Université de Montréal

Les expressions démonstratives en tant que marqueurs de la cohésion textuelle en vue de l'analyse automatique de textes.

par

Pierre-Luc Vaudry

Département de linguistique et de traduction

Faculté des arts et des sciences

Mémoire présenté à la Faculté des études supérieures
en vue de l'obtention du grade de Maître ès arts (M.A.)
en linguistique

Mai 2007

Université de Montréal
Faculté des études supérieures

Ce mémoire intitulé :

Les expressions démonstratives en tant que marqueurs de la cohésion textuelle en vue de l'analyse automatique de textes

présenté par
Pierre-Luc Vaudry

a été évalué par un jury composé des personnes suivantes :

Patrick Drouin
président-rapporteur

Nathan Ménard
directeur de recherche

Richard Kittredge
membre du jury

Mémoire accepté le .....

Résumé

Ce mémoire porte sur les expressions démonstratives en tant que marqueurs de la cohésion textuelle. L'objectif est de contribuer à l'avancement de l'analyse automatique de textes. Une revue des approches utilisées en linguistique informatique et en analyse du discours pour résoudre le problème de la résolution des anaphores est effectuée, suivie d'une récapitulation des études sur les caractéristiques des démonstratifs quant à leur contribution aux phénomènes endophoriques et exophoriques. On teste ensuite une série d'hypothèses sur le comportement de plusieurs classes d'expressions démonstratives à travers une étude de corpus. Le corpus, constitué d'articles de journaux en français, est annoté en XML. L'analyse des résultats fait ressortir certains faits intéressants : le comportement fort régulier d'expressions spécialisées comme CE DERNIER; le nombre plus élevé de COD que de sujets chez les antécédents syntagmes nominaux du pronom CE; ainsi que la proportion importante d'antécédents propositionnels des pronoms CECI, CELA et ÇA.

Mot-clés

résolution des anaphores, linguistique informatique, analyse du discours, anaphore, endophore, exophore, déixis, corpus, XML, antécédent

Abstract

This dissertation discusses demonstrative expressions as textual cohesion markers. Its aim is to contribute to the field of automated text analysis. Approaches to anaphora resolution coming from computational linguistics and discourse analysis are reviewed, followed by a summary of studies on the characteristics of demonstratives in their contribution to endophoric and exophoric phenomena. A set of hypotheses on the behavior of several classes of demonstratives expressions are then tested by conducting a corpus study. The corpus is comprised of newspaper articles written in French and annotated using XML. The results reveal some interesting facts: some specialized expressions like CE DERNIER behave in a strongly regular manner; the pronoun CE has a greater number of direct object noun phrases, rather than subjects, as antecedents; and the pronouns CECI, CELA and ÇA show a large proportion of propositional antecedents.

Keywords

anaphora resolution, computational linguistics, discourse analysis, anaphora, endophora, exophora, deixis, corpus, XML, antecedent

Table des matières

Liste des tableaux

Liste des figures

Liste des abréviations, des sigles et des symboles

majuscules : Un mot écrit en majuscules indique qu'il s'agit d'un nom de lexie.

indice : Un nom de lexie accompagné d'un indice indique qu'il y a plus d'une lexie ayant ce nom. Le numéro sert alors à identifier la lexie de manière univoque.

$\rightarrow$: implication logique

$\equiv$: équivalence logique

$\neg$: négation logique

$\circ$: opérateur de composition de fonctions

CdN : complément du nom

COD : complément d'objet direct

COI : complément d'objet indirect

FL : fonction lexicale

CSS : *Cascading Style Sheet*

DTD : *Document Type Definition*

HTML : *HyperText Markup Language*

iff : *if and only if*

NP : *Noun Phrase*

SG : sujet grammatical

SN : syntagme nominal

S_i : nom typique de l'actant *i*

S_{loc} : nom typique de circonstant de lieu

S_{instr} : nom typique de circonstant d'instrument

S_{med} : nom typique de circonstant de moyen

S_{mod} : nom typique de circonstant de manière

ssi : si et seulement si

TST : Théorie Sens-Texte

XML : *Extensible Markup Language*

XPATH : *XML Path Language*

XSLT : *Extensible Stylesheet Language Transformation*

Remerciements

J'aimerais remercier mon directeur de recherche actuel, Nathan Ménard. Sa grande expérience et son talent manifeste pour la maïeutique m'ont grandement aidé à terminer ce mémoire. J'aimerais également remercier mon ancien directeur de recherche, Richard Kittredge, pour la confiance qu'il m'a accordée au début de ma maîtrise et pour les précieux conseils qu'il m'a prodigués dans les premières phases de ma recherche. Je dois aussi remercier Jean-Yves Morin, pour m'avoir assisté dans la formulation de mon sujet. Je veux encore remercier le directeur du département de linguistique et de traduction, Richard Patry, pour avoir fait le nécessaire pour qu'il me soit possible de terminer ma rédaction. Merci aussi aux autres professeurs et employés de l'université qui m'ont aidé dans mes démarches.

Toute ma gratitude va à ma famille et à mes amis, pour m'avoir soutenu dans les meilleurs moments, comme dans les plus laborieux : ma mère Doris, mon père François, ma sœur Myriam, ainsi que mes amis Maryse, Olivier, Vincent, Julien, Caroline, Malik, Jean-François, Pascal, Marco, Marilène, Maxime, l'autre Olivier, Alexandra, Marc-Étienne, Mathieu. Merci aussi à mes collègues présents et passés du département. Un geste, une parole ou parfois votre simple présence m'ont un jour ou l'autre été d'un grand réconfort : Andréanne, Mylène, Sophie, l'autre Sophie, Andrée, Charles, l'autre Charles, Aimée, Amélie, Émilie, l'autre Émilie, Alexandre, Wajdi, Pascale, Sylvie, Sara-Anne, Philippe, Sébastien, Stephen, Heather, Chantal, Gavin, Elizaveta, Myriam, Lissette. Merci enfin à C. Gagnon et D. Ramirez pour m'avoir aidé dans des moments plus difficiles.

Chapitre 1 Introduction

La présente étude s'inscrit dans deux sous-branches de la linguistique : la linguistique informatique et l'analyse du discours. Du côté de l'analyse du discours, elle prend source dans les travaux sur la cohésion textuelle, les liens anaphoriques et les chaînes de co-référence. Le pendant de ces travaux en linguistique informatique est baptisé «résolution des anaphores».

Parmi les expressions linguistiques jouant souvent un rôle dans la cohésion textuelle, nous avons voulu dans cette étude nous concentrer sur un type d'expression moins étudié que les autres. Il s'agit des expressions démonstratives. Nous entendons par là les syntagmes nominaux dont la tête est un pronom démonstratif, ceux dont la tête est modifiée par un déterminant démonstratif[1], et aussi ceux qui sont construits autour d'un adverbe démonstratif (ICI ou LÀ).

