficile le calcul. Nous avions malgré tout calculé la distance en nombres de mots entre l'antécédent annoté et l'anaphore, mais les résultats n'étaient pas exploitables, car les distances varient

énormément d'un cas à l'autre (les moyennes des distances pour chaque déterminant sont inférieures aux écarts-types entre les distances).

Place dans la chaîne anaphorique Il nous semble très important de dire dès maintenant que l'une des limites de notre travail est liée au fait que nous n'avons considéré que des paires antécédent / anaphore nominale, sans tenir compte de la place occupée par chaque syntagme dans la chaîne anaphorique. Il est difficile de dire dans quelle mesure cette information serait pertinente étant donné les difficultés qu'on trouve pour définir ce qui fait partie ou non d'une chaîne anaphorique [Schnedecker, 1997]. Nous n'excluons malgré tout pas l'idée que les phénomènes sont différents cependant, si on étudie la seconde mention d'un référent ou une mention postérieure.

Il semble désormais qu'une série de paramètres donnés pour justifier le choix entre défini et démonstratif ne soient pas pertinents pris un à un, et que d'autres soient difficiles à formaliser. Si la fonction de l'antécédent n'est pas déterminante, si la relation sémantique entretenue par les deux noms n'est pas pertinente, si la provenance des inférences telles qu'on l'a définie précédemment n'intervient que partiellement et si la capacité à ajouter des informations sur le référent dans l'anaphore ne permet pas de choisir un déterminant, quels peuvent être les paramètres utilisables en génération ?

Dans les sections suivantes, nous étudions les divers critères de choix connus pour les déterminants, nous en présentons de nouveaux, puis nous établissons des contraintes sur la génération de déterminants.

8.2 Contraintes connues sur l'utilisation des déterminants

Pour faire un algorithme de génération, il est nécessaire d'avoir une théorie différentielle de la détermination. Il nous apparaît par ailleurs que les critères donnés par les études théoriques et notre classification doivent être croisés pour pouvoir obtenir des contraintes fiables pour la génération de descriptions définies et démonstratives. Nous allons maintenant faire une liste de ces contraintes, avec des exemples les illustrant. Nous montrerons ensuite que certains exemples, bien que répondant aux contraintes, ne présentent pas le résultat attendu (sections 8.2.1 à 8.2.5).

Nous terminerons en montrant que pour pouvoir générer ces exemples, il faut ajouter et croiser des contraintes sur la forme et le contenu des informations véhiculées dans la reprise (section 8.3).

8.2.1 Unicité et Familiarité

Familiarité Il nous apparaît clairement que la familiarité telle qu'elle est définie par [Dale et Reiter, 1995] puis par [Gardent et Striegnitz, 2003] n'est pas utile en génération pour choisir entre le défini et le démonstratif, les deux déterminants impliquant forcément la mention antérieure des référents des expressions qu'ils déterminent, ou d'un antécédent dans le cas des anaphores associatives. Pour les deux déterminants, le référent doit

être familier, même si cette familiarité est indirecte, comme dans le cas des anaphores associatives.

Unicité L'unicité n'est par ailleurs pas plus déterminante, dans la mesure où le démonstratif isole lui aussi un élément qui peut parfois être unique. L'unicité peut fonctionner pour choisir le défini lorsque les descriptions réfèrent à des objets qui sont non seulement uniques dans le discours, mais aussi dans le monde qui nous entoure.

Ainsi, l'exemple suivant illustre très bien la notion de familiarité et d'unicité :

(c-44) *Impériale,* ***Jeannie Longo*** *n' a laissé aucune chance, vendredi 4 septembre, à ses adversaires lors des championnats du monde de cyclisme sur route, qu' elle a remportés pour la troisième année consécutive. Victorieuse en juillet du Tour de France féminin, puis le mois suivant de la Coors Classic americaine,* ***la sportive grenobloise*** *a terminé détachée sur le circuit autrichien de Villach.*

Dans cet exemple, une série d'inférences est à réaliser pour résoudre la coréférence, mais seul un individu est mentionné dans le discours. La contrainte d'unicité de la description coréférentielle est donc respectée. On notera que l'utilisation du démonstratif est possible, mais semble moins bonne que l'utilisation du défini.

Cependant, on trouve de nombreux exemples de cas où l'unicité n'est manifestement pas le paramètre déterminant pour choisir entre défini et démonstratif. En d'autres mots, il nous faut trouver ce qui différencie de l'exemple c-44 les deux exemples suivants :

(c-45) *Quant a* ***Klaas de Jonge****, il vivait depuis le 29 juillet 1985 dans les locaux, aujourd'hui desaffectés, de l'ambassade des Pays-Bas a Pretoria. Les négociations avec les autorités de La Haye pour faire sortir* ***ce quinquagénaire accusé d'avoir transporté des armes pour le compte de l'ANC*** *n'avaient rien donné .*

(c-46) *Parallèlement, il prendrait la présidence de* ***Ficofrance, la société financière de GMF****. Ce groupe familial, qui a realisé en 1986 un chiffre d'affaires de 5 milliards de francs, figure parmi les candidats les plus sérieux à l'appel d'offres lancé par Albin Chalandon pour la construction de 15000 places de prison supplémentaires d'ici à 1990.*

Dans les exemples c-44 à c-46, l'utilisation des deux déterminants est possible. Pourtant, dans chaque cas, l'auteur a fait un choix qui semble plus naturel que l'autre. Il est donc clair que d'autres critères de choix entrent en jeu, et il nous faut déterminer lesquels.

8.2.2 Opposition notionnelle

Le concept d'opposition notionnelle a été utilisé par [Corblin, 1987] pour expliquer les différences d'interprétation entre le défini et le démonstratif. Le démonstratif s'interprète comme une extraction d' un objet à l'intérieur de sa catégorie, tandis que le défini permet d'identifier l'objet en utilisant sa catégorie pour établir un contraste avec les autres objets du contexte. Dans l'exemple suivant, on retrouve aussi la notion de description uniquement identifiante dans le contexte dans les syntagmes nominaux *le commissaire Hamel* et *l'inspecteur Poisson*. On note que la reprise aurait été possible même si les noms propres

n'avaient pas été mentionnés à nouveau. C'est pourquoi cet exemple illustre l'idée d'opposition notionnelle, puisqu'on oppose les référents par leur catégorie (un inspecteur est opposé à un commissaire). Cependant, cette explication n'est pas satisfaisante pour les exemples c-44, c-45 et c-46, dans lesquels on ne peut pas opposer les référents à un autre référent du contexte par sa catégorie, puisqu'il n'y a pas d'autre référent dans le contexte, et que les deux déterminants sont utilisés.

(c-47) *Ce ne sont là que trois des seize chefs d'inculpation retenus contre* ***Alain Hamel, quarante-cinq ans, commissaire de police****, et* ***Didier Poisson, trente-sept ans, inspecteur de police****. Tous deux ont reconnu avoir « mis en règle », entre juillet et octobre 1986, pour le compte de tiers inconnus, des voitures volées qu'ils ont ensuite revendues à des particuliers ou à des garagistes. Le procureur, M. Jean-Claude Thin, a requis trois ans de prison ferme pour* ***le commissaire Hamel*** *et de dix-huit mois à deux ans, dont une partie avec sursis, pour* ***l'inspecteur Poisson****.*

8.2.3 Rupture discursive

Kleiber [Kleiber, 1986] montre que l'idée d'opposition notionnelle ne fonctionne pas toujours pour expliquer le choix de l'un ou l'autre déterminant, et avance alors la notion de rupture dans les événements décrits par le texte. Nous interprétons cette notion comme, par exemple, un changement d'entité focalisée. Le démonstratif serait alors utilisé pour remettre un référent en focus. Nous pouvons alors ainsi justifier l'emploi du démonstratif dans l'exemple c-46 de la section précédente. En effet, le sujet change, puisqu'on passe d'une phrase où le personnage en focus est un PDG (repris par le pronom *il*) à une phrase où l'élément en focus est la société qu'il reprend.

En revanche, il nous est difficile de voir une rupture discursive dans l'exemple c-48, où le démonstratif est utilisé, et où le coureur cycliste dont il est question reste focalisé, et de ne pas en voir dans l'exemple c-49, où le défini est utilisé, et où le personnage désigné par les termes *le mec* et *le cerbère* est tout d'abord très focalisé puis beaucoup moins.

(c-48) *Apres Colas-Magne, vainqueur en tandem,* ***Richard Vivien*** *a créé une surprise de taille en devenant samedi champion du monde de la catégorie, que l'on définissait il y a peu de temps comme étant celle des « purs ».* ***Ce Normand de vingt-trois ans, remarqué par Yves Hezard****, est peut-être un pur, mais il possède déjà beaucoup de métier si l'on retient la manière dont il a battu l'Allemand de l'Ouest Bolts, le Danois Pedersen, le Polonais Mierzejewski à l'issue d'une course à l'économie, fatale aux Sovietiques .*

(c-49) ***Le mec*** *roule des yeux. Un cliché vivant ? Je range vite mes instruments de travail, en me fendant d'un sourire bête. Repars la queue basse, en pensant qu'un jour ma graphomanie me perdra. Beaucoup plus loin, à distance respectable du* ***cerbère****, sous la protection rapprochée d' un flamboyant, je griffonne les mots que je craignais d'oublier.*

8.2.4 Attribution de propriété et fonction de l'antécédent

D'après [Corblin, 1987], le démonstratif permet d'attribuer une propriété à un référent. Le défini peut le faire aussi, mais essentiellement si l'antécédent est sujet de la phrase antérieure. Nos exemples c-48 et c-49 illustrent cette affirmation. En revanche, les exemples c-50, c-51 et c-52 vont à l'encontre de la théorie de Corblin.

(c-50) *Totalement inconnu il y a deux ans, le nouveau patron de* ***Technip*** *a su en remontrer aux plus endurcis. Si* ***la première société française d'ingénierie*** *a survecu à la crise, elle ressort de l'epreuve totalement transformée au terme d'un traitement de cheval aussi violent qu'inhabituel dans les mœurs industrielles francaises.*

(c-51) *Ce sont les fidèles soldats de* ***Michel Rocard****. Le PS ? Pourquoi parler du PS ? Ils n'ont même pas voulu poser une question à* ***l'ancien ministre*** *sur sa candidature, tant elle leur semble évidente.*

(c-52) *Mais le dernier coup d' Etat - le cinquième en dix-huit mois et le premier aussi sanglant - attise les craintes des investisseurs qui voient là une nouvelle démonstration de l'instabilite politique du* ***pays****. L'étranger avait pourtant recommencé à s'interesser à* ***l'archipel****.*

8.2.5 Source de l'inférence

Notre étude a montré que la source de l'inférence (antécédent, contexte, connaissances lexicales et connaissances encyclopédiques) n'était pas non plus particulièrement déterminante dans le choix du déterminant ; en effet bien que le démonstratif soit plus utilisé dans les reprises nécessitant de faire des inférences sur les connaissances encyclopédiques quand l'antécédent est un nom commun, ce n'est pas toujours le cas. Une série d'exemples viennent prouver que le défini peut être utilisé pour ce type de reprise (cf. exemples c-52 et c-49), et les exemples c-48 et c-45 montrent que le démonstratif sert aussi dans ce type de reprise avec un antécédent sous forme de nom propre.

8.3 Nécessité de combiner les contraintes

Il nous semble, au vu des exemples cités dans la section précédente, que les critères de choix entre un défini et un démonstratif pris un à un ne peuvent fonctionner. Il est probable que certains soient prioritaires sur d'autres, ce que nous allons chercher à déterminer maintenant. Notre étude a donc du être étendue à un croisement de paramètres, et en observant les exemples du corpus, nous sommes arrivée aux généralités suivantes. Les éléments que nous avons mis en valeur ne fonctionnent peut être pas sur la totalité des cas du corpus, mais dans l'ensemble, permettent de générer les anaphores que nous y avons trouvées. Certains cas restent indéterminés (les déterminants se substituent parfaitement), mais nous donnons une préférence en fonction de la fréquence d'apparition du phénomène). Nous montrerons dans cette section que les paramètres à combiner sont les suivants :

L'apport ou non d'information sur le référent et la provenance des inférences se révèlent importants, combinés à des critères d'unicité, et de focalisation. Ceci n'est pas surprenant au vu de l'étude que nous avons menée et des théories antérieures.

En revanche, une étude détaillée de nos exemples nous amènera à montrer que la forme syntaxique des modifieurs et leur sémantique joue aussi un grand rôle dans le choix du déterminant, ce qui est nouveau.

Nous commencerons par des cas transversaux aux catégories DRID et DAIN, et nous montrerons ensuite les différences d'utilisation des déterminants pour chacune de ces catégories. A la fin de chaque paragraphe, nous faisons ressortir typographiquement la contrainte identifiée, en la numérotant, afin de pouvoir établir une liste qui nous amènera à l'algorithme de choix du déterminant.

8.3.1 Contraintes communes aux catégories DRID et DAIN

8.3.1.1 Unicité

Si l'objet est unique, on peut utiliser le défini ou le démonstratif. Cependant, le contexte dans lequel le référent est unique est variable, et lorsque le référent est unique dans le monde ou dans la situation de communication, on utilise le défini. L'exemple c-53 illustre l'unicité dans le monde : nos connaissances du monde nous disent que la LCR est un parti politique, et qu'un parti politique n'a qu'un seul porte-parole. Le référent du syntagme *le porte-parole de la LCR* est donc unique.

Nous parlons d'unicité dans la situation de communication pour l'exemple c-54 pour les raisons suivantes : La description *le chef de l'Etat* n'est valable que dans un contexte national, le monde comptant plusieurs chef de l'Etat au même moment. Il a donc fallu une résolution - au moins implicite - du référent du nom *Etat* sur *Etat français* pour parvenir à une unicité dans le monde. Nos connaissances du monde nous indiquent qu'un Etat n'a qu'un seul chef, et donc si l'Etat dont on parle est la France, l'individu nommé *chef de l'Etat* est unique.

Cette notion de l'unicité est celle utilisée par [Russell, 1905] dans sa théorie sur le défini.

(c-53) ***M. Alain Krivine*** *a appelé, le vendredi 4 septembre, « l'ensemble des candidats de la gauche à s'engager à se désister au second tour en faveur du candidat de gauche qui sera arrivé en tête au premier tour » de l' élection présidentielle. A l'occasion d'une conférence de presse de rentrée,* ***le porte-parole de la LCR*** *a ajouté : « un tel désistement ne devra pas signifier l'octroi d' un blanc-seing à la gauche pour qu'elle recommence l' expérience désastreuse de 1981 - 1986 de capitulation devant la droite.*

(c-54) ***M. Mitterrand*** *constate : « j' ai beaucoup d'admiration pour de Gaulle, mais ça ne m'a jamais conduit, on le sait bien, à me sentir obéissant, prêt à me couler dans le moule, tout aussitôt rallié à la moindre de ses idées, dont certaines étaient mauvaises." Enfin* ***le chef de l' Etat*** *a révélé qu'il n'a « pas toujours aimé la façon dont la classe dirigeante, qu'il [de Gaulle] avait sauvée en 1958, s' est organisée pour le chasser du pouvoir en 1969.*

Contrainte n°1 : Si le référent est unique dans les connaissances du monde, utiliser le défini.

