

Une analyse des emplois du démonstratif en corpus

Hélène Manuélian
LORIA
BP239 Campus Scientifique
F-54506 Vandoeuvre-les-Nancy
helene.manuelian@loria.fr

Mots-clefs – Keywords

Démonstratifs, analyse de corpus, annotation de corpus, génération de texte.
Demonstrative, corpus analysis, corpus annotation, text generation.

Résumé - Abstract

Cet article propose une nouvelle classification des utilisations des démonstratifs, une mise en oeuvre de cette classification dans une analyse de corpus et présente les résultats obtenus au terme de cette analyse. La classification proposée est basée sur celles proposées dans la littérature et étendue pour permettre la génération de groupes nominaux démonstratifs. L'analyse de corpus montre en particulier que la nature "reclassifiante" du démonstratif lui permet d'assumer deux fonctions (une fonction anaphorique et une fonction de support pour de l'information nouvelle) et qu'il existe des moyens variés de réaliser ces fonctions.

This paper presents a new classification for the use of demonstrative descriptions, its application in a corpus analysis and the results of this analysis. The proposed classification is based on the existing literature and extended to support the generation of demonstrative NPs. The corpus analysis shows in particular, that the "reclassifying power" of the demonstrative allows it to perform two functions (anaphora and supporting new information) and that there are various ways of realising these functions.

1 Introduction

Les algorithmes de génération d'expressions référentielles se sont jusqu'à présent préoccupés essentiellement de la génération de descriptions définies. L'un des buts essentiels de ces algorithmes est de produire des descriptions suffisamment informatives pour permettre au locuteur d'identifier le référent décrit par le système (principalement (Dale et Reiter, 1995)). Les travaux de Van Deemter (Van Deemter, 2000; Van Deemter, 2001) étendent les algorithmes de Reiter et Dale à la génération de descriptions plus complexes (vagues ou booléennes), (Krahmer et al., 2001) proposent un algorithme ne se basant plus seulement sur des propriétés internes au référent pour le décrire, mais aussi sur ses relations (spatiales par exemple) avec les autres objets du contexte. Dans une perspective différente, (Gardent, Striegnitz, 2000) présentent un algorithme de génération des anaphores associatives, basé sur un modèle structuré du contexte et sur un moteur d'inférences. Tous ces algorithmes permettent la génération de descriptions définies anaphoriques et rendent compte de l'opposition description définie / pronom. En revanche, ils ne rendent pas compte de l'opposition défini / démonstratif et ne permettent donc pas la génération de descriptions démonstratives, qui, si on se réfère aux études théoriques (Corblin, 1987; Kleiber, 1986; Kleiber, 1988) et aux études empiriques (Vieira et al. 2002), sont employées la plupart du temps en mention subséquente, ce qui les met directement en concurrence avec les descriptions définies et les pronoms dans le choix des expressions référentielles. Une façon d'améliorer les algorithmes de génération d'expressions référentielles serait donc d'y intégrer des possibilités de génération de démonstratifs.

Nous proposons ici un premier pas dans cette direction, en présentant une étude et une classification des utilisations des démonstratifs en corpus. Nous montrons d'abord que les analyses théoriques du démonstratif sont trop générales pour être appliquées en génération automatique d'expressions référentielles. Nous présentons ensuite l'étude de corpus qui nous a permis d'aboutir à une classification des utilisations du démonstratif et de combler certains manques des analyses théoriques. Nous donnons enfin quelques premières pistes pour la génération automatique de descriptions démonstratives.

2 Utilisations du démonstratif

Le démonstratif est classiquement reconnu comme déterminant un groupe nominal dont le référent est présent (physiquement ou linguistiquement) dans le contexte et (ou) focalisé. Les études théoriques identifient essentiellement quatre utilisations du démonstratif en français (Kleiber, 1986; Kleiber, 1988; Corblin, 1987; Apothéloz et Reichler-Béguelin, 1999).

- Utilisation en première mention (déictique) : Le référent est identifié dans le contexte immédiat, et l'expression est parfois accompagnée d'un geste ostensif afin que l'identification se fasse sans ambiguïté.

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

Cette année, la récolte sera bonne.