L'objectif est de déterminer dans quelles conditions les expressions démonstratives jouent le rôle d'exophores ou d'endophores[2]; s'il s'agit d'endophores, où se trouvent leurs antécédents, sur quels indices peut-on se baser pour les localiser et quelles relations entretiennent-ils avec eux, sur les plans lexical, sémantique ou référentiel. Étant donné que les informations recueillies doivent permettre de faire avancer le champ de la résolution automatique des anaphores, sans nécessairement déboucher sur un algorithme complet, il faut que les hypothèses soient formulées de manière suffisamment formelle pour pouvoir être utile dans le cadre du traitement automatique de texte. Plus précisément, c'est la perspective de l'analyse, plutôt que celle de la synthèse ou de la génération, qui sera privilégiée ici.

Pour ce faire, des hypothèses portant sur diverses formes d'expressions démonstratives sont élaborées, à la suite d'une revue de la littérature et d'une série d'observations préliminaires sur un échantillon réduit de textes. Ensuite, ces hypothèses sont mises à l'épreuve par l'étude d'un corpus plus important. Dans

---

[1] Les termes *adjectif démonstratif* et *article démonstratif* sont également utilisés par certains auteurs, de même que par certaines grammaires traditionnelles. Pour nous, ces trois expressions sont équivalentes.

[2] Pour la notion d'endophore, prière de se référer au paragraphe qui débute en bas de la page 11 et se termine en haut de la page 12 de ce mémoire. Pour la notion d'exophore, se référer au dernier paragraphe de la page 28.

celui-ci, chaque démonstratif a d'abord été identifié et étiqueté en notant le genre, le nombre, la partie du discours, etc. Dans un deuxième temps, les limites des syntagmes pertinents ont été déterminés, les antécédents, s'il y lieu, identifiés, et la relation entretenue avec eux qualifiée selon la distance, la présence d'une relation lexicale, la coréférence, etc. Puis, toutes ces informations sont regroupées et les différentes hypothèses sont vérifiées. Des analyses quantitatives et qualitatives appuient les conclusions qui en sont tirées.

Chapitre 2 Question de recherche

2.1 Question de recherche principale

La principale question de recherche à laquelle cette étude tente de répondre est celle de la pertinence, pour la résolution des anaphores, de tenir compte des particularités des démonstratifs quant à leur contribution à la cohésion textuelle. Autrement dit, dans quelle mesure peut-on appliquer formellement à l'analyse automatique de textes[3] des caractéristiques propres au fonctionnement anaphorique ou déictique de certaines formes linguistiques, en l'occurrence les expressions démonstratives? Et quelles sont les limites des principes généraux de cohésion textuelle appliqués à la résolution des anaphores impliquant des expressions démonstratives?

2.2 Questions de recherche secondaires

Pour répondre correctement à la question de recherche principale, il est nécessaire d'en développer certains aspects pour en arriver à une réponse plus précise. Voici donc quelques questions secondaires pertinentes pour cette étude.

Premièrement, il faut se demander dans quelles conditions les expressions démonstratives jouent-elles un rôle dans la cohésion textuelle. Dans quels cas sont-elles impliquées dans des phénomènes endophoriques et dans lesquels s'agit-il plutôt de déixis? Dans les cas d'endophore, comment localiser le ou les antécédents et sur quels indices devrait-on se baser pour ce faire? Il faut aussi se demander quelles relations les expressions démonstratives entretiennent avec leurs antécédents sur les plans lexical, sémantique et référentiel.

Deuxièmement, en ce qui concerne les approches employées en analyse automatique de texte pour résoudre les problèmes de cohésion textuelle, plusieurs questions se posent. Par exemple, quelles sont les limites des approches empiriques,

---

[3] Par **analyse automatique de textes**, nous entendons le sous-ensemble de la linguistique informatique qui s'intéresse à l'analyse (par opposition à la synthèse ou à la génération) de documents textuels écrits, plutôt qu'à la parole ou à la conversation. Il est toutefois à noter que plusieurs applications en linguistique informatique doivent à la fois analyser et synthétiser du texte. La traduction automatique en est un exemple.

par rapport aux approches théoriques? En particulier, de quels types de connaissances a-t-on besoin pour traiter les phénomènes endophoriques et déictiques mettant en jeu des expressions démonstratives? Est-ce les mêmes pour toutes les expressions démonstratives, ou y-a-t-il au contraire des différences importantes entre les diverses formes d'expressions démonstratives? Peut-on faire des recoupements entre le comportement de certaines expressions démonstratives et celui d'autres formes importantes pour la résolution des anaphores, comme les pronoms personnels ou les descriptions définies? Quelles sont les règles qui s'appliquent exclusivement aux démonstratifs, ou à certaines classes de démonstratifs?

Enfin, à la lumière des informations recueillies en corpus sur la fréquence des différents types d'expressions démonstratives, on sera en mesure de répondre à des questions importantes d'un point de vue pragmatique. La première question est celle de la proportion dans laquelle les diverses façons d'utiliser les démonstratifs, du point de vue de la cohésion textuelle, se retrouvent dans les textes pertinents pour l'analyse automatique de textes. On pourra alors se demander : quels sont les cas les plus importants à résoudre? Et aussi, quels sont les cas les plus difficiles à résoudre? Y a-t-il des cas où il est possible de prendre des «raccourcis»? Quelles sont les solutions les plus efficaces? Doit-on combiner différentes approches? Si oui, lesquelles?

Dans le chapitre suivant, nous essayerons de trouver un début de réponse à toutes ces questions dans l'examen de la littérature et des expériences antérieures.

Chapitre 3 Problématique

3.1 Approches générales de la résolution des anaphores

Plusieurs approches, s'appliquant à des niveaux d'analyse linguistiques divers, ont été employées pour la résolution des anaphores. Ces approches proviennent d'une multitude de théories, élaborées dans le cadre de disciplines et sous-disciplines variées. Après en avoir essayé plusieurs indépendamment, les chercheurs en sont venus à la conclusion qu'il valait mieux combiner les techniques les plus efficaces, en commençant par celles qui sont les plus faciles à mettre en oeuvre.

Cependant, pour faciliter la présentation de ces méthodes, l'exposé qui suit les abordera séparément. Les diverses façons d'exploiter les phénomènes de cohésion textuelle peuvent être par le niveau d'analyse auquel elles s'appliquent. Nous en ferons donc un survol en commençant par le niveau morphologique. Nous verrons ensuite successivement les niveaux lexical et syntaxique. Nous terminerons par le niveau discursif.