8.3.1.2 Noms prédicatifs

Si le nom tête du syntagme anaphorique est un nom prédicatif dont l'un des arguments est instancié, on utilise le défini. En revanche, si aucun argument n'est instancié, on utilise le démonstratif. Nous le montrons dans l'exemple c-55, où la répétition de l'argument instancié permet la réalisation du défini. On peut par ailleurs modifier l'exemple de façon, en seconde mention, à ne pas réaliser l'argument, et dans ce cas, le démonstratif est quasiment obligatoire. Ceci est le cas dans l'exemple c-55, où on pourrait dire *obtenir cette libération*, mais pas **obtenir la libération*, ni **obtenir cette libération du jeune coopérant*. Cette observation se fonde sur l'étude des 56 noms prédicatifs coréférentiels employés avec le défini, et des 43 employés avec le démonstratif relevés dans notre corpus. Nous avons trouvé que 73% des définis sont employés avec leurs arguments, tandis que 77% des démonstratifs sont employés sans leurs arguments. Les 27% de définis apparaissant sans leurs arguments apparaissent pour la plupart d'entre eux dans de très longues chaînes de référence en reprise directe (exemple c-56). Les emplois du démonstratif avec des noms prédicatifs réalisant leurs arguments sont des cas de simplification de l'antécédent (exemple c-57).

(c-55) ***La libération de Pierre-Andre Albertini*** *constituerait un « bon point » pour M. Chirac. Elle placerait en revanche le Parti communiste en porte à faux. Le PC, qui a monté une importante campagne en faveur de celui qu'il appelle « l'otage de l'apartheid », a accusé en effet le gouvernement de ne rien faire pour obtenir* ***la libération du jeune cooperant****.*

Dans l'exemple c-56, on note quatre mentions de la *fuite de sodium*, dont une dans le titre. Les deux premières mentions réalisent un argument, et l'emploi du défini est justifié en troisième mention par l'emploi de la relative restrictive. La quatrième mention ne comporte en revanche pas de justification précise, si ce n'est que l'expression désigne une entité déjà parfaitement identifiée par le locuteur à ce stade du texte. Notons pour terminer que la première mention de la *fuite* sans argument est déterminée par un démonstratif.

(c-56) ***La fuite de Superphenix*** *est localisée dans le bas du barillet.* ***La fuite de sodium de Superphenix*** *est enfin localisée. Commencées le 1er septembre, les opérations de vidange du barillet, ce réservoir annexe de sodium utilisé lors des remplacements d'éléments combustibles, ont permis, le samedi 5 septembre, en fin de soirée, de localiser* ***la fuite, qui, depuis mars dernier, laissait s'écouler une vingtaine de litres par heure****.* ***Cette fuite*** *se trouve « au niveau d'un support de tuyauterie situé à l'interieur de la cuve », indique la direction de la centrale de Creys-Malville. Elle n'est donc pas comme on l'avait pensé au niveau du « bec de cafetière », l'ouverture du conduit par lequel les éléments combustibles transitent entre la cuve du surgénérateur et le barillet, mais un peu plus bas. D'après la direction, cette position basse devrait « permettre d'engager assez un examen visuel et d'obtenir rapidement les premières indications quant à la nature et à l'origine du defaut ». La localisation de* ***la fuite*** *puis son examen vont permettre de savoir si elle est réparable ou si un remplacement du barillet est nécessaire.*

Ensuite, l'exemple c-57 illustre la simplification d'un antécédent propositionnel, ce qui semble montrer que la contrainte de simplification de l'antécédent impliquant la présence du démonstratif est plus importante que la contrainte liée à la présence des arguments quand le nom est prédicatif.

(c-57) ***« Le 17 septembre 1983, je faisais mon jogging quand un manège inhabituel a attiré mon attention. Il était près de minuit ... ». Cette déclaration*** *d'un temoin, qui avait été entendu en 1984, mais n'avait pas, semble-t-il, à l'epoque été pris au sérieux par les policiers, pourrait donner une nouvelle dimension à l'enquête sur l'affaire dite des « tueurs fous du Brabant wallon »*

Contrainte n°2 : Si le nom utilisé dans la reprise est un nom prédicatif dont les arguments sont réalisés, utiliser le défini. Si c'est un nom prédicatif sans arguments réalisés, utiliser un démonstratif.

8.3.2 Description Répétant l'Information Donnée

Il est des cas où il nous semble important de conserver l'idée que l'information contenue dans la reprise n'est pas nouvelle. Nous traiterons dans un premier temps les cas où l'information connue provient de l'antécédent, du cotexte ou des connaissances linguistiques, et dans un second temps des cas où l'information connue est inférée des connaissances du monde. Dans ces derniers cas, nous séparerons ceux dont l'antécédent est un nom propre de ceux dont l'antécédent est un nom commun.

8.3.2.1 L'information est explicitement contenue dans l'antécédent, inférée du cotexte et/ou inférée d'une relation lexicale

Focalisation de l'antécédent Si l'antécédent est en position sujet dans la phrase précédente, on reprend le groupe nominal (directement ou non) par un défini (exemple c-58) plutôt que par un démonstratif. Si l'antécédent n'est pas en position sujet, on utilise le démonstratif (exemple c-59). La reprise d'un élément sujet en répétant de l'information donnée provenant du cotexte, de l'antécédent ou des connaissances linguistique représente 75,7% des emplois du défini en DRID, tandis qu'il représente 51,1% des emplois du démonstratif en DRID. Cette contrainte ne semble pas suffire en elle même pour décider du choix entre défini et démonstratif, mais peut constituer un premier filtre avant de faire un choix sur une contrainte plus déterminante.

(c-58) ***L'aviation israélienne*** *a effectué le samedi 5 septembre un raid sur le camp de refugiés palestiniens d'Ain-el-Heloue, dans les faubourgs de Saida, chef-lieu du Liban-sud, ont rapporté les correspondants sur place.* ***Les chasseurs-bombardiers israéliens*** *ont effectué à partir de 10h15 locales plusieurs attaques en piqué sur ce camp qui compte soixante mille habitants, bombardant deux positions militaires palestiniennes.*

(c-59) *Si une banque comme la Citicorp, le plus grand émetteur de cartes aux Etats-Unis, avec 23 %* ***du marché****, ne peut plus se permettre d'avoir sa propre carte de crédit,*

c' est qu' il se passe de drôles de choses sur ce créneau », remarque John Pollock, directeur d'une lettre specialisée. ***Ce marché*** *que l'on dit saturé reste en effet très porteur.*

Contrainte n°3 : Si la reprise est une DRID dont la source de l'inférence n'est pas la base des connaissances du monde, utiliser le défini si l'antécédent est sujet, utiliser le démonstratif si l'antécédent occupe une autre fonction (que la fonction sujet).

Simplification de l'antécédent Dans les cas où le syntagme de reprise simplifie l'antécédent (cas d'antécédents multiples ou d'antécédents propositionnels), on utilise le démonstratif, ce qui est en accord avec [Wiederspiel, 1994]. Nous ne disposons pas de chiffres précis sur ce type de reprise car l'identification des antécédent multiples n'a pu être réalisée automatiquement. Nous dirons donc que la reprise d'antécédents non nominaux représente 18% des utilisations du démonstratif contre 6% des emplois du défini (tableau 8.1). Par ailleurs, un examen informel du corpus nous a confirmé la fréquence plus importante des emplois simplifiant des antécédents multiples avec le démonstratif.

(c-60) *De juin 1986 à juin 1987, la masse monétaire au sens large a augmenté respectivement, dans ces pays,* ***de 10,8%, de 19,2% et de 7,4%****. En termes réels, c'est-à-dire compte tenu de la hausse des prix,* ***ces progressions*** *sont les plus fortes jamais atteintes au cours des dix dernières années.*

Contrainte n°4 : Si la reprise est une DRID dont la source d'inférence n'est pas la base des connaissances du monde, et si l'antécédent est réalisé par plusieurs SN, utiliser le démonstratif.

8.3.2.2 Inférences sur les connaissances encyclopédiques quand l'antécédent est un nom commun

La règle ne semble pas liée à la focalisation. On peut utiliser les deux déterminants, mais le démonstratif est très fortement favorisé (22% des cas contre 8% pour le défini), même si l'entité est unique dans le contexte. Par ailleurs, dans nos exemples, le défini peut être remplacé par un démonstratif.

(c-61) *Soudain, le groupe s'arrête et escalade* ***la clôture****.* ***L'obstacle*** *franchi, les huit hommes parcourent rapidement une dizaine de mètres par la rue Connoly, qui passe sous le pavillon argentin, et débouchent aussitôt devant le pavillon 31 que les Israéliens partagent avec les délégations de Hongkong et de l'Uruguay.*

(c-62) *M. Barre prend ensuite la parole pour immédiatement évoquer* ***l'élection présidentielle****. « Nous sommes entrés, souligne-t-il, dans la phase finale de l'intermède institutionnel que nous vivons depuis mars 1986. (...) Conformément à ce que j' avais annoncé, je n'ai rien fait qui put empêcher le déroulement de cette expérience. Je me suis abstenu de commenter ses épisodes, péripéties et cliquetis... Je souhaite qu'elle garde jusqu'à son terme cette pureté de cristal afin qu'elle puisse contribuer*

à l'édification durable des Français. » Parlant de ***ce rendez-vous de 1988*** *comme d' une « nouvelle donne », M. Barre insiste sur l' « évolution des mentalités » des Français entre 1976 et 1987.*

(c-63) *« Nous sommes entrés, souligne-t-il, dans la phase finale de* ***l'intermède institutionnel que nous vivons depuis mars 1986****. (...) Conformément à ce que j'avais annoncé, je n'ai rien fait qui put empêcher le déroulement de* ***cette expérience****.*

Contrainte n°5 : Si la reprise est une DRID dont la source d'inférence est la base des connaissances du monde, et si l'antécédent est un nom commun, utiliser le démonstratif.

8.3.2.3 Inférences sur les connaissances encyclopédiques quand l'antécédent est un nom propre

Dans cette section, nous distinguons plusieurs cas : les entités sont désignées par leur type, par leur fonction, et les ensembles d'individus ont des possibilités d'être repris différemment. Ensuite, nous étudierons les cas de reprises utilisant un élément mentionnel ou simplifiant des antécédents, et enfin, des reprises par une caractéristique ne permettant pas d'identifier uniquement l'antécédent.

Désignation des entités par leur type Quand le nom tête donne le type de l'entité désignée par un nom propre en première mention, la reprise est faite la plupart du temps par un défini. Sur les 153 syntagmes nominaux définis ayant pour antécédent un nom propre et n'apportant pas d'information, 76 dénotent le type de l'entité à laquelle ils réfèrent (ce qui représente 50% des cas étudiés). Il s'agit majoritairement de noms référant à des lieux (cf. exemples suivants), mais on trouve aussi des humains et des institutions.

(c-64) *Plus de 80% des pluies que reçoit* ***l'Inde*** *se concentrent pendant la saison humide (juin-juillet à septembre-octobre). Or, cette année, le déficit est de 20% à 90% dans plus des deux tiers* ***du pays****.*

(c-65) *Une quinzaine de navires de commerce ont été touchés, amenant notamment les Japonais à suspendre momentanément tout trafic de leurs pétroliers dans* ***le Golfe****. L'affirmation, toutefois, du capitaine d'un navire espagnol croisant dans* ***la région*** *selon laquelle un pétrolier saoudien aurait été coulé dans la nuit de jeudi à vendredi n'a été confirmée par aucune source maritime.*

Contrainte n°6 : Si la reprise est une DRID dont la source d'inférence est la base des connaissances du monde, si l'antécédent est un nom propre et si la reprise dénote le type de l'antécédent, utiliser le défini.

Non unicité et reprise par le type de l'entité On trouve dans le corpus 12 cas de reprises démonstratives de l'antécédent par son type. Sur ces 12 cas, 9 sont des cas où il y a plusieurs entités du même type dans le contexte. En cas de non unicité du type dans le contexte on utilise donc le démonstratif,ce qui est conforme à la théorie de Corblin et à la notion d'opposition interne à la classe.

(c-66) *S'ils ne s'approchent pas de notre frontière et si les forces impérialistes se retirent* ***du Tchad****, la Libye promet, pour sa part, de ne pas intervenir dans* ***ce pays****.*

Contrainte n°7 : Si la reprise est une DRID dont la source d'inférence est la base des connaissances du monde, et si l'antécédent est un nom propre, si la reprise dénote le type de l'antécédent et si plusieurs objets du contexte appartiennent à ce type, utiliser le démonstratif.

Reprise par la fonction Si le nom tête donne la fonction ou le métier, de l'individu désigné par un nom propre en première mention, on utilise le défini. Cette hypothèse est à vérifier, mais il semble de façon plus générale que si le nom tête dénote une propriété qui permet d'inférer rapidement (par relation lexicale par exemple) le type de l'entité, on utilise le défini. (Raymond Barre = ancien premier ministre, hyperonyme = homme). Cette remarque vaut aussi pour les villes désignées dans les mentions suivantes par leur statut dans le pays (capitale, capitale régionale...). Les cas de reprise des entités par le nom de leur fonction sont 58 sur 153, soit près de 38% des cas de reprise d'un nom propre sans ajout d'information. Nous pouvons lier ce phénomène avec la proportion importante d'anaphores associatives fonctionnelles mentionnées dans le chapitre 1.2.4. En effet, la fonction ou le métier d'un individu semble être un élément permettant facilement d'identifier uniquement le référent d'une expression référentielle.

(c-67) ***M. Raymond Barre****, en lançant son « appel d' Hourtin », vient donc de donner un premier et sérieux coup d'accélérateur sur la route de l'Elysée.* ***L'ancien premier ministre*** *n'avait, à vrai dire, plus beaucoup le choix.*

Contrainte n°8 : Si la reprise est une DRID dont la source d'inférence est la base des connaissances du monde, si l'antécédent est un nom propre et si la reprise dénote la profession du référent de l'antécédent, utiliser le défini.

Le syntagme désigne un ensemble Pour la désignation de groupes par un nom propre, on peut utiliser le défini pour désigner l'ensemble des membre des l'organisation, ce qui équivaut à désigner le référent par le terme *organisation*. Ici, l'ETA = l'organisation séparatiste = les séparatistes. Nous avons quatre cas de ce type dans le corpus, et bien que cela représente une faible proportion, cette possibilité nous semble intéressante et à rapprocher de notre étude sur l'anaphore associative, puisque même s'il ne s'agit pas ici de désigner un membre ou une partie des membres d'un ensemble, on désigne la totalité des membres d'un ensemble, ce qui n'est pas un processus éloigné.

(c-68) *Depuis le debut de l' année,* ***l'ETA*** *a successivement perdu ses trois groupes les plus actifs : avant le « commando Barcelone » étaient tombés, en janvier dernier, le « commando Madrid » et, il y a deux mois, le « commando Donosti » (opérant à Saint-Sébastien). La satisfaction du ministère de l'interieur, après le succès policier de samedi, a toutefois été tempérée par une mauvaise nouvelle venue le même jour du*

*Pays basque : dans un communiqué publié par le quotidien Egin, édité près de Saint-Sébastien et porte-parole habituel des **séparatistes**, l'ETA militaire a condamné les tentatives de « dialogue » entre certains de ses dirigeants en exil et le gouvernement de Madrid.*

Contrainte n°9 : Si la reprise est une DRID dont la source d'inférence est la base des connaissances du monde, si l'antécédent est un nom propre désignant un ensemble et si la reprise désigne tous les membres de cet ensemble, utiliser le défini.