- Utilisation coréférentielle anaphorique : La reprise peut être fidèle ou non, mais les expressions utilisées pour désigner le référent entretiennent une relation lexicale connue (hypéronymie, hyponymie, ou synonymie).

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

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

Reprise indirecte : Elle peut prendre les formes suivantes :

Une analyse des emplois du démonstratif en corpus

- 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*" :

Un chat entra dans la pièce. Cet 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*."

Un chat entra dans la pièce. Ce 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.

Le policier entra. Ce flic n'avait pas l'air aimable.

- **Utilisation coréférentielle reclassifiante** : 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. Elle permet au locuteur d'émettre un jugement ou d'apporter une information nouvelle sur le référent ou d'utiliser une figure de style :

Un homme entra dans le café. Cet imbécile commença à provoquer les habitués.

Jean est venu. Ce professeur de philosophie ... (Corblin, 1987)

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

- **Utilisation associative** : On trouve aussi des mentions d'utilisation associative du démonstratif dans (Apothéloz et Reichler-Béguelin, 1999; Gundel et al., 2000; Nissim, 2001) :

Nous arrivâmes dans un village. Cette église, tout de même, quelle horreur ! (Charolles in (Apothéloz et Reichler-Béguelin, 1999))

3 Analyse de corpus

Le corpus dont nous disposons est un extrait du corpus PAROLE¹, comprenant 65 000 mots et 545 syntagmes nominaux démonstratifs. Il est composé d'articles du journal Le Monde datant de septembre 1987 et tirés de toutes les rubriques (Politique, Economie, Sport, Culture et Loisirs). Le corpus est annoté au niveau morphosyntaxique suite au projet PAROLE selon le schéma d'annotation Multext / Multitag de l'action GRACE (Lecomte, 1997; Beaumont et al. 1998). Afin d'automatiser au maximum le travail sur corpus, nous avons réalisé une série de pré-traitements avant l'annotation, décrits dans cette section.

3.1 Pré-traitements

Repérage des descriptions définies. G-search tool (Corley et al., 2001) est un outil qui permet de rechercher des structures syntaxiques dans un corpus non parsé grâce à une grammaire définie par l'utilisateur. Il nous a permis de préparer le corpus à l'annotation en isolant les groupes nominaux démonstratifs, à partir de l'information présente dans les étiquettes Multext / Multitag et d'une grammaire utilisant cette information. Nous avons ensuite écrit un filtre pour adapter le format de sortie de G-search au format requis en entrée de notre outil d'annotation (MMAX), c'est-à-dire du texte avec des balises XML pour chaque mot et des balises supplémentaires pour les syntagmes à annoter, de façon à ce qu'ils soient mis en évidence dans le texte.

Annotation avec MMAX. MMAX (Muller et Strube, 2001) a été conçu spécifiquement pour

¹Ce corpus a été fourni par le laboratoire ATILF (Analyse et Traitement Informatique de la Langue Française, UMR 7118, Nancy), dans le cadre de la collaboration LORIA-ATILF dans le Contrat de Plan Etat-Région

annoter les corpus au niveau référentiel. Il permet d'annoter les relations de coréférence et d'anaphore associative. 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 dans une autre fenêtre de l'application où apparaissent les attributs définis pour l'annotation, on insère des balises XML qui permettent d'identifier le type de relation anaphorique, et de pointer l'antécédent du syntagme. La sortie de MMAX est un fichier XML. Nous avons ensuite écrit des feuilles de style en XSL et des "scripts de shell" afin de réaliser le comptage des phénomènes, d'extraire et de trier les données issues de l'annotation en les présentant dans des fichiers HTML.

3.2 Schéma d'annotation

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 du démonstratif vue en section 2, puis nous avons affiné nos catégories. En effet, bien que les analyses théoriques nous semblent incomplètes pour la génération (cf. section 4), elles donnent des informations sur les moyens utilisés pour faire des reprises par le démonstratif. Nous avons donc annoté les utilisations du démonstratif dans les catégories suivantes : **Utilisations déictiques, utilisations associatives, reprises directes, et reprises indirectes**. Cette dernière catégorie sera sous-typée de la façon suivante :

- **hyperonymie**

- **hyponymie**

- **synonymie** : Nous ajoutons à la définition de la section 2 le fait que parfois, deux groupes nominaux ont le même sens, si on tient compte des modificateurs. Nous en avons tenu compte dans les cas où l'un des noms est un nom prédicatif qui sous-catégorise des compléments.