3.1.1 Approches opérant au niveau morphologique

Dans l'élaboration d'une application informatique, la première préoccupation est souvent l'efficacité. Dans cette perspective, la correspondance des catégories grammaticales nominales (par exemple, le genre et le nombre) est relativement simple à vérifier et permet souvent d'éliminer plusieurs des candidats possibles. Il suffit de déterminer, lorsque cela est possible[4], le genre et le nombre de chaque candidat, ceux de l'expression anaphorique et d'éliminer les candidats qui présentent une incompatibilité.

Voyons un exemple (inventé) :

Ex. 1 : *Les portes du camion de Johanne étant verrouillées,* ***celle-ci*** *sortit ses clés.*

---

[4] Il est impossible de déterminer le genre et le nombre d'un candidat antécédent lorsque ce candidat antécédent ne possède pas les catégories grammaticales en question. Par exemple, un syntagme propositionnel n'a ni genre, ni nombre. Cela ne pose pas nécessairement problème, car un tel syntagme peut être l'antécédent d'un pronom neutre, par exemple.

Supposons que l'on veuille trouver l'antécédent de *celle-ci*. Si l'on considère les syntagmes nominaux précédents en tant qu'antécédents possibles, ***les portes** du camion de Johanne* est éliminé parce qu'il est de nombre pluriel et ***le camion** de Johanne* est éliminé parce qu'il est de genre masculin. Il ne reste donc que *Johanne*. Notez qu'il n'est pas nécessaire de savoir que *Johanne* est un prénom féminin.

Cette démarche est le parfait exemple de l'application de ce qu'on appelle une contrainte. Cela résulte de la règle logique suivante :

> $P \rightarrow Q \equiv \neg Q \rightarrow \neg P$
> 'Si une proposition P implique une proposition Q; alors, la négation de la proposition Q implique la négation de la proposition P; et vice versa.'

Autrement dit, s'il est vrai que la présence d'un lien anaphorique implique l'accord en genre et en nombre, alors l'absence d'accord en genre et en nombre implique l'absence de lien anaphorique.

Notons toutefois que pour certains types de liens anaphoriques, il s'avère que l'accord en genre et en nombre n'est pas obligatoire. Pour le démonstratif, cela se produit de la manière la plus évidente dans les cas de reclassification.[5] Les deux exemples ci-dessous[6] en sont une illustration. Dans l'exemple 2, l'antécédent est au pluriel alors que l'expression anaphorique est au singulier. Dans l'exemple 3, c'est le genre qui varie.

> Ex. 2 : *De la même façon, les anciennes usines d'Alcan à Shawinigana ont été intégralement rénovées pour accueillir depuis 2003 des expositions internationales d'art moderne et d'art contemporain, organisées par le Musée des beaux-arts du Canada.* [...] *C'est dans **cette ancienne aluminerie** que le 20 octobre 1901 fut coulé le premier lingot d'aluminium en sol canadien.* (Voir-3)

> Ex. 3 : *Cette fois, la Semaine de mode se déroulera au coeur du Quartier international, sous un chapiteau spécialement dressé pour l'occasion.* [...] *Depuis quatre ans, **cet événement montréalais** offre une vitrine nécessaire aux créateurs québécois, qui ont là une occasion de présenter leurs collections regroupées à un public spécialisé.* (Voir-9)

Soulignons que le fait que l'accord en genre et en nombre ne soit pas obligatoire dans tous les cas d'anaphore n'enlève rien à la valeur de cette contrainte.

---

[5] Au sujet de la capacité de reclassification du démonstratif, se référer à la page 17 de ce mémoire.
[6] Ces exemples sont tirés de notre étude de corpus. Les références complètes sont disponibles à l'annexe 6.

Elle reste valable dans bien des cas. Cela montre seulement qu'il faut l'utiliser avec discernement.

Mis à part les contraintes, un autre type de règle peut être utilisée : les préférences. Contrairement à ce qui se passe dans le cas d'une contrainte, aucun candidat n'est éliminé par l'application d'une préférence. Si une préférence est satisfaite par un candidat antécédent, on considérera que la probabilité de ce candidat sera plus grande. L'algorithme devra ensuite tenir compte de cette donnée, par exemple en calculant un score pour chaque candidat, score qui dépendra des préférences remplies par chacun. Pour plus d'information sur les notions de contrainte et de préférence, prière de se référer à Mitkov (2002:41-44).

Bien sûr, les catégories grammaticales à vérifier peuvent varier d'une langue à l'autre. De plus, une langue qui intègre davantage ces catégories au niveau morphologique donnera plus d'emprise à une contrainte de ce type. Il est donc à noter que puisque cette méthode est basée sur un facteur qui varie évidemment d'une langue à l'autre, son efficacité varie également d'une langue à l'autre.

Par exemple, le septième chapitre de Mitkov (2002) présente MARS, un système de résolution des pronoms *knowledge-poor*. Conçue à l'origine pour l'anglais, cette approche a été adaptée à d'autres langues. Mitkov (2002:173) considère que la version bulgare, notamment, fonctionne un peu mieux que la version anglaise, à cause de l'existence en bulgare des genres masculin, féminin et neutre et de l'importance du genre pour la résolution des pronoms dans cet algorithme.

3.1.2 Approches opérant au niveau lexical

Au niveau lexical, l'information lexico-sémantique, morphosyntaxique ou paralinguistique véhiculée par les mots-formes dont dépend syntaxiquement l'expression démonstrative, et par ceux qui dépendent d'elle, joue également un rôle dans la localisation de l'antécédent. (Il est ici question de relations de dépendance syntaxique entre mots-formes, et non pas entre constituants.) Autrement dit, les mots-formes voisins du démonstratif ou de l'expression démonstrative dans le texte donnent des indices pouvant mener à l'identification du référent de l'expression

démonstrative. Les sections 3.1.2.1 et 3.1.2.2 développeront davantage ce concept, en présentant les propos de deux auteurs : Cornish (1998) et Tutin (1992).

À titre d'illustration, on observe que dans l'exemple 1, à la page 5, le syntagme verbal *sortit ses clés* permet de conclure que le référent du démonstratif *celle-ci* est une personne. Cela élimine les candidats *les portes du camion de Johanne* et *camion de Johanne*, laissant seulement *Johanne*.

Dans le cas des syntagmes nominaux dotés d'un déterminant démonstratif, on retrouve également cette sorte d'indicateur à l'intérieur du syntagme constitué par l'expression démonstrative. Un cas trivial est celui de l'exemple 4, à la page 8, où le nom *administration*, compris dans l'expression démonstrative *cette administration*, permet de localiser l'antécédent de cette dernière, le syntagme *l'administration d'ecstasy*.