Simplification de l'antécédent Si la description reprend des antécédents multiples par leur type on utilise le démonstratif. Cette possibilité est représentée par 7 cas sur 19 cas de reprise d'un nom propre par un syntagme démonstratif sans ajout d'information dans notre corpus, soit près de 30% des démonstratifs utilisés de cette façon.

(c-69) *Au **Japon, en Grande-Bretagne et en Allemagne fédérale**, les agrégats monétaires sont en pleine explosion. De juin 1986 à juin 1987, la masse monétaire au sens large a augmenté respectivement, dans **ces pays**, de 10,8%, de 19,2% et de 7,4%.*

Contrainte n°10 : Si la reprise est une DRID dont la source d'inférence est la base des connaissances du monde, et si l'antécédent est réalisé par plusieurs SN, utiliser le démonstratif.

8.3.3 Descriptions Ajoutant de l'Information Nouvelle

Nous montrerons dans cette section que la forme syntaxique et le contenu sémantique des modifieurs jouent un rôle dans le choix de déterminant. Enfin, nous étudierons les cas d'ajout d'information en séparant les cas où l'antécédent est un nom commun, des cas où l'antécédent est un nom propre, lorsque l'ajout d'information est réparti dans la tête du syntagme et ses éventuels modifieurs.

8.3.3.1 Dans les modifieurs

Instanciation d'attributs Si les modifieurs instancient un attribut de l'entité, sous forme de relative, de syntagme prépositionnel, d'adjectif relationnel ou de participe on utilise le démonstratif. Nous utilisons ici la même définition d'attribut que dans le chapitre sur les anaphores associatives. Pour mémoire, un attribut est une propriété de l'entité qui peut prendre une valeur. Cette propriété est une propriété que possèdent tous les objets de la catégorie désignée par le nom tête du syntagme, mais peut prendre des valeurs différentes d'un objet appartenant à la catégorie à un autre objet de la catégorie. Dans les exemples suivants, il aurait été possible de reprendre *Ficofrance* par *le groupe*, *les pires incendies* par *la catastrophe*, c'est à dire avec le défini, sans les modifieurs. En revanche, il ne semble pas possible de reprendre *les masques* par *les objets* pour des raisons de focalisation. L'exemple c-70 illustre le cas d'utilisation d'un adjectif relationnel pour instancier un

attribut (*groupe familial* par opposition à *groupe non familial*, on considère ici que l'attribut prend une valeur booléenne) ; l'exemple c-71 illustre l'utilisation d'une relative ; on observe comment l'instanciation d'un attribut par un adjectif fonctionne dans l'exemple c-72, ainsi que l'utilisation d'une expression figée dans un syntagme prépositionnel dans l'exemple c-73. Pour terminer nous voyons l'instanciation de deux attributs dans des modifieurs coordonnés dans l'exemple c-74. Ajoutons que dans tous les exemples que nous citons, le nom tête du syntagme anaphorique est identique ou plus générique que celui du syntagme antécédent. Nous n'avons en fait que peu de cas d'ajout d'information par hyponymie et modifieurs en même temps.

(c-70) *Parallèlement, il prendrait la présidence de* ***Ficofrance, la société financière de GMF***. ***Ce groupe familial, qui a realisé en 1986 un chiffre d'affaires de 5 milliards de francs****, figure parmi les candidats les plus serieux à l'appel d'offres lancé par Albin Chalandon pour la construction de 15000 places de prison supplémentaires d'ici à 1990.*

(c-71) *Depuis dimanche 30 août, la Californie, devastée par* ***les pires incendies de forêt de son histoire****, a vu la moitié de son territoire placé en état d' urgence. Cinq autres Etats de l'Ouest américain sont également menacés à des degrés divers (Idaho, Arizona, Montana, Wyoming et Etat de Washington) par* ***cette catastrophe, qui a déjà ravagé 200000 hectares de forets***

(c-72) *On ne déplace pas* ***les masques*** *de leur terroir, sauf cas de force majeure, par exemple pour sauver la nation en péril. Il convient de respecter les interdits et les tabous autour de* ***ces objets sacrés*** *.*

(c-73) *Et pourtant ! Totalement inconnu il y a deux ans,* ***le nouveau patron de Technip*** *a su en remontrer aux plus endurcis. Si la première société française d' ingénierie a survécu à la crise, elle ressort de l'épreuve totalement transformée au terme d'un traitement de cheval aussi violent qu'inhabituel dans les mœurs industrielles francaises. « J'ai pris une entreprise en état de choc. J' ai fait une opération sans anesthésie dans un corps prêt à être opéré. Mais aujourd'hui la convalescence est faite », reconnait sans détours* ***ce nouveau PDG de choc****.*

(c-74) *Apparemment oui, et* ***le courtier international Enskilda*** *est de ceux-là. Ses experts escomptent, d' ici à la fin décembre, une progression du marché parisien, et ils s'intéressent aux valeurs financières très convoitées par les investisseurs anglo-saxons.* ***Ce courtier suédois, opérant principalement à Londres pour l'instant****, a également jeté son dévolu sur les secteurs des biens d'équipement.*

Le même phénomène est constatable pour les adjectifs de jugement :

(c-75) *Cela a permis en realite au RPR d'exercer une véritable main-mise et de nommer a leur tête des hommes à lui, sans que* ***les nouveaux actionnaires*** *aient eu leur mot à dire. Si nous gagnons les élections, nous remettrons en cause ces blocs de contrôle. Ce qu'une loi a fait, une autre loi peut le défaire. J' ai d'ailleurs d'ores et déjà demandé a mes collaborateurs d'étudier les problèmes juridiques que cela va poser. Aujourd'hui, on se moque de* ***ce fameux actionnariat populaire****.*

Dans notre corpus, nous trouvons seulement deux exemples de définis correspondant à la description que nous venons de donner, et pourtant impossibles à utiliser avec le démonstratif. Nous donnons l'extrait en c-76. Nous expliquons l'utilisation du défini avec les noms « carnet » et « stylo » par le fait que leur antécédent soit un syntagme nominal préfixé par un possessif. Le possessif les rend uniquement identifiable comme étant les objets du locuteur, et ils deviennent des attributs du locuteur. On ne peut alors plus les désigner par un syntagme démonstratif, de la même façon qu'il est difficile de désigner des parties d'objets par un démonstratif.

(c-76) *Je sors* ***mon carnet***, ***mon stylo***, *me prépare à noter l'inscription exacte qu'on peut lire sur la façade du bâtiment. Une voix retentit, menaçante : « on n'écrit pas. » Je reste un moment avec* ***le stylo suspendu*** *au-dessus* ***du carnet ouvert***.

Contrainte n°11 : Si la reprise est une DAIN et si l'information nouvelle instancie un attribut du référent dans les modifieurs uniquement, utiliser le démonstratif.

Reprise par le nom de rôle thématique et ajout d'information Nous n'avons que deux cas de ce type dans le corpus reproduits en c-78 et c-77. Il est donc difficile d'en tirer des conclusions générales, mais le fait que ces deux seuls cas soient utilisés avec le démonstratif et qu'il soit difficile d'après nous d'utiliser le défini de façon équivalente va selon nous dans le sens de nos observations : certaines informations ajoutées, qui sont des instanciations d'attribut ne permettent pas l'utilisation du défini. Par ailleurs, il s'agit dans le premier cas d'une simplification de l'antécédent qui implique l'utilisation du démonstratif.

(c-77) *Certes, lors du dernier « Club de la presse » d'Europe1, il a admis* ***que la hausse des prix à la consommation atteindra « un peu plus de 3 % » en 1987 (contre 2,1% en 1986 et un objectif initial de 2,4% cette année)***. *Mais le ministre d'Etat a souligné* ***qu'au cours des cinq derniers mois la France a réussi à diminuer de plus d'un point son écart d'inflation avec l'Allemagne fédérale (il est passé de 3,9% à 2,7%)***. ***Ces propos apaisants*** *vont-ils suffire à desarmer les craintes manifestées par de nombreux observateurs - pas seulement en France - sur les risques d' un regain des tensions inflationnistes ?*

(c-78) *Cette ville de province, cernée par la montagne, sans passé prestigieux et peuplée pour l'essentiel de gens venus d'ailleurs, a besoin de* ***grandeur***. *Seule, parmi les quelques villes françaises (Montpellier, Rennes, Toulouse...) qui prétendent lui faire concurrence, à ne pas avoir le rang de métropole régionale, elle se prête volontiers elle-même le titre de capitale : capitale des Alpes, de la houille blanche ou de la matière grise, mais jamais rien de moins. Lors des élections municipales de 1983, le jeune leader de la droite locale, Alain Carignon, avait saisi, d' instinct,* ***cette aspiration commune***.

Apposition Si l'antécédent est focalisé et si l'apport d'information se fait par l'intermédiaire d'une apposition on utilise le défini, comme l'illustrent les exemples c-79 et c-80.

(c-79) ***Deux complices des deux malfaiteurs qui, le 1er septembre, avaient pris en otage six personnes après l'attaque à main armée d'une agence bancaire à Alençon (Orne)*** *ont été inculpés de complicité et d'association de malfaiteurs, et écroués le vendredi 4 septembre.* ***Les deux hommes, Michel Maison, vingt-huit ans, et Robert Dubray, quarante-cinq ans, interpellés mercredi****, ont admis avoir loué un véhicule utilisé pendant l'attaque.*

(c-80) *Pour parachever cet itinéraire breton, on aurait pu imaginer* ***le musée de Quimper*** *nous sortant (pourquoi pas ?) Emile Bernard. Mais c'est Rohner qui s'y déploie dans toutes ses prétentions. On peut tout de même y faire un saut.* ***Le lieu, une aile de la mairie ayant depuis 1976 fait l'objet d'aménagements sérieux****, est agréable*

Contrainte n°12 : Si la reprise est une DAIN et si l'information nouvelle est réalisée dans des modifieurs apposés, utiliser le défini.

8.3.3.2 Apport d'information dans tout le syntagme

Quand l'antécédent est un nom commun De façon générale, on utilise le démonstratif qui grâce à sa capacité reconnue pour reclassifier les référents, permet de forcer le lien entre l'antécédent et l'anaphore (exemples c-81 et c-82). Les cas où le syntagme nominal complet apporte de l'information sur l'antécédent et où il est déterminé par un défini dans notre corpus sont des cas où la description est complètement identifiante, bien qu'apportant de l'information (exemples c-83), ou alors un cas limite entre l'apport d'information et la production d'une anaphore sur la base des connaissances encyclopédiques (exemple c-84).

(c-81) *Mais à Roubaix où la résistance s'organise,* ***le personnel*** *a l' impression de seulement compter les points. La Lainière va peut-être supprimer des cars de ramassage ! Pour* ***ces ouvrières du bassin houillier dont quelques-unes ont déjà trois heures de transport par jour****, la nouvelle pour l'instant simple rumeur a relegué au second plan les manœuvres boursières dont leur entreprise fait l'objet depuis deux mois.*

(c-82) *L'exposition retrospective de Morlaix, où est mort* ***le peintre*** *en 1927, il y a soixante ans, n'est donc pas inutile, d'autant qu'elle réunit beaucoup d'œuvres de collections privées susceptibles de révéler la complexité cachée de* ***cet intellectuel impenitent, pris dans l'imbroglio des idées et des sources qui ont fait le symbolisme****.*

(c-83) *Soudain,* ***le groupe*** *s'arrête et escalade la clôture. L'obstacle franchi,* ***les huit hommes*** *parcourent rapidement une dizaine de mètres par la rue Connoly, qui passe sous le pavillon argentin, et débouchent aussitôt devant le pavillon 31 que les Israéliens partagent avec les délégations de Hongkong et de l'Uruguay.*

(c-84) *Mais le dernier coup d'Etat - le cinquième en dix-huit mois et le premier aussi sanglant - attise les craintes des investisseurs qui voient là une nouvelle démons-*

tration de l'instabilité politique du ***pays****. L'étranger avait pourtant recommencé à s'interesser à* ***l'archipel****.*

Le démonstratif permet aussi d'apporter un changement de point de vue, comme dans l'exemple c-85 où le locuteur exprime ce qu'il pense être une erreur de la part des journalistes du *Monde* ou dans l'exemple c-86 où le journaliste rapporte les propos d'un autre et requalifie les événements vécus par cet autre journaliste par le terme *attente*.

(c-85) *Dans votre article du 19 août consacré à la mort de Hess, j' ai relevé à deux reprises la formule* ***« l'incroyable médiocrité du personnel politique nazi »****. En 1939, je partageais* ***ces illusions****.*

(c-86) *A partir de là, commence ce que Jean Lacouture, envoyé special du Monde, appellera* ***« une nuit de sang et de mensonge »****. Au village olympique, où les centaines de journalistes aux aguets sont privés d'informations, les rumeurs les plus diverses circulent. Dans cette atmosphère de fièvre et d'angoisse, on croit même, un moment, au miracle. Lacouture décrit* ***cette attente****.*

Contrainte n°13 : Si la reprise est une DAIN, si l'information nouvelle est réalisée dans tout le syntagme et si l'antécédent est un nom commun, utiliser le démonstratif.

Quand l'antécédent est un nom propre Ici, les mêmes contraintes que pour les DRID s'appliquent. Si la spécification est complète (i.e. si la description rend le référent identifiable de façon unique dans les connaissances du monde), et/ou si le nom tête dénote le type de l'objet, on utilise le défini :

(c-87) *Totalement inconnu il y a deux ans, le nouveau patron de* ***Technip*** *a su en remontrer aux plus endurcis. Si* ***la première société française d'ingénierie*** *a survécu à la crise...*

(c-88) *« Puis* ***M. Barre*** *aborde plus précisement le thème de ces universités d'été centristes, celui de l'ouverture, livrant sa définition d'une société ouverte (...). Mettant en exergue la nécessité pour la France de « refuser l'isolement et le protectionnisme », le député de Lyon en vient à évoquer plus longuement la question des rapports Nord-Sud.*

De la même manière si le nom tête dénote une profession ou une nationalité on utilise le défini dans la mesure où il permet de retrouver le type de l'objet par inférence sur relations lexicales.

(c-89) *La libération de* ***Pierre-André Albertini*** *constituerait un « bon point » pour M. Chirac. Elle placerait en revanche le Parti communiste en porte à faux. Le PC, qui a monté une importante campagne en faveur de celui qu'il appelle « l'otage de l'apartheid », a accusé en effet le gouvernement de ne rien faire pour obtenir la libération du* ***jeune cooperant****.*

(c-90) *Impériale,* ***Jeannie Longo*** *n'a laissé aucune chance, vendredi 4 septembre, à ses adversaires lors des championnats du monde de cyclisme sur route, qu'elle a*

remportés pour la troisième année consecutive. Victorieuse en juillet du Tour de France féminin, puis le mois suivant de la Coors Classic américaine, ***la sportive grenobloise*** *a terminé detachée sur le circuit autrichien de Villach.*

Contrainte n°14 : Si la reprise est une DAIN, si l'information nouvelle est réalisée dans tout le syntagme et permet d'identifier uniquement le référent et si l'antécédent est un nom propre, utiliser le défini.