"Nous sommes dans l'ignorance la plus totale de ce qui se passe et de ce que veulent les Chargeurs. Mais, après tout, c'est le fonctionnement normal du marché boursier", constate un cadre. Cette pénurie d'informations, les ouvriers des usines quasi désertes de Tourcoing et de Cambrai n'en ont pas trop souffert.

- **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. Un syntagme anaphorique est classé ici lorsque l'antécédent (nominal ou non) est désigné par son rôle dans l'événement décrit précédemment (Manuélian, 2002).

Jack a vendu du vin à Bill. Ce vendeur est compétent.

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

Jack a vendu un livre à Bill. Cet imbécile a oublié qu'il me l'avait réservé.

Deux arbres encadraient l'entrée. Ces sentinelles dormaient.

- **autres** : La reprise n'entre pas dans les catégories décrites précédemment.

Ces classes ont constitué une première base à notre annotation. Nous avons ensuite, dans une deuxième passe d'annotation, affiné les catégories "autres" (trop importante) et "reclassification" (trop hétérogène), et ajouté les catégories suivantes :

- **métalinguistique** : L'antécédent (nominal ou non) est repris par un nom désignant sa nature linguistique, ou par un élément mentionnel du type "ce dernier". Ces reprises avaient été classées au départ dans la catégorie "autres".

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 nou-

veaux Etats sont venus s'ajouter en 1987 à *cette liste*.

- **antécédent = nom propre** : Nous avons classé à part les reprises de noms propres (au départ dispersées dans les catégories "hypéronyme", "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 dans la catégorie "métalinguistique", parce que comme pour les noms communs, il n'y a pas de relation lexicale entre l'antécédent et l'anaphore.

- **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 "theta" et "métalinguistique".

Notre corpus contient 545 descriptions démonstratives, dont nous avons exclu 17 cas d'appositions, ou de groupes nominaux liés à l'antécédent par une copule. Nos résultats (figure 1) sont proches de ceux des autres annotations connues (Vieira et al. 2002). Nous y retrouvons 59% de reprises indirectes, dont 24% passent par une relation lexicale. On constate aussi qu'il y a peu d'emplois en première mention et en association (21% et 2%), ce qui distingue bien l'emploi du démonstratif de l'emploi du défini (Poesio, Vieira 1998)².

| Relation | Nombre | Pourcentage |
|-------------------------|--------|-------------|
| Première mention | 113 | 21% |
| Association | 9 | 2% |
| Reprises Directes | 94 | 17% |
| Reprises Indirectes | 312 | 59% |
| hyponymes | 16 | 3% |
| hyperonymes | 74 | 14% |
| synonymes | 21 | 4% |
| theta | 18 | 3% |
| reclassifications | 73 | 14% |
| antécédent = nom propre | 43 | 8% |
| antécédent non nominal | 52 | 10% |
| métalinguistique | 22 | 4% |
| autres | 3 | 0,6% |
| TOTAL | 528 | 100 |

Figure 1: Résultats de l'annotation

4 Classification des utilisations du démonstratif

Pour générer des SN démonstratifs, on a besoin de connaître les moyens linguistiques utilisés pour faire des reprises avec un déterminant démonstratif et les fonctions de ces reprises. En observant les données de notre corpus, nous distinguons deux fonctions sémantiques du démonstratif utilisé en coréférence :

- il détermine un syntagme anaphorique n'ajoutant pas d'information sur le référent ;

²S'agissant d'une étude préliminaire, cette annotation a été réalisée par une seule personne. Nous ne pouvons donc pas donner de taux d'accord, mais précisons que l'annotation a ensuite été vérifiée par deux personnes, jusqu'à accord complet sur les catégories.

- il détermine un syntagme anaphorique qui, par le biais de la reclassification (ou de l'hyponymie), peut donner de l'information supplémentaire sur le référent.