3.1.2.1 LE SEGMENT INDEXICAL

Selon Cornish (1999:98), on trouve un grand nombre d'indices pouvant mener à l'identification du référent dans la proposition anaphorique (la proposition contenant l'anaphorique). Ces indices restreignent le type et la nature du référent d'une expression anaphorique. Ces informations s'ajoutent à celles fournies par l'anaphorique lui-même. Cornish (1999:98) utilise l'expression *indexical segment* («segment indexical» [N.T.]) pour désigner l'entité formée par l'environnement lexical immédiat de l'anaphorique et l'expression anaphorique elle-même.

3.1.2.2 LES FONCTIONS LEXICALES APPLIQUÉES À L'ANAPHORE

D'un autre côté, Tutin (1992) applique le formalisme des fonctions lexicales (FL) à la caractérisation de certains liens anaphoriques. Pour une introduction aux fonctions lexicales, prière de se référer à Mel'čuk et al. (1995:125-151), par exemple. Voyons tout d'abord la définition d'un type de lien anaphorique, l'anaphore lexicale coréférentielle, où le mécanisme des FL s'illustre aisément (Tutin, 1992:59).

> *Un élément textuel est une* ***anaphore lexicale coréférentielle*** *d'un antécédent si :*
> *1) cet élément a le même référent que son antécédent,*
> *2) cet élément appartient à une classe lexicale ouverte,*

3) l'anaphore et son antécédent partagent une composante sémantique non triviale.

Notez que Tutin (1992) utilise ici le terme **anaphore** pour désigner un élément textuel, alors que nous préférons réserver ce terme pour désigner le phénomène ou le lien. Dans cet exposé, l'élément textuel en question sera appelé **expression anaphorique** ou, plus simplement, **anaphorique**.

Les fonctions lexicales peuvent être utilisées pour formaliser cette «composante sémantique» qui dépend des lexies utilisées. Tutin (1992:60) commence par montrer que les fonctions lexicales couvrent les cas que Halliday & Hasan (1976) regroupent sous le terme de «réitération» : répétitions, synonymes (FL Syn), synonymes plus larges (FL $Syn_{\supset}$) et hyperonymes (FL Gener).

L'exemple 4 illustre un cas de répétition partielle :

Ex. 4 : *Pour tenter de répondre à cette question, l'équipe de Sylvie Chalon et Laurent Galineau (INSERM U619, Université François Rabelais à Tours) a étudié les conséquences de l'administration d'ecstasy à des femelles de rat gestantes sur le fœtus à naître. Et sur le rat,* ***cette administration*** *n'est pas sans conséquence.*[7]

La coréférence lexicale peut aussi être réalisée par la verbalisation (V_0) d'un nom ou la nominalisation (N_0) d'un verbe. Souvent, ces opérations s'accompagnent de processus relevant de l'anaphore grammaticale (pronominalisation ou ellipse). L'adjectivation (A_0) d'un nom peut aussi être utilisée, mais un substantif doit alors être ajouté à l'adjectif obtenu pour former un syntagme nominal autonome.

Ces anaphores lexicales coréférentielles, où «l'antécédent correspond au mot-clé de la FL» et où l'anaphorique correspond à une valeur de cette FL, sont qualifiées de «directes» par Tutin (1992:63). Elle y oppose les anaphores coréférentielles lexicales indirectes, où la relation de coréférence est établie «entre un actant ou un circonstant du mot-clé et la valeur de la FL». Pour distinguer les deux types «d'antécédent», impliqués dans une anaphore de ce type, Tutin (1992:63) emploie les termes de «coréférent» et «d'antécédent sémantique», respectivement.

Les fonctions lexicales permettant la coréférence lexicale indirecte diffèrent de celles utilisées pour la coréférence lexicale directe. Il s'agit des noms typiques

---

[7] PÉTRY, Françoise (2005). «De l'ecstasy chez les petits rats», in *Pour la Science*, no 330, avril 2005.

d'actants (S_i, où i est l'indice actantiel), FL applicable autant à des noms qu'à des verbes, mais n'ayant que des substantifs comme valeur. De même, les FL de noms typiques de circonstants de lieu (S_{loc}), d'instrument (S_{instr}), de moyen (S_{med}) et de manière (S_{mod}) produisent des substantifs qui peuvent donner lieu à reprises anaphoriques indirectes. Le nom de résultat typique (S_{res}) apparaît aussi dans une fonction similaire; toutefois, on ne peut parler dans ce cas, selon Tutin (1992:64), de coréférence au sens stricte, puisqu'il y a forcément eu, alors, transformation du référent.

La **compositionnalité** est une propriété importante des fonctions lexicales. Par exemple, en utilisant une composée de fonctions telle que Gener°Gener[8], on obtient un hyperonyme très générique. Notons que cet exemple de composition en particulier pourrait être utile pour décrire l'usage de certaines expressions démonstratives, du même type que celle de l'exemple 5. Tutin (1992:66) note aussi que certaines compositions de FL sont réductibles (par exemple : Syn°Gener est équivalent à Gener), mais nous ne nous attarderons pas sur ce point plutôt technique pour notre propos.

> Ex. 5 : *Le gouvernement* ***vendit*** *la société d'état.* ***Cette action*** *fut fort critiquée.*[9]

Tutin (1992:69-73) parle aussi des emplois non coréférentiels, c'est-à-dire de ce qu'on appelle des **anaphores associatives**. On peut faire un parallèle avec la coréférence lexicale indirecte, car dans les deux cas, il y a présence d'un «antécédent sémantique», excepté que dans les cas d'association, il n'y a pas mention d'un coréférent. Il ne faut donc pas se surprendre si Tutin (1992:71) donne comme FL impliqués dans ce phénomène les noms typiques d'actants et de circonstants, les mêmes FL que celles citées pour la coréférence indirecte. Nous pensons d'ailleurs que la frontière entre les deux est parfois difficile à établir. Par exemple, il arrive qu'un coréférent soit éloigné de l'anaphorique dans le texte, mais qu'un antécédent sémantique soit tout proche. Comme Tutin (1992:69) exclut d'emblée les démonstratifs de l'anaphore associative (voir la citation en bas de la page 20), ses

---

[8] Où «°» est l'opérateur de composition de FL, tel que $FL_1°FL_2(\textit{mot-clé}) = \textit{valeur}$ est équivalent à $FL_1(FL_2(\textit{mot-clé})) = \textit{valeur}$.

[9] Cet exemple a été inventé par nous.

exemples présentent seulement des cas où les expressions anaphoriques sont introduites par des articles définis et indéfinis.

Tutin (1992:68) remarque enfin que «toutes les relations coréférentielles ne sont pas formalisables en termes de FL». Elles peuvent relever de connaissances extra-linguistiques, de figures de style, de prototypes, etc. De plus, plusieurs relations tout de même sémantiques ne correspondent à aucune FL. Ces relations sémantiques peuvent par contre se retrouver dans la définition même de la lexie. Par exemple, on peut définir la «**main**» comme «une partie du **bras**...» (Tutin, 1992:73), ce qui introduit une relation de type «partie-tout» entre ces deux lexies.