Contrainte n°15 : Si la reprise est une DAIN, si l'information nouvelle est réalisée dans tout le syntagme et si la reprise dénote le type, la profession ou la nationalité du référent et si l'antécédent est un nom propre, utiliser le défini.

Si la spécification n'est pas complète, si l'instanciation d'attributs se fait sous forme de participe ou de syntagme prépositionnel, où si on introduit des informations subjectives on utilise le démonstratif.

(c-91) *Quant a* ***Klaas de Jonge****, il vivait depuis le 29 juillet 1985 dans les locaux, aujourd'hui desaffectés, de l'ambassade des Pays-Bas à Pretoria. Les négociations avec les autorites de La Haye pour faire sortir* ***ce quinquagénaire accusé d'avoir transporté des armes pour le compte de l'ANC*** *n'avaient rien donné.*

(c-92) *Après Colas-Magne, vainqueur en tandem,* ***Richard Vivien*** *a créé une surprise de taille en devenant samedi champion du monde de la catégorie, que l'on définissait il y a peu de temps comme étant celle des « purs ».* ***Ce Normand de vingt-trois ans, remarqué par Yves Hezard****, est peut-être un pur, mais il possède déjà beaucoup de métier si l'on retient la manière dont il a battu l'Allemand de l'Ouest Bolts, le Danois Pedersen, le Polonais Mierzejewski à l'issue d'une course à l'économie, fatale aux Sovietiques .*

(c-93) *Et si* ***Carl Lewis*** *était condamné à se battre sans cesse contre les chimères du sport moderne ?* ***Ce petit garçon qui avait une mauvaise croissance*** *est devenu adulte, un athlète prodigieusement doué.*

Contrainte n°16 : Si la reprise est une DAIN, si l'information nouvelle est réalisée dans tout le syntagme, et si la reprise ne satisfait pas les contraintes n°14 et 15 et si l'antécédent est un nom propre, utiliser le démonstratif.

8.3.4 Synthèse

Dans cette section, nous synthétisons les résultats détaillés tout au long de ce chapitre sur le contenu informationnel des descriptions définies et démonstratives employées dans des mentions subséquentes. Nous reviendrons tout d'abord sur les critères habituellement utilisés pour expliquer la différence entre le défini et le démonstratif en français. Nous ferons ensuite la liste de toutes les contraintes identifiées sur l'utilisation des déterminants décrites dans la dernière section de ce chapitre.

8.3.4.1 Statut de l'unicité et de la saillance du référent.

L'unicité et la saillance du référent sont généralement les contraintes invoquées pour justifier le choix d'un défini par rapport à un démonstratif, et inversement. Pourtant, l'étude de corpus montre que ces contraintes sont insuffisantes, bien que fondamentales. Nous revenons une dernière fois sur ces contraintes dans cette section, afin d'expliquer leur rôle dans notre étude.

Unicité Le critère d'unicité de la description reste fondamental, même si sa définition n'est pas assez précise. La notion d'opposition notionnelle décrite par [Corblin, 1987] semble en dériver assez directement, avec pour but de l'affiner. Notre point de vue sur l'unicité est le suivant : les reprises coréférentielles impliquent une série de raisonnements sur diverses bases de connaissances, et le problème est de savoir *dans quelle base de connaissances le référent doit-il être unique pour justifier l'emploi d'un démonstratif.* Suite à l'étude de corpus, nous pensons que l'unicité doit être satisfaite dans la base des connaissances du monde, c'est à dire la base la plus large de toutes celles que nous utilisons en génération. Dans des bases de connaissances plus restreintes, il semble que le critère d'unicité n'entre plus en ligne de compte.

Saillance De la même façon, les critères de saillance du référent restent vagues. Pour une discussion complète de la notion et des problèmes qu'elle pose en traitement automatique des langues, nous renvoyons à [Landragin, 2003]. Nous ne la discutons que partiellement ici, avec les paramètres que nous pouvons appréhender en génération suite à notre étude de corpus. En cas de non unicité de l'objet satisfaisant la description, il est clair que le démonstratif seul peut être utilisé, et pour désigner le dernier élément mentionné. Sinon, il faut introduire un élément mentionnel dans le groupe nominal anaphorique. Dans les cas où le référent est unique, mais où il est reclassifié, la fonction grammaticale semble déterminante, et porte le locuteur à choisir le défini si le groupe nominal antécédent est en position sujet. Le problème que posent les définitions de la saillance est donc le suivant : dans certains cas, on donne comme élément pour définir la saillance la récence de la mention, tandis que dans d'autres cas, on utilise la fonction grammaticale de l'antécédent, et dans ce cas le sujet est le plus saillant. Dans une langue comme le français, ces critères sont forcément contradictoires et donc problématiques. Nous n'utiliserons donc pas la notion de saillance dans notre algorithme, mais uniquement la fonction grammaticale de l'antécédent.

8.3.4.2 Liste des contraintes identifiées au cours de l'étude de corpus

Les contraintes que nous avons identifiées au cours des sections précédentes présentent la particularité de combiner des paramètres syntaxiques et sémantiques. Il nous semble intéressant de noter l'importance des données syntaxiques dans le choix du déterminant, données qui ont souvent été oubliées dans les analyses théoriques du déterminant. Nous dressons maintenant la liste de ces contraintes, de façon à pouvoir écrire un algorithme de choix du déterminant.

Contrainte n°1 : Si le référent est unique dans les connaissances du monde, utiliser le défini.
Contrainte n°2 : Si le nom utilisé dans l'anaphore est un nom prédicatif dont les arguments sont réalisés, utiliser le défini. Si c'est un nom prédicatif sans arguments réalisés, utiliser un démonstratif.
Contrainte n°3 : Si la reprise est une DRID dont la source de l'inférence n'est pas la base des connaissances du monde, utiliser le défini si l'antécédent est sujet, utiliser le démonstratif si l'antécédent occupe une autre fonction (que la fonction sujet).
Contrainte n°4 : Si la reprise est une DRID dont la source d'inférence n'est pas la base des connaissances du monde, et si l'antécédent est réalisé par plusieurs SN, utiliser le démonstratif.
Contrainte n°5 : Si la reprise est une DRID dont la source d'inférence est la base des connaissances du monde, et si l'antécédent est un nom commun, utiliser le démonstratif.
Contrainte n°6 : Si la reprise est une DRID dont la source d'inférence est la base des connaissances du monde, et si l'antécédent est un nom propre, et si la reprise dénote le type de l'antécédent, utiliser le défini.
Contrainte n°7 : Si la reprise est une DRID dont la source d'inférence est la base des connaissances du monde, et si l'antécédent est un nom propre, si la reprise dénote le type de l'antécédent, et si plusieurs objets du contexte appartiennent à ce type, utiliser le démonstratif.
Contrainte n°8 : Si la reprise est une DRID dont la source d'inférence est la base des connaissances du monde, et si l'antécédent est un nom propre, et si la reprise dénote la profession du référent de l'antécédent, utiliser le défini.
Contrainte n°9 : Si la reprise est une DRID dont la source d'inférence est la base des connaissances du monde, et si l'antécédent est un nom propre désignant un ensemble, et si la reprise désigne tous les membres de cet ensemble, utiliser le défini.
Contrainte n°10 : Si la reprise est une DRID dont la source d'inférence est la base des connaissances du monde, et si l'antécédent est réalisé par plusieurs SN, utiliser le démonstratif.
Contrainte n°11 : Si la reprise est une DAIN et si l'information nouvelle instancie un attribut du référent dans les modifieurs uniquement, utiliser le démonstratif.
Contrainte n°12 : Si la reprise est une DAIN et si l'information nouvelle est réalisée dans des modifieurs apposés, utiliser le défini.
Contrainte n°13 : Si la reprise est une DAIN, si l'information nouvelle est réalisée dans tout le syntagme, et si l'antécédent est un nom commun, utiliser le démonstratif.
Contrainte n°14 : Si la reprise est une DAIN, si l'information nouvelle est réalisée dans tout le syntagme, et permet d'identifier uniquement le référent et si l'antécédent est un nom propre, utiliser le défini.
Contrainte n°15 : Si la reprise est une DAIN, si l'information nouvelle est réalisée dans tout le syntagme, et si la reprise dénote le type, la profession ou la nationalité du référent et si l'antécédent est un nom propre, utiliser le défini.
Contrainte n°16 : Si la reprise est une DAIN, si l'information nouvelle est réalisée dans tout le syntagme, et si la reprise ne satisfait pas les contraintes n°14 et 15, et si l'antécédent

est un nom propre, utiliser le démonstratif.

8.4 Algorithme de choix du déterminant

Notre algorithme de choix du déterminant se base sur notre extension de l'algorithme de [Gardent et Striegnitz, 2003], dans la mesure où il prend en entrée, la sortie de l'algorithme de détermination du contenu de la description et où il s'appuie sur les mêmes bases de connaissances[13]. Nous présentons tout d'abord les éléments qu'il prend en entrée, puis son déroulement en trois parties.

8.4.1 Entrée de l'algorithme de choix du déterminant :

Contexte

r : Référent cible.

Bases de connaissances :

WKL : connaissances du monde.

DM : modèle de discours.

SM : modèle du locuteur.

LEX : bases de données lexicales, où sont représentées les relations lexicales standard (hyperonymie, hyponymie, synonymie).

Informations syntaxiques

T : arbre syntaxique représentant le $\overline{N}$ (i.e. le syntagme sans déterminant) décrivant r.

F(A) : fonction grammaticale de l'antécédent.

C(A) : catégorie syntaxique de l'antécédent.

Informations sémantiques

ϕ : représentation du contenu sémantique déterminé par l'algorithme de Gardent et Striegnitz.

F(L) : statut informationnel de la reprise. Deux valeurs possibles : DRID ou DAIN.

SI : la source de l'inférence qui a permis de construire ϕ et T à partir de l'antécédent. SI peut prendre plusieurs valeurs : DM (modèle du discours), WKL (connaissances du monde) ou LEX (connaissances lexicales).

8.4.2 Algorithme

L'algorithme est présenté de la façon suivante pour plus de lisibilité. Nous présentons dans un premier temps le corps de l'algorithme principal, avec deux parties qui sont résumées entre < >. Ces parties résumées correspondent aux cas de DRID et aux cas de DAIN, que nous déroulons séparément dans des paragraphes indépendants.

[13]Comme nous réutilisons les mêmes bases de données que Gardent et Striegnitz, nous conservons aussi les abréviations anglaises qu'elles utilisent dans leurs articles pour les désigner.

8.4.2.1 algorithme principal

Le corps de l'algorithme se présente de la manière suivante :

Si `r appartient à DM`
Alors
  Si `r unique dans WKL`
  Alors `DEFINI`
  Sinon
    Si `tête (T) = Nom prédicatif`
    Alors
      Si `arguments`
      Alors `DEFINI`
      Sinon `DEMONSTRATIF`
    Sinon
      Si `<DRID>`
      Sinon `<DAIN>`
Sinon `INDEFINI`

8.4.2.2 Sous - algorithme DRID

L'algorithme pour les reprises ne contenant pas d'information nouvelle sera le suivant :

Si `F(L) = DRID`
Alors
  Si `SI = DM ou SI = LEX :`
  Alors
    Si `Antécédent multiple`
    Alors `DEMONSTRATIF`
    Sinon
      Si `F(A) = sujet`
      Alors `DEFINI`
      Sinon `DEMONSTRATIF`
        Sinon
          Si `SI = WKL`
          Alors
            Si `C(A) = nom commun`
              Alors `DEMONSTRATIF`
              Sinon
                Si `C(A) = nom propre`
                Alors
                  Si ϕ `= type de l'entité ou profession de l'individu`
                  Alors `DEFINI`
                  Sinon `DEMONSTRATIF`
Sinon `<DAIN>`

8.4.2.3 Sous - algorithme DAIN

Enfin, voici la forme que prennent sous forme algorithmique les contraintes identifiées pour les reprises ajoutant de l'information sur le référent.

Si F(L) = DAIN
  Si Si ajout d'information dans les modifieurs de T
  Alors
    Si modifieurs de T = apposition
    Alors DEFINI
    Sinon
      Si modifieurs = attributs de r
      Alors DEMONSTRATIF
      Sinon DEFINI
  Sinon
    Si ajout d'information dans tout T entier
    Alors
      Si C(A) = nom commun
      Alors DEMONSTRATIF
      Sinon
        Si C(A) = nom propre
        Alors
          Si ϕ = profession de l'individu
          Alors DEFINI
          Sinon DEMONSTRATIF

Chapitre 9

Conclusions et Perspectives

Pour conclure, rappelons tout d'abord les objectifs que nous nous étions fixés au début de notre thèse :

Nous souhaitions parvenir à une étude assez fine des descriptions définies et démonstratives, afin d'exhiber des contraintes utiles pour leur génération automatique. Nous avons tenu compte des travaux réalisés antérieurement, et nous avons proposé trois directions pour étendre les algorithmes existants :

- pour le traitement de tous les types d'anaphore associative,
- pour le traitement des reprises définies et démonstratives ajoutant de l'information sur le référent,
- pour intégrer aux algorithmes de génération la possibilité de choix entre le déterminant défini et le démonstratif.

Nous avons souhaité fonder notre étude sur une analyse de corpus, pour éviter les biais intrinsèques à l'analyse des phénomènes par introspection. Nous avons donc utilisé une analyse semi-automatique de corpus pour mener notre étude et atteindre nos objectifs, et avons montré que l'étude empirique des phénomènes sur un ensemble large de données devient rapidement nécessaire pour appréhender globalement un phénomène linguistique. En effet, nous avons pu recenser un grand nombre d'occurrences du phénomène étudié, sans avoir à nous poser le problème de l'acceptabilité de ces occurrences. De plus, même si parfois les déterminants sont substituables l'un à l'autre, une analyse de corpus permet de dégager des tendances sur ce que les locuteurs font spontanément, ou tout au moins, sans *a priori* théorique sur la formulation de leurs idées.

Bien que notre étude de corpus n'ait pu être réalisée totalement dans les règles de l'art, bien que nous n'ayons pu pour des raisons techniques, extraire tous les résultats que nous aurions pu espérer - particulièrement les résultats concernant les chaînes anaphoriques, de nombreux éléments nouveaux se sont fait jour au terme de notre analyse.

La thèse présentée a donc atteint les objectifs que nous nous étions fixés, et nous développons une dernière fois nos conclusions avant de définir des perspectives pour nos recherches futures.

Dans un premier temps, nous avons réalisé une première étude de corpus confirmant un grand nombre de données théoriques sur l'utilisation des descriptions définies et démonstratives. Nous avons alors pu montrer en quoi ces données étaient insuffisantes dans une perspective de formalisation de la production d'expressions référentielles. Dans le deuxième volet de notre travail qui a consisté en une ré-annotation du corpus et en une nouvelle classification des phénomènes référentiels impliquant des descriptions définies et démonstratives, nous avons montré que certains paramètres, jusqu'à présent peu pris en compte dans les études sur le même sujet, étaient observables directement sur les données et formalisables. Notre ré-annotation s'est déroulée en deux phases : une ré-annotation des anaphores associatives, et une ré-annotation des description définies et démonstratives coréférentielles.