Ces deux utilisations nous semblent fondamentalement différentes, mais la littérature ne répond pas aux questions suivantes : Par quels moyens linguistiques se réalisent ces fonctions ? L'opération de reclassification est-elle toujours utilisée pour ajouter de l'information ? D'où viennent les informations qui permettent d'interpréter l'anaphore ou de produire cette nouvelle information ? Par ailleurs, la littérature ne s'intéresse qu'aux liens entre les têtes nominales des syntagmes. En génération, nous avons besoin de connaître le contenu sémantique global du syntagme, modificateurs inclus, parce que nous devons générer une description suffisante pour identifier le référent. Quels sont alors les liens entre les deux syntagmes pris dans leur totalité ? Notre but sera donc de répondre à ces questions, fondamentales pour la génération, en nous appuyant sur une étude de corpus.

Dans notre corpus, la fonction anaphorique sans ajout d'information apparaît dans 258 cas (73,5% des reprises), et la fonction d'ajout d'information dans 93 cas (26,5% des reprises). Après une analyse des reprises de notre corpus, nous avons établi une classification mettant en évidence à la fois la fonction de l'anaphore et la provenance des informations permettant sa génération ou son interprétation. Nous ne considérons pas pour le moment les reprises classées dans les catégories "antécédent non nominal" et "autres". Nous présentons donc nos résultats sur 351 cas de reprises démonstratives.

4.1 Reprises sans ajout d'information

Il s'agit de la catégorie la plus représentée dans le corpus (3 reprises sur 4). L'information communiquée par le syntagme démonstratif est déjà connue et donnée. Cependant, elle n'est pas forcément identique à celle contenue dans l'antécédent, et peut provenir de sources diverses. D'après notre analyse de corpus, ces sources sont les suivantes (tous les exemples cités dans cette section sont extraits du corpus analysé) :

1. L'information utilisée dans la reprise provient de l'antécédent :

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.

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

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.

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

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.

4. L'information est inférée à partir de l'antécédent grâce au savoir lexical et à partir du contexte :

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.

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

"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 Horizon, après que le chef d'Etat lui-même eut requis la "tranquillité" pour cette partie "privée" de son voyage en République fédérale.

| | Antec. | Antec. + ctxt | Antec + lex | Antec. + lex + ctxt | C. En- cycl. |
|--------------|-------------|------------------|----------------|---------------------------|-----------------|
| Direct + mod | 7 | 4 | 0 | 0 | 0 |
| Direct | 72 | 0 | 0 | 0 | 0 |
| hyper + mod | 0 | 0 | 7 | 5 | 2 |
| hyper | 0 | 0 | 49 | 0 | 0 |
| hypo | 0 | 0 | 0 | 2 | 0 |
| syn + mod | 1 | 0 | 3 | 4 | 1 |
| syn | 0 | 0 | 8 | 0 | 0 |
| theta + mod | 0 | 0 | 0 | 6 | 0 |
| theta | 0 | 0 | 0 | 8 | 0 |
| recl + mod | 0 | 0 | 0 | 0 | 14 |
| recl | 0 | 0 | 0 | 0 | 19 |
| Npropre | 0 | 0 | 0 | 0 | 24 |
| metaling | 0 | 22 | 0 | 0 | 0 |
| total | 80 (21%) | 26 (10%) | 67 (26%) | 25 (10%) | 60 (23%) |

| | Sans in- férences | Avec in- férences |
|-------------------------|----------------------|----------------------|
| Direct | 72 | 0 |
| Direct + modifieur | 7 | 4 |
| Indirect | 0 | 132 |
| Indirect + modifieur | 1 | 42 |
| total | 80 (21%) | 178 (69%) |

Figure 2: Répartition des reprises sans ajout d'information

Nous pouvons constater (figure 2, tableau de gauche) les diverses façons dont sont réalisées les reprises (les catégories de l'annotation ainsi que la provenance de l'information sont abrégées dans le tableau et la mention "+mod" signifie que la reprise est modifiée). La reclassification ne signifie pas systématiquement qu'un ajout d'information est effectué. Les informations contenues dans la reprise proviennent de plusieurs sources :

- de l'antécédent (majoritairement : 57%), dans le cas des reprises directes ou avec une relation lexicale connue,
- du contexte, (20%)
- des connaissances encyclopédiques. (23%)

Le tableau de droite (figure 2) présente un résumé du tableau de gauche. On peut y lire clairement que les reprises sans ajout d'information sont produites majoritairement à partir d'inférences (69% des cas).