3.1.2.3 LISTES DE TERMES FRÉQUENTS PAR DOMAINE

Une manière d'exploiter les particularités d'un type de texte donné est d'extraire une liste des termes fréquemment utilisés pour référer (ou désigner) à des thèmes ou des entités importantes dans le domaine. Cette liste sert ensuite à suggérer des candidats antécédents / coréférents. Par exemple, Boudreau (2004:65) mentionne que certains noms communs se retrouvent plus souvent dans les chaînes de coréférence. La liste de ces noms dépend du type de texte. Cette idée fait d'ailleurs penser à la notion de «champ lexical», utilisée en analyse littéraire. Une typologie fonctionnelle est ici utilisée pour classer les texte du corpus exploité en trois catégories : articles de journaux portant sur des fusions de compagnies, critiques de cinéma et documentation de type «How-To» pour le système d'exploitation Linux (en informatique). Boudreau (2004:65) donne quelques exemples de «noms communs associés au domaine» :

> *Fusion de compagnie : mariage, société, compagnie, conseil;*
> *Critiques de films : mère, acteur;*
> *HOWTO Linux : fichier, disque, serveur.*

3.1.3 Approches opérant au niveau syntaxique

Les algorithmes basés sur des règles opérant sur la structure syntaxique, lorsque l'on dispose de cette information, sont elles aussi efficaces pour restreindre le champ des possibilités. Basés sur certaines théories développées au sein de la grammaire générative, comme la théorie «Gouvernement-Liage», ces algorithmes utilisent des règles qui permettent d'éliminer les cas où la coréférence est impossible

entre deux éléments à cause de la relation syntaxique dans laquelle ils se trouvent, directement ou indirectement.

3.1.3.1 L'ALGORITHME DE HOBBS

L'approche syntaxique «naïve» de Hobbs (1976, 1978) est un exemple classique d'algorithme de ce type. Mitkov (2002:72-77) rapporte qu'elle est encore utilisée de nos jours comme référence pour évaluer les performances des algorithmes résultant de la recherche actuelle. L'algorithme syntaxique de Hobbs se concentre sur la résolution des pronoms en anglais. Il se base pour cela sur un arbre syntagmatique de surface de chaque phrase du texte. Ces arbres syntagmatiques comprennent des éléments élidés tels que des anaphores-zéro (ellipses) et des antécédents-zéro. L'algorithme parcourt l'arbre à la recherche d'un syntagme nominal (SN) de genre et de nombre appropriés. Il commence par la zone limitée où devrait être utilisé un pronom réfléchi. Ensuite, il parcourt le reste de la phrase courante, puis les phrases précédentes, en commençant par la plus proche.

Toujours selon Mitkov (2002:75), l'algorithme de Hobbs adopte deux contraintes syntaxiques empruntées à Langacker (1969). La première vise à tirer parti de l'usage complémentaire des pronoms réfléchis vis-à-vis des autres pronoms. Elle stipule que ces derniers ne peuvent apparaître dans la même proposition que leur antécédent. La deuxième règle est formulée de manière à tenir compte des cas de cataphore comme ceux d'anaphore stricto sensu. Selon cette règle, «the antecedent of a pronoun must precede or command the pronoun.» (Mitkov 2002:75) (« l'antécédent d'un pronom doit soit le précéder, soit le commander » [N.T.]) Le terme **commander** est ici défini comme suit :

> *A node NP_1 is said to command node NP_2 if neither NP_1 nor NP_2 dominates the other and if the S node which most immediately dominates NP_1 dominates but does not immediately dominate NP_2.*
> (Un nœud SN_1 commande un autre nœud SN_2 si aucun des deux ne domine l'autre et si le nœud P (proposition ou phrase) qui domine le plus immédiatement SN_1 domine non immédiatement SN_2. [N.T.])
> (Mitkov 2002 : 75)

Rappelons que si l'anaphore est un phénomène qui se produit quand un segment textuel dépend, pour son interprétation, d'un autre segment situé devant lui

dans le texte, alors la cataphore est simplement le phénomène inverse. C'est-à-dire que la mention dont dépend l'expression cataphorique est situé postérieurement dans le texte, au lieu d'antérieurement. L'anaphore et la cataphore sont collectivement appelés **endophore**, quoiqu'on parle souvent d'anaphore en incluant les deux phénomènes. C'est probablement parce que l'anaphore est beaucoup plus courante que sa cousine, la cataphore.

La figure 1 illustre le fonctionnement de l'algorithme de Hobbs. Elle est tirée de Mitkov (2002:74). On cherche ici à trouver l'antécédent du pronom *it*. L'algorithme commence au nœud NP_1. La ligne pointillée indique le chemin emprunté par l'algorithme pour parcourir l'arbre syntagmatique. Celui-ci s'arrête finalement au nœud NP_4, qui représente le syntagme *the residence of the king*.

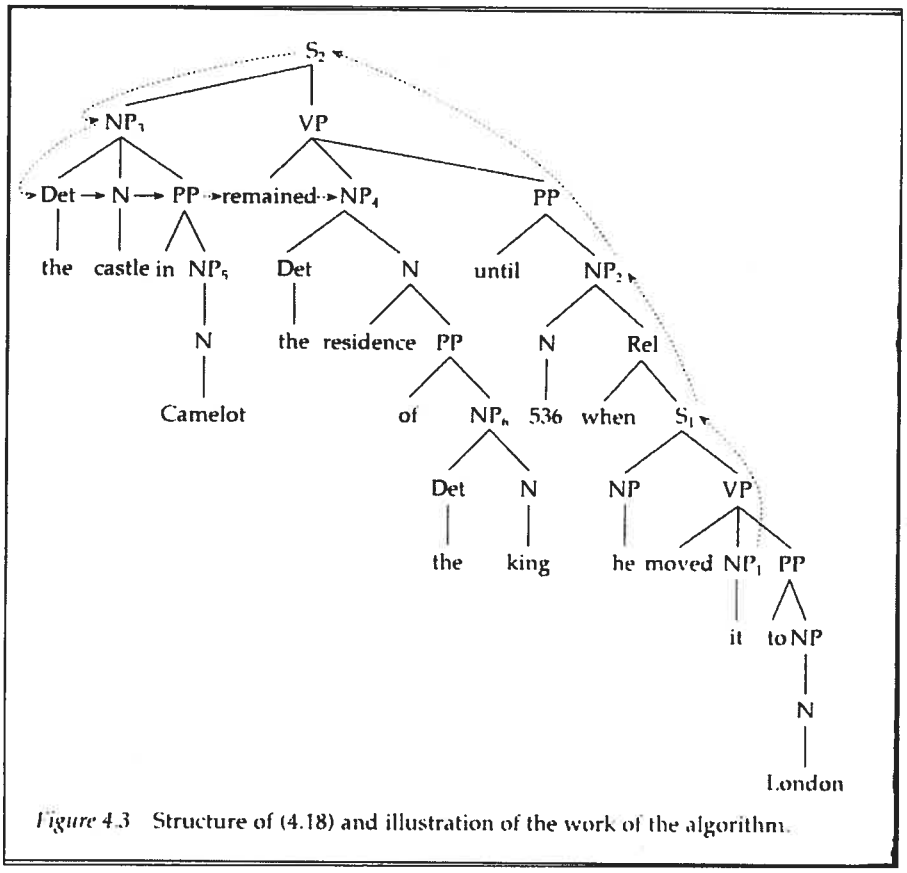

Figure 4.3 Structure of (4.18) and illustration of the work of the algorithm.