Notre étude plus approfondie des anaphores associatives a montré que la série de relations entre l'antécédent et l'anaphore étudiées jusqu'à présent étaient soient trop vagues, soit trop restreintes. En recherchant les sources permettant de construire la relation entre un antécédent et une anaphore, nous avons établi une classification générique du phénomène. En effet, ce qui est important pour générer une anaphore associative n'est pas tant la nature de la relation entre les deux référents, mais bien la source de l'inférence permettant de la produire et de l'interpréter. Aussi, cette façon d'étudier le phénomène le rend possible à appréhender dans le cadre d'un algorithme de génération. Nous avons ensuite montré comment utiliser les outils de TAL existant dans un algorithme de génération des descriptions définies pour générer l'intégralité des anaphores associatives.

L'étude des reprises définies et démonstratives s'est déroulée en deux temps : dans un premier temps, nous avons établi une classification de ces descriptions en nous basant sur leur contenu sémantique. Dans un deuxième temps, nous avons recherché des contraintes justifiant le choix du défini ou du démonstratif, en nous appuyant sur les données théoriques connues, et sur notre classification.

Dans cette partie de notre étude de corpus, nous avons présenté une classification des expressions coréférentielles de deux points de vue : le premier point de vue est un point de vue informationnel, qui nous permet d'aborder la capacité d'ajout d'information des reprises coréférentielles en génération. Il s'agit alors de déterminer la fonction du groupe nominal : apporte-t-il ou non de l'information sur le référent ? Ensuite, il nous a été nécessaire d'identifier la provenance des inférences permettant de faire le lien entre les deux groupes nominaux, toujours dans le but de permettre une génération plus complète des phénomènes. L'étude de corpus basée sur cette classification a permis d'établir une extension aux algorithmes de génération, en y intégrant la possibilité de générer des reprises n'apportant pas d'information sur le référent, mais dont le contenu est inféré de toutes les bases de connaissances disponibles dans le générateur. Nous y avons aussi intégré la possibilité de générer des reprises ajoutant de l'information sur le référent, et nous avons pu observer que la réalisation linguistique de ce type de reprise passe essentiellement par des reprises dont le nom tête n'entretient aucune relation sémantique avec son antécédent.

Enfin, nous avons repris notre étude de corpus pour étudier plus finement les contextes

d'apparition des déterminants définis et démonstratifs. Nous avons montré que les contraintes impliquant la génération de l'un ou l'autre des deux déterminants sont variées : il semble que la composition et la structuration du contexte entre en jeu, autant que le contenu sémantique de la reprise et sa capacité à apporter de l'information nouvelle sur le référent. Nous avons aussi montré que le contenu sémantique des modifieurs était important. Le résultat le plus surprenant pour nous a été de montrer que des contraintes syntaxiques interviennent elles aussi pour le choix du déterminant : il semble en effet que la fonction grammaticale de l'antécédent, sa forme syntaxique, la nature grammaticale des modifieurs de la reprise ont une grande influence sur le choix du déterminant. Nous avons alors établi une liste de seize contraintes, et construit un algorithme de choix du déterminant.

Bien entendu, il reste une série de problèmes à résoudre, et nous nous proposons à court et moyen terme de continuer nos recherches dans plusieurs axes :

Tout d'abord, nous souhaitons poursuivre le travail entamé dans le corpus PAROLE de plusieurs façons. Nous voudrions dans un premier temps faire ré-annoter la totalité du corpus par deux annotateurs « naïfs » afin de confirmer les résultats obtenus lors de notre propre annotation. Ceci serait à notre avis le seul moyen de valider statistiquement nos résultats. Par ailleurs, afin de mieux identifier les inférences qu'ils produisent en annotant le texte, nous souhaitons, pour tous les cas où les connaissances encyclopédiques sont impliquées, leur demander de décrire le raisonnement qu'il font pour relier l'antécédent et l'anaphore. Ceci serait intéressant pour connaître précisément les éléments qui permettent de faire des inférences, et surtout de voir le nombre de pas nécessaires dans le raisonnement qui permet d'établir la coréférence. Nous avions commencé cette expérience, qui n'a pu être menée à son terme faute de temps, mais pensons pouvoir la reprendre.

Sur les anaphores associatives, deux perspectives s'ouvrent à nous : tout d'abord, nous souhaiterions faire une comparaison avec d'autres langues. Le corpus NEGRA (un corpus en allemand) est actuellement annoté avec notre classification des anaphores associatives, et il serait intéressant de vérifier que les relations sémantiques identifiées dans notre étude de corpus se retrouvent dans d'autres langues.

Par ailleurs, des études de l'anaphore associative sont actuellement menées à partir de données statistiques, et de recherches d'antécédent par apprentissage [Markert et al., 2003]. Il serait intéressant de voir si des analyses statistiques permettent de retrouver les anaphores associatives impliquant les relations que nous avons identifiées, et si elles ne le peuvent pas, quelles relations sont ignorées. Nous pouvons aussi envisager une recherche orientée par les relations que nous avons identifiées.

Concernant les reprises définies et démonstratives, nous souhaitons orienter nos recherches dans trois directions :

Nous souhaitons réaliser un programme qui permettrait d'extraire automatiquement les chaînes de référence afin d'obtenir des données plus précises sur les mentions subséquentes des référents. Nous demeurons en effet persuadée que la place dans une chaîne anaphorique

longue influe sur la façon de désigner les référents, sur la possibilité d'ajouter de l'information à leur sujet, et sur les déterminants employés. Il nous faudrait alors prendre en compte les reprises pronominales qui interviennent entre les descriptions nominales faisant partie de la chaîne de référence.

Nous souhaitons aussi réaliser une nouvelle annotation sur un type de textes différent (roman, texte juridique ou scientifique), afin de voir si notre étude n'a pas été biaisée par l'analyse de textes journalistiques. Bien que les auteurs des textes soient différents, bien que les sujets abordés soient variés, nous avons conscience que des contraintes de style pèsent sur la rédaction des textes journalistiques, et particulièrement sur les textes publiés dans *Le Monde*. Il serait intéressant alors de mener la même étude sur d'autres journaux, sur des textes littéraires, et éventuellement sur des transcriptions de dialogue oraux, qui ne manipulent pas les expressions référentielles de la même façon que les textes écrits.

A plus long terme, nous envisageons de relier nos travaux aux travaux sur la structuration des discours. En effet, une des tâches de la génération de texte est la planification et la structuration de document. Nous ne pouvons pas nier l'influence des relations de discours sur la production d'expressions référentielles, aussi, nous souhaitons intégrer nos travaux dans des recherches plus larges sur la génération de textes et sur l'organisation du discours.

Enfin, nous avons pu établir une série de contraintes intervenant dans le choix des déterminants. Cependant, même si nous avons réussi à ordonner ces contraintes, et à définir des priorités entre elles, il nous semble intéressant de pouvoir prolonger notre étude, afin de vérifier si cet ordonnancement est valable. On peut par exemple imaginer d'établir des contraintes sur le principe de la théorie de l'optimalité, en définissant des priorités absolues de certaines contraintes sur d'autres [Beaver, 2002]. Ceci nécessiterait alors une étude sur des données encore plus nombreuses, et une annotation de corpus encore plus précise, apportant une réponse pour chaque contrainte définie dans notre étude.

Annexe A

Premiers manuels d'annotation

Les deux manuels d'annotation que nous présentons dans cette annexe ont été rédigés pour la première annotation du corpus (présentée au chapitre 5). Le premier porte sur l'annotation des descriptions définies, le second sur l'annotation des descriptions démonstratives. Il était destiné à l'annotation de notre corpus (dont nous étions l'unique annotatrice) et d'un autre corpus, en allemand, annoté à Sarrebrück sous la directions de Kristina Striegnitz, avec qui nous l'avons rédigé. Il est basé sur l'état de l'art que nous avons présenté dans le premier chapitre de notre thèse.

Le manuel s'organise de la façon suivante : dans un premier temps, nous décrivons la démarche générale que doit faire l'annotateur. Ensuite, nous donnons une série de définitions et d'exemples pour déterminer précisément les relations anaphoriques entretenues entre les syntagmes à annoter.

A.1 Le manuel d'annotation pour les définis

A.1.1 Déroulement des actions

Les étapes présentées dans ce paragraphe sont résumées dans l'arbre de décision présenté figure A.1.

L'annotation se déroule en quatre grandes étapes :

Étape 1. Lorsqu'on annote un syntagme nominal, on doit répondre à la question suivante : Puis-je trouver un antécédent à ce syntagme nominal, qui soit un syntagme nominal ni elliptique ni pronominal, ou un verbe sans ses compléments ?

Étape 2. Si la réponse est non, on annote le syntagme comme une première mention, on clique sur « free » dans la rubrique « type », et l'annotation est terminée. Si la réponse est oui, on doit trouver le type de relation qu'il entretient avec l'antécédent et choisir entre « bridging » (anaphore associative), « direct » et « indirect » (coréférent) (comme indiqué dans le paragraphe étape 3) .

Étape 3. Les syntagmes sont-ils strictement coréférents ou non (réfèrent-ils au même objet) ?
Si oui, on coche « direct » ou « indirect » dans la catégorie « type ». Le moyen de choisir entre les deux est expliqué dans les sections suivantes.
Si non, on les annote comme anaphores associative et on coche « bridging » dans la catégorie « type ».
On doit alors choisir le sous-type de relation anaphorique (Etape 4).

Étape 4. On choisit alors le type de relation entre l'antécédent et le syntagme annoté. La relation « coréférence directe » n'a pas à être sous-typée, mais on cliquera sur « coref » dans la rubrique « relation » (les raisons sont purement techniques). Ensuite, pour la relation indirecte, on devra choisir entre les relations « syn » « hyp » « theta » « other », explicitées dans les sections suivantes. De même, pour les anaphores associatives (type = « bridging »), on se reportera aux sections suivantes pour choisir le sous-tyê de relation entre « has », « theta » et « other ».

A.2 Comment choisir le type de relation entre les syntagmes ?

A.2.1 coréférence directe (Point n°5 dans l'arbre)

Les deux syntagmes nominaux réfèrent exactement à la même entité. L'antécédent est un syntagme avec la même tête nominale que le syntagme anaphorique. La présence de modifieurs différents dans les deux syntagmes n'est pas à prendre en compte.

A.2.2 coréférence indirecte (Point 6 dans l'arbre)

Les deux syntagmes (antécédent et anaphore) réfèrent au même objet. L'antécédent et l'anaphore n'ont pas la même tête nominale, comme dans les exemples suivants :

(105) *Un chien - Le chihuahua*

(106) *Un policier - Le flic*

(107) *Un chien - L'animal*

(108) *Bill - le vendeur*

Les trois premiers exemples illustrent des cas de coréférence indirecte par des relations lexicales connues, mais parfois, comme dans le dernier exemple, la relation entre les syntagmes nominaux est possible à saisir grâce à un verbe. Pour plus de clarté nous l'illustrons dans l'exemple suivant :

(109) ***Jack*** *a vendu un livre à Bill.* ***Le vendeur*** *était content.*

On ignorera les cas où l'antécédent n'est pas un syntagme nominal ou un verbe sans ses compléments.

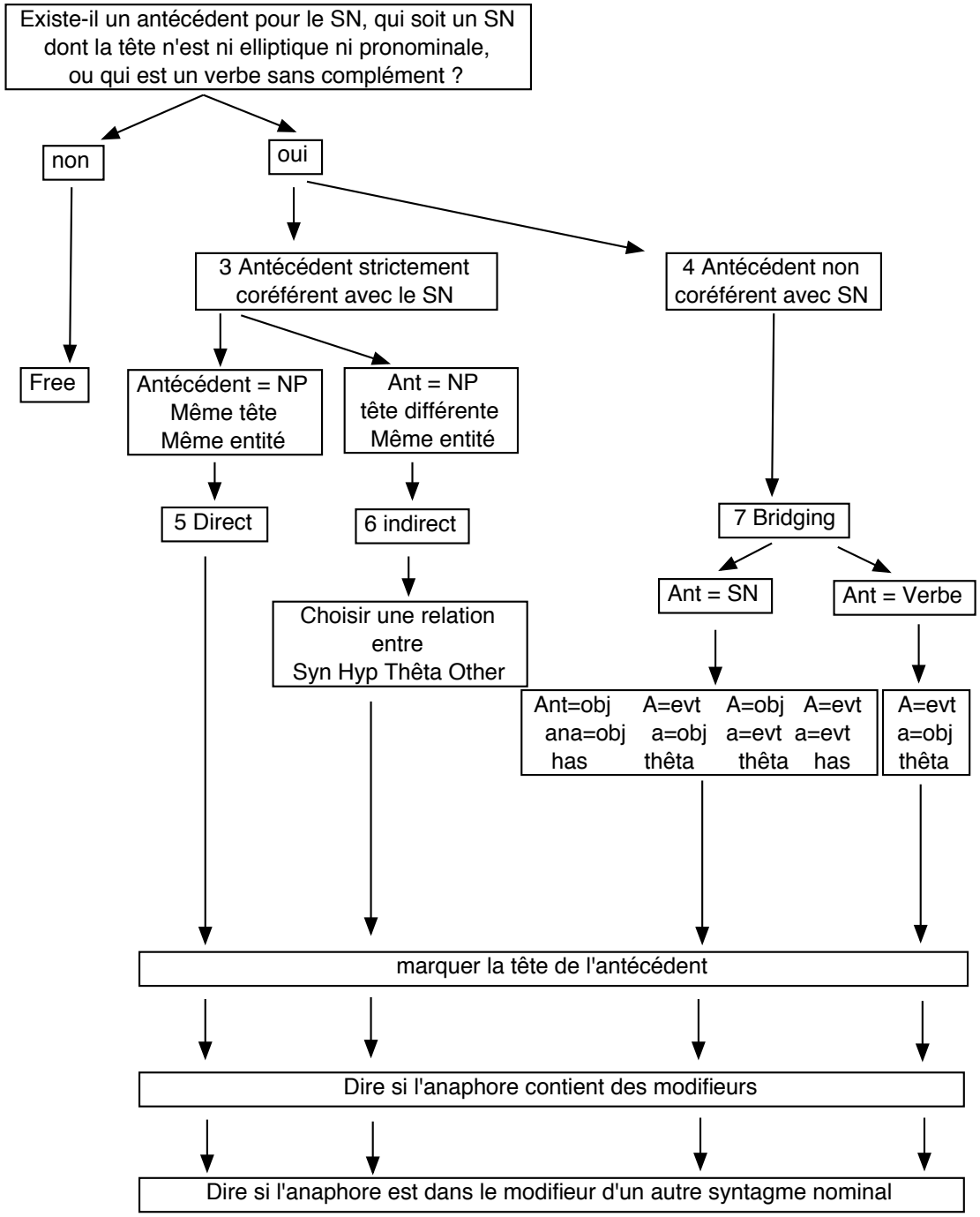

FIG. A.1 – Arbre de décision - schéma d'annotation

A.2.3 Bridging (Point n° 7 dans l'arbre de décision)

On parle dans ce cas d'anaphore associative. Les deux syntagmes nominaux ou le syntagme nominal et le verbe réfèrent à des entités différentes reliées par des savoirs encyclopédiques. L'entité considérée n'a pas encore été mentionnée dans le discours. On trouve différentes configurations pour cette possibilité, énumérées dans les paragraphes suivants :

A.2.3.1 Le syntagme anaphorique réfère à un objet et l'antécédent est un syntagme référant à un objet

Les exemples suivants entrent dans cette catégorie :

(110) *Je suis entré dans* ***la maison***. ***La porte*** *était ouverte.*

(111) *J'ai rencontré* ***Marie***. *Elle avait une blessure sur* ***le nez***

(112) *Je dois réparer* ***ma voiture.*** ***Le moteur*** *est cassé.*

(113) ***Le club de football de Bastia*** *a perdu la finale.* ***Le président*** *a dit que c'était la faute des supporters.*

(114) ***Cette famille*** *est étrange.* ***Les parents*** *semblent fous.*

(115) ***Un couple*** *entra dans la pièce.* ***L'homme/la femme*** *portait un chapeau.*

(116) *Je suis entré dans* ***un village***. ***L'église*** *était sur une colline.*

(117) *J'entrai dans* ***la cuisine***. ***Le réfrigérateur*** *était ouvert.*

(118) *J'aime* ***cette chemise***. ***La matière*** *est douce.*

(119) *C'est* ***une valise*** *très chère.* ***Le cuir*** *est très beau.*

(120) *Jean a fait* ***un gâteau***. *Marie a mangé* ***la dernière part*** *et a été malade.*

(121) ***La classe*** *prépare un spectacle.* ***Les filles*** *vont chanter.*

(122) *J'aime* ***les chiens***. *Mes préférés sont* ***les épagneuls***.