4.2 Reprises avec ajout d'information

La capacité "reclassifiante" du démonstratif peut parfois servir à ajouter de l'information. Dans cette section, nous montrons par quels moyens linguistiques sont réalisés les ajouts d'information dans les reprises par un syntagme démonstratif.

1. Relation lexicale spécifiante :

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.

2. Relation lexicale spécifiante et modifieurs :

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 houillier 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.

3. L'information nouvelle est dans les modifieurs :

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**, (...).

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 (...).

4. L'information nouvelle est dans tout le syntagme et ne passe pas par une relation lexicale :

Métaphore : 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 (...).

Introduction du point de vue du locuteur : 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**.

Changement de point de vue : 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.

Introduction de propriété: **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 (...).

Le tableau (figure 3) nous montre les éléments suivants : L'information nouvelle est ma-

| Relation | syntagme complet | rel. lex. | modifieur | mod. + rel. lex. |
|--------------|------------------|-----------|-----------|------------------|
| Direct mod | 0 | 0 | 11 | 0 |
| hyper mod | 0 | 0 | 11 | 0 |
| hypo mod | 0 | 2 | 0 | 8 |
| hypo | 0 | 4 | 0 | 0 |
| syn mod | 0 | 0 | 4 | 0 |
| theta mod | 0 | 0 | 4 | 0 |
| recl mod | 17 | 0 | 0 | 0 |
| recl | 13 | 0 | 0 | 0 |
| N propre mod | 5 | 0 | 14 | 0 |
| total | 35 (38%) | 6 (6,5%) | 44 (47%) | 8 (8,5%) |

Figure 3: Répartition des reprises avec ajout d'information

ajoritairement introduite par les modifieurs de l'anaphore (47% des cas). Dans 38% des cas, l'information nouvelle est apportée par le biais de la reclassification sans lien lexical identifié, dans la totalité du syntagme. Enfin, on constate que l'hyponymie est assez peu utilisée pour ajouter de l'information (15% des cas).

4.3 Synthèse

Les résultats de ces analyses de corpus nous amènent aux conclusions suivantes : Pour générer une reprise démonstrative sans ajouter d'information, plusieurs choix sont offerts, dans l'ordre de préférence suivant (nous nous basons pour le moment sur la fréquence d'apparition des formes de reprises, n'ayant pas encore identifié d'autres critères de choix) :

On génère un syntagme sans modifieur :

- La tête de l'anaphore est inférée du contexte, du savoir lexical et des connaissances du monde.
- La tête de l'anaphore répète la tête du syntagme antécédent.

On génère un syntagme nominal modifié :

- La description démonstrative anaphorique est inférée dans sa totalité du contexte, du savoir lexical et des connaissances du monde.
- La description anaphorique est une répétition partielle ou totale de l'antécédent.

Pour utiliser une reprise démonstrative comme support à de l'information nouvelle, on utilise (dans l'ordre de préférence, toujours établi à partir de la fréquence d'apparition dans le corpus) :

- Une description définie sans relation lexicale avec l'antécédent,
- Les modifieurs dans une description démonstrative dont la tête est inférée de la tête de l'antécédent
- Une relation d'hyponymie
- Les modifieurs dans une description démonstrative dont la tête est identique à celle de l'antécédent.