Figure 1 Exemple d'application de l'algorithme de Hobbs

Veuillez noter que certaines «*selectional constraints*» (Mitkov, 2002:74, «contraintes de sélection» [N.T.]) sont utilisées pour éliminer certains candidats suggérés par l'algorithme. L'algorithme poursuit ensuite son parcours jusqu'à ce

qu'un autre candidat adéquat soit trouvé. Par exemple, selon Mitkov (2002:74), le syntagme *the castle in Camelot*, identifié par le nœud NP_3, est rejeté par une contrainte de sélection telle que «*places can't move*» ou «*large objects don't move*» («les lieux ne peuvent pas bouger», «les gros objets ne bougent pas» [N.T.]).

3.1.3.2 LA C-COMMANDE

On pourrait également trouver d'autres formulations de ces contraintes sur l'antécédent des pronoms réfléchis et non réfléchis définis en fonctions d'arbres de constituants syntaxiques. Parmi, celles-ci, citons la fameuse «c-commande». On trouve dans Reinhart (1983:18) cette définition simplifiée de la relation structurelle de «*c-command*» («*constituent-command*») :

> *Node A c(constituent)-commands node B iff the branching node most immediately dominating A also dominates B.* (Le nœud A c-commande le nœud B ssi [si et seulement si] le nœud branchant dominant le plus immédiatement A domine aussi B. [N.T.])

Reinhart (1983:19) développe ensuite la notion de «*domain*» («domaine» [N.T.]) pour pouvoir comparer les régions syntaxiques circonscrites par les ensembles de nœuds en relation de «c-commande» et d'autres relations concurrentes. Par exemple, elle arrive à la conclusion que le sujet d'une phrase simple aura pour domaine l'ensemble de cette phrase, à la fois pour la relation de «c-commande» et pour celle connue sous le nom de «*precede-and-command*».

> *The domain of a node A consists of all and only the nodes c-commanded by A.* (Le domaine d'un nœud A consiste en l'ensemble des nœuds c-commandés par A. [N.T.]) (Reinhart, 1983:19)

Poursuivant l'analyse, Reinhart (1983:22) conclut que le «domaine» défini par la «c-commande» se réduit à : «*for any given node the minimal branching constituent that contains it*» («pour un nœud donné, le plus petit constituant branchant qui le contient» [N.T.])

Enfin, Reinhart (1983:30) en arrive à définir une contrainte sur la coréférence des syntagmes nominaux en fonction de la notion de «domaine» : «*A given NP must be interpreted as non-coreferential with any non-pronoun in its domain.*» («Un SN donné doit être interprété comme étant non coréférentiel avec tout non-pronom se trouvant dans son domaine.» [N.T.]) Elle ajoute qu'une autre règle est nécessaire

pour couvrir les cas où un pronom non réfléchi est situé dans un «*reflexivisation environment*». Cette expression fait référence à la zone restreinte où un pronom réfléchi doit être utilisé, s'il y a coréférence avec le sujet d'un verbe.

3.1.3.3 LA DISJONCTION RÉFÉRENTIELLE

Également en ce qui a trait à la structure syntaxique, Corblin (1995:187-191) formule des recommandations sur la manière dont devrait se faire la résolution automatique des anaphores. Selon lui, celle-ci devrait se faire par «lots» d'expressions anaphoriques se trouvant dans un même «domaine syntaxique» (Corblin, 1995:188). La «disjonction référentielle» (Corblin, 1995:191), c'est-à-dire l'impossibilité pour deux expressions situées dans le même domaine syntaxique de coréférer, permettrait alors de réduire progressivement le nombre de candidats antécédents. On procéderait ainsi par élimination, en commençant par l'anaphore la plus facile à résoudre.

3.1.4 Approches opérant au niveau discursif

Au-delà des niveaux morphologiques et syntaxiques, plusieurs chercheurs ont travaillé à mettre au point des théories qui tiennent compte de l'enchaînement des sujets d'une phrase à l'autre, à l'intérieur de textes suivis ou d'échanges entre locuteurs.

Mitkov (2002:53) rapporte que la théorie du «centering» (Grosz et al. 1983; Grosz et al. 1995) porte sur l'idée que certaines des entités mentionnées dans un texte sont plus importantes que d'autres et que cela crée des contraintes sur le choix des expressions référentielles. Concrètement, ces contraintes s'appliquent plus particulièrement aux référents introduits par un syntagme nominal et repris par un pronom.

3.1.5 Synthèse

En fait, c'est une synthèse de toutes ces approches qui est visée de nos jours. Du point de vue de l'informatique, les approches empiriques restent les plus simples à implémenter, car elles exploitent ce qui est la véritable force de l'ordinateur : effectuer un grand nombre d'opérations simples à très grande vitesse. Mais il est

possible d'y ajouter certaines approches théoriques qui peuvent se réduire également à des calculs individuellement peu complexes.

3.2 Caractéristiques du démonstratif en cohésion textuelle

Toutefois, bien que les méthodes utilisées à ce jour permettent de résoudre bien des cas, pour pouvoir progresser, il faut maintenant faire un pas de plus du général au spécifique. Plus précisément, nous nous penchons ici sur le (ou les) démonstratif(s).

3.2.1 Le démonstratif en général dans la cohésion textuelle

Tout d'abord, nous nous pencherons sur ce qui distingue le démonstratif en général des autres moyens linguistiques qui sont utilisés dans les processus référentiels, anaphoriques ou déictiques. Ensuite, nous verrons comment on peut dégager des classes de démonstratifs et les particularités de chacune d'elles, de même que les inévitables exceptions et cas particuliers.