A.2.3.2 L'antécédent dénote un événement et le syntagme anaphorique dénote un objet

L'événement peut être dénoté par un syntagme nominal ou par un verbe.

(123) ***La construction*** *a duré deux ans.* ***Les constructeurs*** *n'avaient pas assez d'argent pour payer plus d'ouvriers.*

(124) *J'ai aimé* ***cet opéra***. ***La chanteuse*** *m'a impressionné.*

(125) ***Un crime*** *a été commis.* ***Le meurtrier*** *court toujours.*

(126) *Quelqu'un* ***a acheté*** *ce livre.* ***L'acheteur*** *sera mécontent.*

(127) *J'ai commencé à* ***beurrer*** *la tartine.* ***La margarine*** *était périmée.*

A.2.3.3 L'antécédent réfère à un objet et l'anaphore à un événement

(128) ***Cette maison*** *est belle.* ***La construction*** *a duré deux ans.*

A.2.3.4 Le syntagme anaphorique réfère à un événement et l'antécédent à un événement

(129) *J'aime* ***cet opéra****.* ***Le duo final*** *est merveilleux.*

A.2.3.5 Le syntagme anaphorique réfère à un objet, l'antécédent est un verbe

(130) *Jean* ***a été assassiné*** *hier.* ***Le meurtrier*** *a été arrêté.*

A.2.3.6 Le syntagme anaphorique dénote un événement et l'antécédent est un verbe

On n'annotera pas ces cas de figure, qui ne seront pas des anaphores associatives mais des cas de coréférence événementielle, très complexes à annoter.

A.3 Comment choisir le sous-type de relation entre les syntagmes ?

A.3.1 Pour la coréférence indirecte

On choisira une relation entre les suivantes, que nous définissons dans les sections suivantes. Le choix s'effectue en cochant l'une des possibilités dans la rubrique « relation ».

- « hyp » pour hyponymie ou hyperonymie
- « thêta » pour reprise par le nom de rôle thématique
- « syn » pour synonymie
- « other » pour les cas n'entrant pas dans les trois catégories citées précédemment.

A.3.1.1 Hyponyme ou hyperonyme

Ces cas devront entrer dans la définition classique de l'hyponymie ou de l'hyperonymie. La reprise devra correspondre à un terme plus générique que l'antécédent dans les cas d'hyperonymie :

(131) ***Un chien*** *a été perdu hier.* ***L'animal*** *est considéré comme dangereux.*

Elle correspondra à un terme plus spécifique pour les cas d'hyponymie.

(132) ***Un chien*** *a été perdu hier.* ***Le chihuahua*** *appartient à une grand-mère du quartier.*

A.3.1.2 Synonyme

Ces cas entrent dans la définition suivante de la synonymie. La tête nominale de l'antécédent doit être aussi spécifique que la tête du syntagme nominal anaphorique.

(133) ***Un policier*** *m'a arrêté hier.* ***Le flic*** *m'a pris pour un suspect.*

A.3.1.3 Thêta

Les reprises par le nom de rôle thématique sont définies de la façon suivante : On considère qu'on syntagme anaphorique est une reprise par un nom de rôle thématique lorsque l'entité est désignée par le rôle qu'elle joue dans l'événement décrit précédemment.

(134) ***Jack*** *a vendu beaucoup de fruits hier.* ***Le vendeur*** *est très apprécié des clients.*

A.3.2 Anaphores associatives (bridging)

On devra choisir le type d'anaphore associative parmi les catégories suivantes. Le choix sera effectué en cochant l'une des trois solutions dans la rubrique « relation ».

Has : l'entité à laquelle réfère le syntagme anaphorique est une partie de l'antécédent

Thêta : l'entité à laquelle réfère le syntagme anaphorique joue un rôle dans l'événement décrit précédemment dans le discours, mais n'a pas été mentionnée.

A.3.2.1 La relation has

L'entité doit être une partie de l'antécédent. Sa présence doit être impliquée ou facilement déduite de la présence de l'antécédent. Ces relations peuvent être de plusieurs types que nous ne détaillons pas dans l'annotation.

Il existe des anaphores associatives méronymiques, ou l'anaphore est réellement une partie de l'antécédent :

(135) *Je suis entré dans la maison. La porte était ouverte.*

(136) *J'ai rencontré Marie. Elle avait une blessure sur le nez.*

(137) *Je dois réparer ma voiture. Le moteur est cassé.*

La relation peut être fonctionnelle :

(138) *Le club de football de Bastia a perdu la finale. Le président a dit que c'était la faute des supporters.*

La relation peut être de type ensembliste, c'est à dire exprimer une relation entre un ensemble et l'un de ses éléments, ou l'un de ses sous-ensembles :

(139) *Cette famille est étrange. Les parents semblent fous.*

(140) *Un couple entra dans la pièce. L'homme/la femme portait un chapeau.*

(141) *La classe prépare un spectacle. Les filles vont chanter.*

(142) *J'aime les chiens. Mes préférés sont les épagneuls.*

La relation peut aussi être de type locative, c'est à dire que l'antécédent doit être localisé dans l'anaphore.

(143) *Je suis entré dans un village. L'église était sur une colline.*

(144) *J'entrai dans la cuisine. Le réfrigérateur était ouvert.*

La relation peut être une relation entre un objet et la matière dont il est composé.

(145) *J'aime cette chemise. La matière est douce.*

(146) *C'est une valise très chère. Le cuir est très beau.*

La relation peut être entre une tout et l'un de ses morceaux :

(147) *Jean a fait un gâteau. Marie a mangé la dernière part et a été malade.*

Enfin, la relation peut être une relation entre un événement et l'une de ses sous-parties temporelles :

(148) *J'aime cet opéra. Le duo final est merveilleux.*

A.3.2.2 La relation thêta :

L'entité joue un rôle dans l'événement dénoté par l'antécédent. Elle peut être un complément (facultatif ou obligatoire - sujet, objet, objet prépositionnel ou circonstant prépositionnel) du verbe (ou du nom dérivé d'un verbe).

(149) *La construction a duré deux ans. Les constructeurs n'avaient pas assez d'argent pour payer plus d'ouvriers.*

(150) *J'ai aimé cet opéra. La chanteuse m'a impressionné.*

(151) *Un crime a été commis. Le meurtrier court toujours.*

(152) *Quelqu'un a acheté ce livre. L'acheteur sera mécontent.*

(153) *J'ai commencé à beurrer la tartine. La margarine était périmée.*

(154) *Cette maison est belle. La construction a duré deux ans.*

(155) *Jean a été assassiné hier. Le meurtrier a été arrêté.*

A.3.3 Dans tous les cas

On pointera pour tous les syntagmes **anaphoriques** annotés (coréférence directe ou indirecte et anaphore associative) la tête de l'antécédent. (Si c'est un verbe, on marque seulement le verbe, sans les auxiliaires). Si l'antécédent n'est pas marqué préalablement (si l'antécédent n'est pas un défini), on le marquera avant de le pointer. S'il y a plusieurs antécédents possibles, on choisit le plus proche.

On marquera en plus si le syntagme annoté a des modifeurs ou s'il fait partie d'un modifieur.

A.4 Manuel d'annotation pour les démonstratifs

Comme pour l'annotation des définis, on doit d'abord savoir si le syntagme considéré a un antécédent dans le texte ou non. Le schéma d'annotation des démonstratifs est très proche de celui des définis, aussi, nous renverrons l'annotateur aux exemples cités dans le manuel pour les définis.

A.4.1 Syntagmes sans antécédent

Il existe des syntagmes nominaux démonstratifs sans antécédent dans le texte antérieur. Parmi ces syntagmes, on distingue deux cas : Il s'agit la plupart du temps de syntagmes à valeur déictique, mais il peut aussi s'agir de cataphores.

A.4.1.1 Déictiques

Si le syntagme se rapporte à un élément contextuel qui n'a pas été mentionné dans le texte, on dira qu'il s'agit d'un emploi déictique du démonstratif. On annotera alors que le démonstratif est de type déictique.

(156) *Cette année, le nombre de naissances a augmenté.*

A.4.1.2 Cataphores

Dans le cas des cataphores, l'antécédent ne se trouve pas dans le contexte antérieur au syntagme, mais dans le contexte postérieur. Les sous-types de cataphores sont les mêmes que les sous-types d'anaphore. On se reportera alors à la section suivante pour le sous-typage, mais on annotera le fait qu'il s'agit d'une cataphore.

A.4.2 Syntagmes avec antécédent

Les syntagmes nominaux démonstratifs ont la plupart du temps un antécédent dans les textes que nous traitons. Cet antécédent ne peut pas être implicite. On ne peut pas avoir :

(157) **Un attentat a été commis hier. Cette victime / ces tueurs / ces terroristes...*

On annotera alors la relation lexicale qui unit le démonstratif et son antécédent.
Cette relation peut être de deux types :
directe : si les têtes nominales des syntagmes sont identiques
indirecte : si les têtes nominales sont différentes

A.4.3 Reprises directes

On considère que la reprise est directe si les têtes nominales des deux syntagmes sont identiques :

(158) *Un chat est entré. Ce chat s'appelle Grouchat.*

A.4.4 Reprises indirectes

Comme pour le défini, on trouve plusieurs types de reprises indirectes : hyponymie, hyperonymie, synonymie, thêta, reclassification, autre.

A.4.4.1 Hyponymie et hyperonymie

On se reportera aux définitions de l'hyponymie et de l'hyperonymie de la section sur les définis.

(159) *Le chat est encore sort. Cet animal a besoin de courir.*

(160) *L'animal était superbe. Ce chat était un chartreux.*

A.4.4.2 Synonymie

On se reportera à la définition de la synonymie de la section sur les définis.

(161) *Le policier entra.Ce flic n'avait pas l'air aimable.*

A.4.4.3 Relation thêta

Ici encore, on se reportera à la section des reprises indirectes par un défini pour trouver une définition du sous-type thêta.

(162) *Jack a vendu du vin à Bill. Ce vendeur est compétent.*

A.4.4.4 Reclassification

Pour Corblin, le démonstratif a une valeur reclassifiante. Cette valeur se retrouve dans les trois catégories vues précédemment, mais peut aussi se faire par des procédés différents, où l'antécédent et l'anaphore n'ont pas de rapport lexical clair (qui peut aller jusqu'à un rapport métaphorique).

(163) *Jack a vendu un livre à Bill. Cet imbécile a oublié qu'il me l'avait réservé.*

(164) *Deux arbres encadraient l'entrée. Ces sentinelles dormaient.*

A.4.4.5 Autres

On annotera la relation « other » si la reprise ne correspond à aucune des catégories décrites ici.

Annexe B

Deuxième manuel d'annotation : groupes nominaux définis et démonstratifs coréférentiels

B.1 Déroulement des actions

Le corpus dont vous disposez est déjà annoté au niveau référentiel. Vous disposez des informations suivantes :

- le type de relation entre l'antécédent et l'anaphore (coreférentiel ou associatif)
- la présence de modifieurs
- le sous-type de relation entre l'antécédent et l'anaphore (type de relation lexicale, reclassification, etc.)
- la fonction de l'antécédent.

Nous vous demandons de travailler **uniquement sur les syntagmes nominaux coréférents**.

De façon générale, votre tâche consistera à dire si le groupe nominal anaphorique apporte de l'information nouvelle ou répète de l'information déjà connue à propos du référent. (Il faut bien entendu lire tout le texte qui précède pour répondre à cette question. Une série d'exemples illustrant les question suit ce paragraphe.) Ensuite, vous devrez déterminer la provenance de l'information contenue dans le syntagme anaphorique. Votre travail d'annotation se déroulera de la manière suivante :

Etape 1 Le groupe nominal anaphorique apporte-il de l'information que vous ignoriez à propos du référent ?

- Si non, cochez la case IRA de la ligne « informational status » et rendez vous à l'étape 2.
- Si oui, cochez la case IAA de la ligne « informational status » et rendez vous à l'étape 2bis.

Etape 2 Dans le cas où le groupe nominal répète de l'information : d'où vient-elle ? Pour répondre, vous devrez cochez l'une des possibilités de la ligne « source répétition ».

- De l'antécédent : cochez la case AO
- De l'antécédent et du contexte : cochez la case AO contexte
- D'une relation lexicale : cochez la case Lex Rel
- D'une relation lexicale et du contexte : cochez la case Lex Rel Contexte
- Des connaissances encyclopédique : cochez la case WKL

Etape 2bis Dans le cas où le groupe nominal ajoute de l'information sur le référent, d'où vient- elle ? Pour répondre, vous devrez cochez l'une des possibilités de la ligne « source répétition ».

- Des modifieurs : cochez la case modifiers
- D'une relation lexicale spécifiante : cochez la case Lex Rel
- D'une relation lexicale spécifiante et des modifieurs : cochez la case Lex Rel + modifier
- D'un groupe nominal complet sans lien particulier avec l'antécédent : cochez la case noun phrase

B.2 Exemples

Dans cette section, vous trouverez des exemples commentés permettant de mieux comprendre les questions auxquelles vous devez répondre, et la façon dont vous devez rechercher la réponse aux questions posées pour l'annotation.

B.2.1 Les anaphores qui répètent de l'information donnée (Information Repeating Anaphors - IRA)

B.2.1.1 IRA : l'information vient de l'antécédent

(c-94) *Celle-ci, (...) aurait en effet tissé un réseau de liens ambigus dans la gendarmerie, la sûreté de l'Etat, les clubs de tir. Le procès, (...) de deux membres d'une organisation néo-nazie, (...), avait permis de mettre ces liens en relief.*

Dans cet exemple, le mot « lien » est retrouvé explicitement dans l'antécédent et l'anaphore.