5 Conclusions et perspectives

Les analyses existantes ne tiennent pas compte de toutes les données que nous avons pu réunir et qui sont nécessaires en génération de texte. Tout d'abord, en génération, on a besoin de connaître le but communicatif de l'expression à générer. Nous en avons identifié deux pour les reprises par une description démonstrative : une fonction purement anaphorique, et une fonction d'ajout d'information. Ensuite, pour générer une description démonstrative, on doit connaître le contenu sémantique global du syntagme, la provenance possible des informations connues qu'il contient s'il est anaphorique et la façon dont elles peuvent être exprimées. On doit aussi pouvoir donner des indications sur la façon dont l'information nouvelle est réalisée. Notre classification prend en compte ces contraintes et donne des indications sur les moyens linguistiques de faire une reprise par un syntagme démonstratif. Cependant, cette étude est encore incomplète, et pour parvenir à un algorithme de génération efficace, il nous reste encore à donner des contraintes sur le choix de la réalisation de l'expression référentielle. Ceci pourra se faire si on étudie :

- Comment fonctionnent les inférences sur le contexte dans la résolution d'anaphore, et quel type d'information est réutilisable pour la reprise anaphorique.
- S'il y a des restrictions (liées au discours ou à la position dans la chaîne anaphorique) sur l'utilisation de tel ou tel moyen de faire la reprise.
- Les différences avec le défini afin de déterminer dans quel cas on utilise quel déterminant et s'il y a des cas où le choix est impératif. L'étude de corpus sur les définis présentée dans (Gardent et al. 2003) est une première étape dans cette direction.

Références

- Apothéloz D., Reichler-Béguelin M.J., (1999) Interpretations and Functions of Demonstrative NPs in Indirect Anaphora, *Journal of Pragmatics* 31, pp363-397.
- Beaumont C., Lecomte J., et Hatout N., (1998) *Etiquetage morpho-syntaxique du corpus "Le Monde" pour les besoins du projet PAROLE*, Rapport Technique, INALF, Nancy.
- Corblin F. (1987), *Indéfini, Défini et Démonstratif*, Genève, Paris, Droz.
- 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.
- Cosse M., (2001), *Sur Ce N*, non publié.
- Dale R. Reiter E., (1995), Computational Interpretations of the Gricean Maxims in the Generation of Referring Expressions *Cognitive Sciences* 19(2), pp233-263.
- Gardent C., Striegnitz K., (2000), Generating Indirect Anaphora, Actes de *IWCS'00 (International Workshop on Computational Semantics)*.
- Gardent C., Manuélian H., Kow E., (2003), Which Bridges for Bridging Descriptions, à paraître dans les Actes de *EACL Workshop on Linguistically Interpreted Corpora*.
- Gundel J., Hedberg N., et Zacharski R., (2000), Statut cognitif et forme des anaphoriques indirects, *Verbum*, 22, pp 79-102.
- 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 G., (1988), Reprise immédiate et théorie des contrastes, *Studia Romanica Posnaniensa*, 13, pp 67-83.
- Krahmer E., van Erk S., Verleg A., (2001), A Meta-Algorithm for the Generation of Referring Expressions, Actes de *8th European Workshop on Natural Language Generation* pp 29-39.
- Lecomte J., (1997) *Codage Multext - GRACE pour l'action GRACE / Multitag*, Rapport Technique, INALF, Nancy.
- Manuélian H., (2002), Annotation des descriptions définies : le cas des reprises par les rôles thématiques, Actes de *RECITAL*, pp455-467.
- Muller C., Strube M., (2001) Annotating Anaphoric and Bridging Relations with MMAX, Actes de *2nd SIGDial Workshop on Discourse and Dialogue*, pp90-95.
- Nissim M., (2001) *Bridging Definites and Possessives: Distribution of Determiners in Anaphoric Noun Phrases*, PhD Thesis, University of Pavia.
- Poesio M., Vieira R., (1998), A Corpus Based Investigation of Definite Description Use, *Computational Linguistics*, 24-2 pp183-216.
- Vieira R., Salmon-Alt S., Gasperin C., Schang E., Othéro G., (2002), Coreference and Anaphoric Relations of Demonstrative Noun Phrases in a Multilingual Corpus, Actes de *DAARC*.
- Van Deemter K., (2000), Generating Vague Descriptions, Actes de *First International Conference on Natural Language Generation*, pp 179-186.
- Van Deemter K., (2001), Logical Form Equivalence : the Case of Referring Expressions Generation, Actes de *8th European Workshop on Natural Language Generation* 21-29.