3.2.1.1 DÉFINITIONS DU DÉMONSTRATIF ET DE L'EXPRESSION DÉMONSTRATIVE

Voyons d'abord ce que nous entendons par **démonstratif**. Il s'agit des pronoms démonstratifs, des déterminants démonstratifs (aussi appelés traditionnellement «adjectifs démonstratifs») et des adverbes démonstratifs.

Par **expression démonstrative**, nous entendons les syntagmes nominaux constitués d'un pronom démonstratif seul ou accompagné d'un modificateur, de tout syntagme nominal dont le déterminant est un déterminant démonstratif, ainsi que des adverbes démonstratifs seuls. (Par opposition aux adverbes démonstratifs clitiques «-ci» et «-là» qui sont toujours adjoints à un autre démonstratif.)

3.2.1.2 CARACTÉRISTIQUES DU DÉMONSTRATIF DANS LA COHÉSION TEXTUELLE

Ce qui caractérise les démonstratifs, ce sont la façon dont ils saisissent leur référent, le(s) type(s) d'antécédent qui leur sont associés, la distance à l'antécédent et le(s) type(s) de relation(s) anaphorique(s) qu'ils peuvent entretenir.

Le mode de saisie du référent

Plusieurs auteurs ont fait des recherches sur la manière dont les démonstratifs fonctionnent pour saisir leur référent. Corblin (1995:78) écrit : «Le démonstratif saisit son référent par emprunt à une désignation (ou pointage) identifiable en vertu de critères externes (proximité, saillance).» Il parle ici du *déterminant* démonstratif, mais cela s'applique probablement encore mieux au pronom démonstratif, puisque celui-ci est dépourvu, ou presque, de contenu «descriptif», élément dont l'utilisation caractérise le mode de saisie du référent du défini. Même s'il n'a presque pas de contenu descriptif associé, l'environnement immédiat du pronom démonstratif peut influencer le choix de son référent de manière similaire à la description qui accompagne le déterminant. Cornish (1999:98) réfère à cette notion sous le nom de «segment indexical».

Pourtant, bien que le démonstratif possède plus de liberté par rapport à son contenu descriptif, il peut tout de même utiliser ce moyen. Dans ce cas, le démonstratif peut avoir une fonction constrastive, auquel cas il semble plus adapté que le défini. En guise d'illustration, voici un exemple tiré de Tutin (1992:34) :

Marguerite a rencontré un charcutier. Ce / ? Le charcutier est français.

Kleiber (1986:175-176) (cité dans Tutin (1992:36)) trouve quant à lui que le démonstratif est le «connecteur anaphorique» par excellence. En effet, selon lui, il désigne directement son objet de référence, sans avoir besoin de vérifier une propriété, dans une circonstance d'évaluation donnée, comme doit le faire le défini.

C'est cette propriété du démonstratif de désigner directement, au besoin, son référent, qui lui permet d'adopter, au niveau conceptuel, une fonction de classification, de reclassification ou même de déclassification. Au niveau textuel, cette même caractéristique lui permet d'opérer un changement de point de vue sur un sujet déjà abordé, de dégager un sujet d'une situation où il n'était, au mieux, qu'implicite, ou même de redonner un caractère indistinct à un sujet antérieurement défini selon un certain point de vue.

Le type d'antécédent

Tout ceci fait que le démonstratif, quand il joue un rôle anaphorique, a tendance à avoir des antécédents d'un type différent. Dans les travaux sur la

résolution des anaphores, le cas le plus simple, peut-être le plus courant, et celui sur lequel on se consacre principalement, est celui où l'anaphorique est un pronom personnel de 3^{e} personne et où l'antécédent est un syntagme nominal, souvent coréférent. C'est possiblement l'archétype de l'anaphore. Cornish (1986:7) mentionne :

> *[N]ominal pronominal anaphora involving a relation of coreference [...] is most frequently taken as representative of anaphora as a whole in the majority of studies of the phenomenon.*
> («L'anaphore nominale pronominale [(où un nom ou SN est l'antécédent d'un pronom)] impliquant une relation de coréférence est le plus fréquemment considérée comme représentative de l'anaphore en entier dans la majorité des études de ce phénomène.» [N.T.])

Or, ce qui différencie le démonstratif d'autres moyens linguistiques comme le pronom personnel et le SN introduit par un article défini, c'est sa plus grande propension à avoir des antécédents qui ne sont pas des syntagmes nominaux. On retrouve parmi ceux-ci des verbes, des syntagmes verbaux, des propositions, des phrases et même des paragraphes en entier.

Cornish (1986:8-18), pour illustrer les divers types d'anaphores grammaticales, énumère toute une série de syntagmes pouvant faire office d'antécédent : nom (sans déterminant), syntagme nominal, verbe, syntagme verbal (au sens de prédicat syntaxique), phrase (au sens de proposition syntaxique), syntagme prépositionnel, syntagme adjectival. Il fait également remarquer que c'est moins la forme syntaxique de surface de l'antécédent qui importe que sa valeur sémantique en contexte. Corblin (1995:175), quant à lui, met de l'avant que «l'hétérogénéité catégorielle» n'est pas un obstacle à la formation de liens anaphoriques ou coréférentiels : «c'est ce qu'on observe quand un pronom reprend un énoncé entier ou même une suite d'énoncés.»

Voici quelques exemples d'expressions anaphoriques démonstratives ayant un (ou des) «antécédent» («coréférent» ou «antécédent sémantique») non nominaux. Ils sont tirés du magazine de vulgarisation scientifique *Pour la Science*.

> Ex. 6 : *Ils seraient <u>anhédoniques</u>, c'est-à-dire indifférents au plaisir. Leur système de la récompense, celui qui libère de la dopamine, l'hormone du plaisir, fonctionne peut-être en « sous-régime ».* ***Cette anhédonie*** *résulte-t-*

elle du traitement subi par leur mère ?[10]

Ex. 7 : *Chez le rat nouveau-né dont la mère n'a pas reçu d'ecstasy, on constate un pic important de la concentration en sérotonine juste après la naissance. En revanche, chez le nouveau-né d'une mère ayant reçu de l'ecstasy, l'augmentation de la concentration en sérotonine est très faible.* ***Ce premier résultat*** *semble indiquer un dysfonctionnement du système sérotoninergique.*[11]

Ex. 8 : *Par quel prodige ?*
C'*est la question que s'est posée Claude Cyr, de la Faculté de médecine de l'Université de Sherbrooke, au Québec.*[12]

Dans l'exemple 6, le nom utilisé dans l'expression démonstrative, «anhédonie», est une nominalisation de l'adjectif «anhédoniques» rencontré antérieurement. En y regardant d'un peu plus près, il semble que «cette anhédonie» pourrait référer à la proposition exprimée par la phrase précédente, ou encore à l'ensemble du concept et de l'explication qui en est donnée dans les deux phrases précédentes.