B.2.1.2 IRA : L'information vient de l'antécédent et du contexte

(c-95) *Le patronat avait très sensiblement modifié son comportement. (...) La clé de ce nouveau comportement tient en deux chiffres.*

Nouveau n'est pas une information nouvelle sur le comportement dont il est question puisque comportement se retrouve explicitement dans l'antécédent, et le verbe *modifier* implique que le comportement soit nouveau. Parfois, les informations peuvent provenir d'un contexte très antérieur à l'antécédent, voire de la première mention du référent.

B.2.1.3 IRA : L'information vient de la relation lexicale entre antécédent et anaphore

(c-96) *L'Inde paie un tribut sans cesse plus lourd à la sécheresse (...). Ce phénomène a été accentué par des choix économiques erronés.*

Ici, on trouve un hyperonyme (un terme plus générique) dans l'anaphore. Cette anaphore n'apporte pas d'information, le lien vient de la relation lexicale d'hyperonymie, qui ne permet pas l'apport d'information.

B.2.1.4 IRA : L'information vient d'une relation lexicale et du contexte

(c-97) *La municipalité s'est dotée récemment d'un somptueux Palais des concerts. C'est dans ce bâtiment confortable et flambant neuf qu'a eu lieu l'inauguration.*

Dans cet exemple, *bâtiment* est un hyperonyme de *Palais des concerts* (relation lexicale avec l'antécédent) ; Le terme *confortable* est une conséquence de l'adjectif *somptueux* (qui est dans l'antécédent) L'expression *flambant neuf* est inférée de l'information contenue dans le groupe verbal « s'est dotée récemment » (contexte)

B.2.1.5 IRA : L'information est inférée à partir des connaissances du monde

(c-98) *Les journalistes ne feront pas de reportage sur la visite de M. Honecker au cimetière de Neunkirchen, dans la Sarre, où sont enterrés ses parents. Ainsi en a-t-il décidé, explique Otto Schwabe, (...) après que le chef d'Etat eut requis la « tranquillité » pour cette partie « privée » de son voyage en République fédérale.*

Ici grâce au contexte et aux connaissances du monde, on sait qu'une visite sur la tombe des parents est une visite privée. On sait aussi que cette visite a lieu au cours d'un voyage officiel, il ne s'agit donc que d'une partie de la visite du chancelier est-allemand.

B.2.2 Les anaphores servant de support à de l'information nouvelle (IAA - Information adding anaphors)

B.2.2.1 IAA - L'information vient d'une relation lexicale spécifiante

(c-99) *Ce document souligne la gravité croissante des conséquences médicales de la consommation de tabac, (...). Les auteurs de ce rapport formulent une série de propositions (...).*

Ici, le mot *rapport* est plus spécifique que le mot *document*. Il s'agit d'une relation d'hyponymie.

B.2.2.2 IAA - L'information vient des modifieurs

(c-100) *L' aviation israélienne a effectué (...) un raid sur le camp de refugiés palestiniens d'Ain-el-Heloue, dans les faubourgs de Saida, chef-lieu du Liban-sud, ont rapporté les correspondants sur place. Les chasseurs-bombardiers israéliens ont effectué (...) plusieurs attaques en piqué sur ce camp qui compte soixante-mille habitants.*

Ici, le nom tête est identique dans les deux syntagmes, mais la relative apporte de l'information qui n'était pas encore mentionnée dans le texte.

B.2.2.3 IAA - L'information vient d'une relation lexicale spécifiante et des modifieurs

(c-101) *Mais à Roubaix (...), le personnel a l'impression de seulement compter les points. (...) Pour ces ouvrières du bassin houillier dont quelques-unes ont déjà trois heures de transport par jour, la nouvelle (...) a relégué au second plan les manoeuvres boursières dont leur entreprise fait l'objet.*

Ici, non seulement ouvrières est un mot plus spécifique que personnel mais en plus, il est au féminin. Par ailleurs, tous les modifieurs apportent une information qu'il n'est pas possible d'inférer à partir du texte.

B.2.2.4 IAA - l'information nouvelle est apportée par un syntagme sans relation lexicale avec l'antécédent

(c-102) *(...) je tombe sur un article intitulé « Pourquoi les maris prennent le large ». Je me dis : cherche pas, ils se débinent pendant que tu t'échines à faire des pompes et des flexions, ces salauds-là.*

Ici, il n'y a pas de lien particulier entre maris et salaud, et l'anaphore apporte le jugement du locuteur sur le référent.

Annexe C

Fichiers Schemefile

Les fichiers que nous présentons ici sont les reflets informatiques des schémas d'annotation décrits dans les annexes A et B. Le premier correspond au schéma d'annotation présenté en annexe A, et le second au schéma présenté en annexe B.

```
"type" "none"
"modifiers" "none"
"relation" "none"
"cont_inf" "none"

"type" "free"
"modifiers" "none" "yes" "no"
"in_modifier" "none" "yes" "no"
"relation" "none" "free"
"cont_inf" "non" "oui"

"type" "direct"
"modifiers" "none" "yes" "no"
"in_modifier" "none" "yes" "no"
"relation" "none" "coref"
"ante_funct" "none" "subj" "obj" "ind_obj" "circ"
"ante_status" "none" "adjunct" "tete"

"type" "indirect"
"modifiers" "none" "yes" "no"
"in_modifier" "none" "yes" "no"
"relation" "none" "syn" "hypo" "hyper" "theta" "reclass" "other"
"ante_funct" "none" "subj" "obj" "ind_obj" "circ"
"ante_status" "none" "adjunct" "tete"

"type" "bridging"
"modifiers" "none" "yes" "no"
"in_modifier" "none" "yes" "no"
"relation" "none" "has" "theta" "other"
"ante_funct" "none" "subj" "obj" "ind_obj" "circ"
"ante_status" "none" "adjunct" "tete"
```

Fig. C.1 – Premier fichier Schemefile

```
"type" "none"
"modifiers" "none"
"relation" "none"

"type" "free"
"modifiers" "none" "yes" "no"
"in_modifier" "none" "yes" "no"
"relation" "none" "free" "cont_inf"

"type" "coref"
"modifiers" "none" "yes" "no"
"in_modifier" "none" "yes" "no"
"ante_funct" "none" "subj" "obj" "ind_obj" "circ"
"ante_status" "none" "adjunct" "tete"
"relation" "none" "direct" "syn" "hypo" "hyper" "theta" "re-
class" "other" "ant_nprop"
"ant_non_nom" "metaling" "attr-app"
"informational_status" "none" "IRA" "IAA"
"source_repetition" "none" "AO" "AO_contexte" "LexRel" "LexRel_ctxt" "wkl"
"ling_mean_add" "none" "LexRel" "modifier" "LexRel_modifier" "noun_phrase"

"type" "bridging"
"modifiers" "none" "yes" "no"
"in_modifier" "none" "yes" "no"
"relation" "none" "set" "theta" "def_Ind_attr" "def_ind_assoc"
"def_mero" "copart" "circ" "wkl"
"ante_funct" "none" "subj" "obj" "ind_obj" "circ"
"ante_status" "none" "adjunct" "tete"
```

FIG. C.2 – Deuxième fichier Schemefile

Annexe D

Algorithme de Génération

Entrée :
WKL (connaissances du monde) : ensemble de règles reliant les entités les unes aux autres
DM (modèle de discours) : ensemble de formules atomiques
SM (modèle du locuteur) : ensemble de formules atomiques
t : entité cible, t appartient aux termes de SM et de DM.
F(L) : la fonction de la description (apporter ou non de l'information, L étant le nom de la description). Elle peut prendre deux valeurs : DRID ou DAIN.
ϕ : contenu sémantique de la description de l'antécédent.
LEX : les ressources lexicales.

Initialisation :
1. buts $\leftarrow$ pile avec l'élément t
2 N $\leftarrow$ structure syntaxique initiale avec une place vide pour un nom

Check success :
3. Si buts est vide, alors retourner <uniquely identifying, N>
4. but-courant $\leftarrow$ but en sommet de pile
5. Si IA(but-courant) $\nsubseteq$ PA(but-courant, L(N)), alors retourner <unfamiliar, N>
6. Si PA(but-courant, L(N)) = IA(but-courant) et $\forall$ a $\in$IA(t) : t est unique selon a étant donné L(N) alors empiler but en sommet de pile ; aller en 4.

Etendre la description :
7. Si but-courant $\in$ terms(DM) alors R $\leftarrow$ DM
8'. Si F(L) = DRID, alors choisir une formule atomique p applicable telle que :
- $\phi \models$ p
- WKL $\models$ p
- LEX $\models$ p
- ϕ + Ctxt $\models$ p
- LEX + Ctxt $\models$ p

8". Sinon, F(L) = new, alors sélectionner une formule atomique p applicable telle que Ctxt+WKL $\nvDash$ p et SM $\vDash$ p

7'. Sinon Ctxt $\leftarrow$ DM + SM

8'". Essayer de sélectionner une formule atomique p applicable telle que Ctxt+WKL $\vDash$ p, en recherchant prioritairement la relation de la façon suivante :

Ctxt+Wordnet $\vDash$ p

Ctxt+base lexicographique $\vDash$ p

Ctxt+ connaissances du monde $\vDash$ p

Ctxt+ Framenet $\vDash$ p

9. S'il n'existe pas de tel p alors retourner <non identifying, N>

10 pour chaque o $\in$ termes(p) - termes(L(N)) dépiler(o, buts)

11. N $\leftarrow$ N' tel que L(N') = L(N) $\cup$ {p}

12. Aller en 4.

Bibliographie

[Aarts, 1990] Aarts J., (1990) Corpus linguistics : an appraisal, in *Computers in literary research*, Hammesse J., et Zampolli A. (eds), pp. 13-28, Champion-Slatkine, Paris-Genève.

[Apothéloz et Reichler-Béguelin, 1999] Apothéloz D., Reichler-Béguelin M.J., (1999) Interpretations and Functions of Demonstrative NPs in Indirect Anaphora, *Journal of Pragmatics* 31, pp363-397.

[Appelt, 1985] Appelt D., (1985), *Planning English Referring Expressions* Cambridge University Press, New York.

[Baker et al., 1998] Baker C.F., Fillmore C.J., Lowe J.B. (1998), The Berkeley Framenet Project in *Proceedings of the thirty-sixth Annual Meeting of the ACL and Seventeenth International Conference on Computational Linguistics*

[Beaumont et al., 1998] Beaumont C., Lecomte J., et Hatout N., (1998) *Etiquetage morpho-syntaxique du corpus "Le Monde" pour les besoins du projet PAROLE*, Technical Report, INALF, Nancy.

[Beaver, 2002] Beaver D., (2002) The Optimization of Discourse Anaphora, à paraître dans *Linguistics and Philosophy.*

[Blackburn, 2003] Blackburn P., (2003) *Inférence et sémantique computationnelle*, Conférence invitée, TALN'03, Batz-sur-mer.

[Blackwell, 1993] Blackwell S., (1993), From dirty data to clean language, in *English language corpora : design, analysis and exploitation*, Aarts J., de Haan P., Oostdijk N. (eds), pp. 201-222, Rodopi, Amsterdam.

[Boas, 1940] Boas F., (1940) *Race, Language and Cultures*, Macmillan, New-York.

[Bonhomme, 2000] Bonhomme P., (2000), Codage et normalisation de ressources textuelles in *Ingénierie des langues*, J-M. Pierrel (ed.), Information, Commande, Communication, Hermès Science, Paris.

[Bruneseaux et Romary, 1997] Bruneseaux F., Romary L., (1997), Codage des références et des coréférences en DHM, *Actes de ACH-ALLC'97*, Kingston.

[Carletta, 1996] Carletta J., (1996) Assessing agreement on classification tasks : the Kappa statistics *Computational Linguistics*, Vol. 22 pp. 249-254.

[Charolles, 1990] Charolles M., (1990), L'anaphore associative : problèmes de délimitation, *Verbum*, n°13, Vol. 3, pp. 119-148.

[Chastain, 1979] Chastain C., (1979), Reference and Context, in *Language, Mind and knowledge*, Gunderson K. (ed), University of Minnesota Press, Minneapolis.

[Chinchor et Hirschmann, 1997] Chinchor N., Hirschmann L., (1997), MUC-7 Coreference Task Definition (Version 3.0), *Actes de MUC-7*, http ://www-nlpir.nist.gov/related_projects/muc/proceedings/co_task.html.

[Chomsky, 1956] Chomsky N., (1956), Three models for the description of language, *IRE Transactions on Information Theory*, IT-2(3), 113 : 124, New-York.

[Chomsky, 1957] Chomsky N., (1957), *Syntactic Structures*, Mouton, The Hague.

[Clark, 1977] Clark H.H., Bridging *Thinking : Readings in Cognitive Science*, Johnson-Laird P.N., Wason P.C. (eds), Cambridge, Cambridge University Press.

[Corblin, 1987] Corblin F. (1987), *Indéfini, Défini et Démonstratif*, Genève, Paris, Droz.

[Corblin, 1995] Corblin F. (1995), *Les formes de reprise dans le discours*, Presses Universitaires de Rennes.

[Corblin, 1999] Corblin F. (1999), Les références mentionnelles : le premier, le dernier, celui-ci. In *La référence (2) Statut et processus*, Mettouch A. and Quinyin H. (eds.), Travaux linguistiques du CERLICO, Rennes, PUF.

[Corley et al., 2001a] Corley S., Corley M., Keller F., Crocker M.W., et Trewin S., (2001) Finding Syntactic Structure in Unparsed Corpora : The Gsearch corpus query system, *Computer and Humanities*, 35(2), pp81-94.

[Corley et al., 2001b] Corley M., Corley S., Crocker M.W., Keller F., et Trewin S., (2001) Gsearch User Manual, Revision 1.3, http ://www.hcrc.ed.ac.uk/gsearch/

[Cosse, 2001] Cosse M., (2001), *Sur Ce N*, non publié.

[Dale et Reiter, 1995] Dale R. Reiter E., (1995), Computational Interprerations of the Gricean Maxims in the Generation of Referring Expressions *Cognitive Sciences* 19(2), pp233-263.

[Dale, 1992] Dale R., (1992), *Generating Referring Expressions*, MIT press, Cambridge, Mass.

[Danlos, 1985] Danlos L., (1985), *Génération automatique de textes en Langue Naturelle* Etudes et recherches en informatique, Masson, Paris.

[Danlos, 1987] Danlos L., (1987) *The linguistics basis of text generation*, Cambridge University Press.

[Danlos, 1999] Danlos, L., (1999), Sur la coréférence événementielle, *Actes des Journées LTT de l'AUPELF-UREF*, Beyrouth.

[Danlos et Gaiffe, 2000] Danlos, L., Gaiffe, B., (2000), Coréférence événementielle et relations de discours, in *Actes du Colloque TALN'2000*, Lausanne.

[Danlos et Roussarie, 2000] Danlos L., Roussarie L., (2000), Génération automatique de textes, in *Ingénierie des langues*, J-M. Pierrel (ed), Information, Commande, Communication, Hermès Science, Paris.

[Davies et al., 1998] Davies S., Poesio M., Bruneseaux F., Romary L. (1998), *Annotating Coreference in Dialogues : Proposal for a Scheme for MATE*, first draft, http ://www.hcrc.ed.ac.uk/ poesio/MATE/anno_manual.html

[Davies et Poesio, 1998] Davies S., Poesio M. (1998), *MATE Deliverable D1.1 Supported Coding Schemes, Coding Schemes for Coreference* http ://www.cogsci.ed.ac.uk/ poesio/MATE/coreference.html

[Deschizeaux et Reb, 1995] Deschizeaux P. et Reb G., (1995), Analyse syntaxique et sémantique d ?un sous-ensemble du français : de la théorie à la programmation, in *Langages et ordinateurs*, Reb G., (éd.), pp. 77- 97, Scolia 4, Université des Sciences humaines de Strasbourg.

[Donnellan, 1966] Donnellan K. (1966), Reference and Definite Descriptions, *Philosophical Review*, 75, pp. 281-304.

[Errenati, 2001] Errenati M., (2001), *Etude du démonstratif en corpus*, Mémoire de DEA de l'Université Paris 7.

[Fellbaum, 1998] Fellbaum C., *Wordnet. An electronic lexical database*, MIT Press, Cambridge, Mass.

[Fraurud, 1990] Fraurud, K, (1990), Definiteness and the processing of NPs in natural discourse, *Journal of Semantics*, Vol. 7, pp. 395-433.

[Fries, 1952] Fries C., (1952), *The structure of English : an introduction to the construction of sentences*, Hartcour-Brace, New-York.

[Gaiffe, 1992] Gaiffe B., (1992) *Référence et Dialogue Homme-Machine : vers un modèle adapté au multimodal*, thèse de Doctorat, Université de Nancy I.

[Gardent et Striegnitz, 2000] Gardent C., Striegnitz K., (2000), Generating Indirect Anaphora, proceedings of *IWCS'00 (International Workshop on Computational Semantics).*

[Gardent et Striegnitz, 2003] Gardent C., Striegnitz K., (2003), Generating Bridging Descriptions, à paraître dans *Computing Meaning* , Volume 3. H. Bunt and R. Muskens (eds). Studies in Linguistics and Philosophy Series Kluwer Academic Publishers, 2003.

[Gardent et al., 2003] Gardent C., Manuélian H., Kow E., (2003), Which Bridges for Bridging Descriptions, in *EACL Workshop on Linguistically Interpreted Corpora* proceedings.

[Grice, 1975] Grice H.P., (1975), Logic and conversation, in *Syntax and Semantics, Vol. 3, Speech Acts*, Cole and Morgan (eds) pp. 43-58, New York Academic Press.

[Gundel et al., 2000] Gundel J., Hedberg N., et Zacharski R., (2000), Statut cognitif et forme des anaphoriques indirects, *Verbum*, 22, pp 79-102.

[Habert et al., 1997] Habert B., Nazarenko A., Salem A., (1997), *Les linguistiques de corpus*, Armand Colin, Paris.

[Hawkins, 1978] Hawkins J.A., (1978), *Definiteness and Indefiniteness : a study in reference and grammaticality prediction*, Croom Helm, London.

[Heim, 1982] Heim I., (1982), *The Semantics of Definite and Indefinite Noun Phrases*, Ph. D. University of Massachusetts-Amherst.

[Imbs, 1971] Imbs P., (1971), *Trésor de la Langue Française. Dictionnaire de la langue des XIX et XXe siècles*, Editions du CNRS, Paris.

[Kleiber, 1986] Kleiber G., (1986), Pour une explication du paradoxe de la reprise immédiate un N - le N / un N - Ce N *Langue Française*, 72, pp 54-79.

[Kleiber, 1988] Kleiber G., (1988), Reprise immédiate et théorie des contrastes, *Studia Romanica Posnaniensa*, 13, pp 67-83.

[Kleiber, 1997] Kleiber G., (1997), Des anaphores associatives méronymiques aux anaphores associatives locatives, *Verbum*, 19, pp 25-67.

[Kleiber, 2001] Kleiber G., (2001), *Anaphore associative, lexique et référence, ou un automobiliste peut-il rouler en anaphore associative ?* , in Walter De Mulder, Co Vet, Carl Vetters (eds.), Anaphores pronominales et nominales. Amsterdam , New-York, éditions Rodopi, Collection Faux Titre, 2001.

[Krahmer et al., 2001] Krahmer E., van Erk S., Verleg A., (2001), A Meta-Algorithm for the Generation of Referring Expressions, proceedings of *8th European Workshop on Natural Language Generation* pp 29-39, Toulouse, France.

[Landragin, 2003] Landragin F., (2003) *Modélisation de la communication multimodale, Vers une formalisation de la pertinence*, Thèse de Doctorat, Université de Nancy 1.

[Lecomte, 1997] Lecomte J., (1997) *Codage Multext - GRACE pour l'action GRACE / Multitag*, Technical Report, INALF, Nancy.

[Leech, 1991] Leech G., (1991) The State of the art in corpus linguistics, *English corpus linguistics* Aijmer K., Altenberg B. (eds) pp. 8-29, Longman, London.

[Leech, 1993] Leech G., (1993) Corpus annotation schemes, *Literary and Linguistic Computing* Vol.8(4) pp. 275-281.

[Levelt, 1989] Levelt W., (1989), *Speaking : from Intention to Articulation* MIT Press, Cambridge, MA.

[Löbner, 1985] Löbner S., (1985), Definites, *Journal of Semantics*, Vol.4, pp. 279-326.

[Manuélian, 2002] Manuélian H., (2002), Annotation des descriptions définies : le cas des reprises par les rôles thématiques, proceedings of *RECITAL 2002, Nancy, France*, pp455-467.

[Manuélian, 2003] Manuélian H., (2003), Une analyse du démonstratif en corpus, proceedings of *TALN 2003, Batz sur Mer, France.*

[Markert et al., 2003] Markert K., Nissim, M., et Modjeska N.N., (2003), Using the Web for Nominal Anaphora Resolution, in Dale, R. and van Deemter, K. and Mitkov, R. (eds.), *Proceedings of the EACL Workshop on the Computational Treatment of Anaphora*, pp. 39-46, Budapest, Hungary.

[McEnery et Wilson] McEnery T., Wilson A., *Part Two : What is a Corpus, and what is in it ?*, Web pages to be used to supplement the book "Corpus Linguistics" published by Edinburgh University Press ISBN : 0-7486-0808-7 (cased) and 0-7486-0482-0 (paperback) http ://www.ling.lancs.ac.uk/monkey/ihe/linguistics/corpus2/2fra1.htm

[Mengel et al., 2000] Mengel A., Dybkjaer L., Garrido J.M., Heid U., Klein M., Pirrelli V., Poesio M., Quazza S., Schiffrin A., Soria C., (2000) *MATE dialogue annotation guidelines*, http ://www.ims.uni-stuttgart.de/projekte/mate/mdag/.

[Michea, 1964] Michea R., (1964), Les vocabulaires fondamentaux. *Recherches et techniques nouvelles au service de l'enseignement des langues vivantes*, pp. 21-36, Université de Strasbourg.

[Milner, 1982] Milner J.-C., (1982), *Ordres et Raisons de Langue*, Seuil, Paris.

[Mueller] Mueller M., *A very gentle introduction to TEI*, http ://www.tei-c.org/Sample_Manuals/mueller-main.

[Müller et Strube, 2001a] Muller C., Strube M., (2001) Annotating Anaphoric and Bridging Relations with MMAX, proceedings of *2nd SIGDial Workshop on Discourse and Dialogue*, pp90-95.

[Müller et Strube, 2001b] MMAX : A tool for the annotation of multi-modal corpora *Proceedings of the 2nd IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems*, Seattle, Wash., 5 August 2001, pp. 45-50.

[Nissim, 2001] Nissim M., (2001) *Bridging Definites and Possessives : Distribution of Determiners in Anaphoric Noun Phrases*, PhD Thesis, University of Pavia.

[Le Petit Robert] Le Petit Robert, dictionnaire alphabétique et analogique de la langue Française, sous la direction d'Alain Rey, 23° édition, 1976.

[Poesio et Vieira, 1998] Poesio M., Vieira R., (1998), A Corpus Based Investigation of Definite Description Use, *Computational Linguistics*, 24-2 pp183-216.

[Poesio, 2002] Poesio M., (2002), Scaling up anaphora interpretation, *Proc. of the Workshop on Scalable Natural Language Processing*, Heidelberg.

[Prince, 1981] Prince E.F., (1981), Towards a taxonomy of given-new information, in P. Cole (ed) *Radical Pragmatics*, pp. 223-256 Academic Press, New York.

[Quirk, 1960] Quirk R., (1960), Towards a description of English usage *Transactions of the philological society*, pp. 40-61.

[Reboul et al., 1997] Reboul A., Balkanski C., Briffault X., Gaiffe B., Popescu-Bellis A., Robba ., Romary R., Sabah G., (1997), *Le projet CERVICAL : Représentations mentales, référence aux objets et aux événements*, Rapport interne .

[Reiter, 1990] Reiter E., (1990), Generating Descriptions that exploit User's Domain Knowledge, *Current Research in Natural Language Generation*, Dale R., Mellish C. and Zock M. (eds), pp 257-285.

[Reiter et Dale, 1992] Reiter E., Dale R. (2000), A Fast Algorithm for the Generation of Referring Expressions, *Actes de COLING'92, Nantes*, pp 232-238.

[Reiter et Dale, 2000] Reiter E., Dale R. (2000), *Building Natural Generation Systems* Studies in Natural Language Processing, Cambridge University Press.

[Reiter et Sripada, 2002] Reiter E., Sripada S., (2002), Should Corpora Texts Be Gold Standards for NLG ? In *Proceedings of INLG-02* pp. 97-104.

[Russell, 1905] Russell, B. (1905), On denoting *Mind*,14 pp. 479-493.

[Salmon-Alt, 2001] Salmon-Alt S., (2001), Entre corpus et théorie : l'annotation (co)référentielle, *Linguistique de corpus, TAL*, Vol. 42- n°2/2001 pp. 459-485.

[Salmon-Alt et Vieira, 2002] Salmon-Alt S., Vieira R., (2002), Nominal Expressions in Multilingual Corpora : Definites and Demonstratives, *proceedings of LREC 2002*.

[Schnedecker, 1997] Schnedecker C., (1997), Nom propre et Chaines de Référence, *Recherches linguistiques*, n°21, Université de Metz, Klincksieck, Paris.

[Sidner, 1979] Sidner C., (1979), *Towards a computationa theory of definite anaphora comprehension in English discourse.*, PhD Thesis, MIT.

[Silberztein, 1993] Silberztein M., (1993), *Dictionnaires électroniques et analyse automatique de textes. Le système INTEX*, Informatique linguistique, Masson, Paris.

[Strand, 1997] Strand K., (1997), *A Taxonomy of Linking Relations*, Manuscript.

[Theissen, 2001] Theissen A., (2001), *La concurrence entre SN défini fidèle et SN défini totalement fidèle*, in Walter De Mulder, Co Vet, Carl Vetters (eds.), Anaphores pronominales et nominales. Amsterdam , New-York, éditions Rodopi, Collection Faux Titre, 2001.

[Traum et al., 2003] Traum D., Romary L., et Strube M., (2003), *Best Practice in Empirically-based Dialogue Research*, Diabruck 2003 Tutorial, http ://www.coli.uni-sb.de/diabruck/pages/tutorial.htm.

[Trésor de la langue française informatisé] Trésor de la Langue Française Informatisé, (2002), CNRS, http ://atilf.atilf.fr/tlf.htm.

[Van Deemter, 2000] Van Deemter K., (2000), Generating Vague Descriptions, proceedings of *First International Conference on Natural Language Generation*, pp 179-186.

[Van Deemter, 2001] Van Deemter K., (2001), Logical Form Equivalence : the Case of Referring Expressions Generation, proceedings of *8th European Workshop on Natural Language Generation* pp 21-29, Toulouse, France.

[Van der Sluis, 2001] Van der Sluis I., (2001), An Empirically Motivated Algorithm for the Generation of Multimodal Referring expressions *Student Workshop of ACL 2001* pp 67-72, Toulouse, France.

[Veronis, 2000] Veronis J., (2000), Annotation automatique de corpus : panorama et état de la technique, in *Ingénierie des langues*, J-M. Pierrel (ed), Information, Commande, Communication, Hermès Science, Paris.

[Veronis, 1998] Veronis J., (1998), *ARCADE-ROMANSEVAL, Data from the 1998 evaluation exercice*, http ://www.up.univ-mrs.fr/ veronis/data/arcroman98/Documentation/.

[Vieira, 1998] Vieira, R. (1998), A review of the Linguistic literature on definite descriptions. *Acta Semiotica et Lingvistica*, Vol. 7 : 219-258

[Vieira et al., 2002] Vieira R., Salmon-Alt S., Gasperin C., Schang E., Othero G., (2002), Coreference and Anaphoric Relations of Demonstrative Noun Phrases in a Multilingual Corpus, proceedings of *DAARC 2002*.

[Wiederspiel, 1994] Wiederspiel B., (1994), *Descriptions démonstratives anaphoriques : interprétations et stratégies référentielles*, Thèse de Doctorat, Université de Strasbourg II.

[Winston et al., 1987] Winston M., Chaffin R., et Herrmann D., (1987), A Taxonomy of Meronymic Relations, *Cognitive Science*, 11 pp. 417- 444.

[Wolters, 2002] Wolters M., (2002), *Working with Corpora*, ESSLLI'02 Lectures Notes, Trento, Italy.

Index

Résumé

L'objectif de la thèse est de parvenir, grâce à une étude de corpus, à la génération de descriptions définies référant à des entités nouvelles et de descriptions démonstratives. La première partie présente un état de l'art et la seconde expose les résultats de notre étude. Le premier chapitre expose les données théoriques et empiriques sur les expressions référentielles et les limites de ces analyses. Le deuxième chapitre présente la problématique de la génération d'expressions référentielles, et le troisième présente la linguistique de corpus et le traitement de corpus électroniques. La première partie s'achève par une synthèse reliant les trois domaines abordés. Le cinquième chapitre présente les travaux réalisés sur le corpus, des pré-traitements à l'extraction des résultats. Les sixième et septième chapitres exposent les résultats d'une étude des anaphores associatives et des SN coréférentiels annotés dans notre corpus et deux extensions de l'algorithme de Gardent et Striegnitz. Le dernier chapitre présente les contraintes sémantiques et syntaxiques identifiées à l'aide du corpus sur le choix du déterminant des descriptions.

Mots-clés: Analyse de corpus, génération automatique de textes, descriptions définies et démonstratives, coréférence, anaphore associatives.

Abstract

The objective of the thesis is to generate demonstrative descriptions and definite descriptions that refer to new referents, by carrying out a corpus study of 10,000 definite and demonstrative NPs. The first part of the thesis is a state-of-the-art overview and the second shows the results of the present work. The first chapter reviews some theoretical and empirical results on referential NPs and shows the limitations of these analyses. The second chapter discusses the problems involved in the generation of referential expressions, and the third presents the concepts behind corpus linguistics and the processing of digital corpora. The first part concludes by showing the relationships between the above three domains. The fifth chapter of this thesis presents the corpus study that was carried out, from automated pre-processing to the extraction of results. The sixth ans seventh chapters discuss findings on bridging and coreferential NPs and two extensions to the Gardent-Striegnitz algorithm. The final chapter presents the semantic and syntactic constraints on the choice of determinant in referring NPs which were identified through the corpus study.

Keywords: Corpus analysis, Automatic generation, definite and demonstrative descriptions, coreference, bridging.