Dans l'exemple 7, «ce premier résultat» réfère à l'ensemble de ce qui est expliqué dans les deux phrases précédentes et qu'on peut identifier comme étant la description des résultats d'une expérience. L'utilisation du nombre singulier regroupe ces renseignements sous la forme d'un tout, pour mieux les replacer dans un ensemble d'objets similaires dont l'adjectif «premier» laisse entendre que ce n'en est que le début.

Dans l'exemple 8, le SN «la question», parce qu'il est l'attribut du pronom «c'», aide clairement à déterminer que ce pronom réfère à la phrase interrogative précédente, si l'on l'interprète au sens littéral. Si l'on considère plutôt qu'il s'agit ici d'indiquer, par une tournure intéressante, le sujet de recherche de l'individu nommé «Claude Cyr», on dira plutôt que le pronom «c'» coréfère avec la proposition exprimée par la phrase précédente. Dans tous les cas, la première interprétation facilite l'évocation de la deuxième.

---

[10] PÉTRY, Françoise (2005). «De l'ecstasy chez les petits rats», in *Pour la Science*, No 330, avril 2005
[11] Ibid.
[12] GROUSSON, Mathieu (2005). «Surprenante jeunesse», in *Pour la Science*, No 330, avril 2005

Parmi les antécédents qui ne sont pas des syntagmes nominaux, le plus important est probablement l'antécédent propositionnel. Celui-ci peut être explicite, mais il peut également être implicite. Cornish (1986:126) considère que CELA et ÇA peuvent entretenir une relation anaphorique avec «*an entity discourse referent (a pre-supposed or asserted proposition) without the necessary mediation of a textually co-present or inferrable linguistic antecedent-trigger.*» («un référent discursif de type «entité» (une proposition présupposée ou énoncée) sans que soit nécessaire la médiation par un antécédent présent ou déductible du co-texte.» [N.T.]) L'exemple 9 montre un pronom CELA qui a comme antécédent la proposition infinitive «[je][13] être plus transparente et vertueuse».

> Ex. 9 : *Je ne peux pas me soigner moi-même. Je veux être plus transparente et plus vertueuse, mais pour* ***cela*** *j'ai besoin de balises et d'être encadrée.*[14]

La distance à l'antécédent

Une autre tendance du démonstratif est la distance en moyenne plus courte que celui-ci entretient avec son antécédent, par rapport aux anaphores où intervient l'article défini. Selon Mitkov (2002:18), la distance à l'antécédent, pour les descriptions définies (c'est-à-dire les syntagmes nominaux dotés d'un article défini), est plus grande que pour les syntagmes nominaux dotés d'un déterminant démonstratif. Cela tient apparemment de la fonction reclassifiante du démonstratif et de la manière plus directe qu'il a de désigner (selon Kleiber, 1986:175-176), ainsi que de la plus grande dépendance qu'il a sur des critères de proximité (selon Corblin, 1995:78).

Le type de relation anaphorique

Également, le type de relation anaphorique qui est privilégié est plus direct, c'est-à-dire que les relations anaphoriques coréférentielles sont plus fréquentes chez les démonstratifs. Certains auteurs affirment qu'il ne peut pas être impliqué dans une anaphore dite associative (et/ou indirecte). Tutin (1992:69) mentionne que «s'il n'y a pas de coréférent, mais une association sémantique (cas de l'ensemble des

---

[13] Le sujet implicite de cette proposition est 'je'. Le sens de l'expression «cela» inclut cette idée.
[14] *Quartier Libre*. 22 mars 2006. vol. 13, no 14.

associations), le démonstratif n'apparaît pas employable, à la différence du défini». Elle cite ensuite l'exemple suivant (Tutin 1992:70) :

> Ex. 10 : *Marguerite donnait une conférence. * **Ce / le** sujet portait sur l'acquisition du langage.*

Apothéloz & Reichler-Béguelin (1999) se sont penchés sur des cas où cette règle semble faire exception.

Plusieurs auteurs classifient les expressions démonstratives comme pouvant seulement être anaphoriques dans le cas de l'anaphore directe. Cependant, Apothéloz & Reichler-Béguelin (1999:370-378) montrent qu'en français, les expressions démonstratives, dans certains cas, sont employées en tant qu'anaphores associatives. La définition de l'anaphore associative employée ici est celle d'une anaphore non coréférentielle, c'est-à-dire dont le contexte ne mentionne pas le référent, mais où le référent est déduit grâce à l'information fournie par le contexte.

Voici quelques exemples rapportés par Apothéloz & Reichler-Béguelin (1999:366 et 371) :

> Ex. 11 : *Un gros chat blanc, qui appartient au jardinier, sauta sur mes genoux, et, de **cette secousse**, ferma le livre que je posai à côté de moi pour caresser la bête.*
> (G. de Maupassant, Sur les chats. In : Contes fantastiques. Paris : Marabout, 1992 : 241)

> Ex. 12 : *Mais quand se décidera-t-on à imposer les gros revenus et fortunes des multi-millionnaires comme sont imposés les petits revenus qui le sont jusqu'au dernier franc?*
> *Exemple : si l'on percevait un impôt de 700,000 francs (commune, canton et Confédération) sur un revenu d'un million (il en est qui 'gagnent' encore plus) **ce contribuable** aurait à disposition encore 300'000 francs.*
> (L'Impartial, 27.12.1993)

> Ex. 13 : *Le Marquis de Cuevas avait épousé la petite-fille de Rockefeller. Avec **cet argent**, il a créé un ballet.*
> (Radio, France-Musique, 7.2.1993)

Apothéloz & Reichler-Béguelin (1999:384-392) montrent aussi que les utilisations associatives des syntagmes démonstratifs peuvent remplir certaines fonctions au-delà de la fonction référentielle. Ils peuvent également résulter de la divergence d'intérêt communicatif entre le locuteur et l'interlocuteur.

3.2.2 Classes de démonstratifs et caractéristiques textuelles

Voyons maintenant quelles subdivisions nous pouvons faire parmi la classe des démonstratifs. Ici, ne perdons pas de vue que l'objectif que nous poursuivons est d'établir des divisions selon des critères formels, lexicaux et syntaxiques, susceptibles de coïncider avec des frontières au niveau du comportement en contexte de cette forme en ce qui concerne la coréférence et le comportement anaphorique ou déictique. De plus, ces deux types de classement doivent être décrites et pensées en des termes suffisamment précis et des critères suffisamment pragmatiques pour pouvoir présenter un intérêt pour l'analyse automatique de textes.

Nous devons donc faire correspondre deux ensembles de catégories. Le premier ensemble représente un classement selon l'interprétation que l'on doit donner à une expression démonstrative. Le deuxième ensemble réunit les informations dont on peut disposer au niveau formel, aux différents stades d'analyse. La figure 2 schématise le problème.

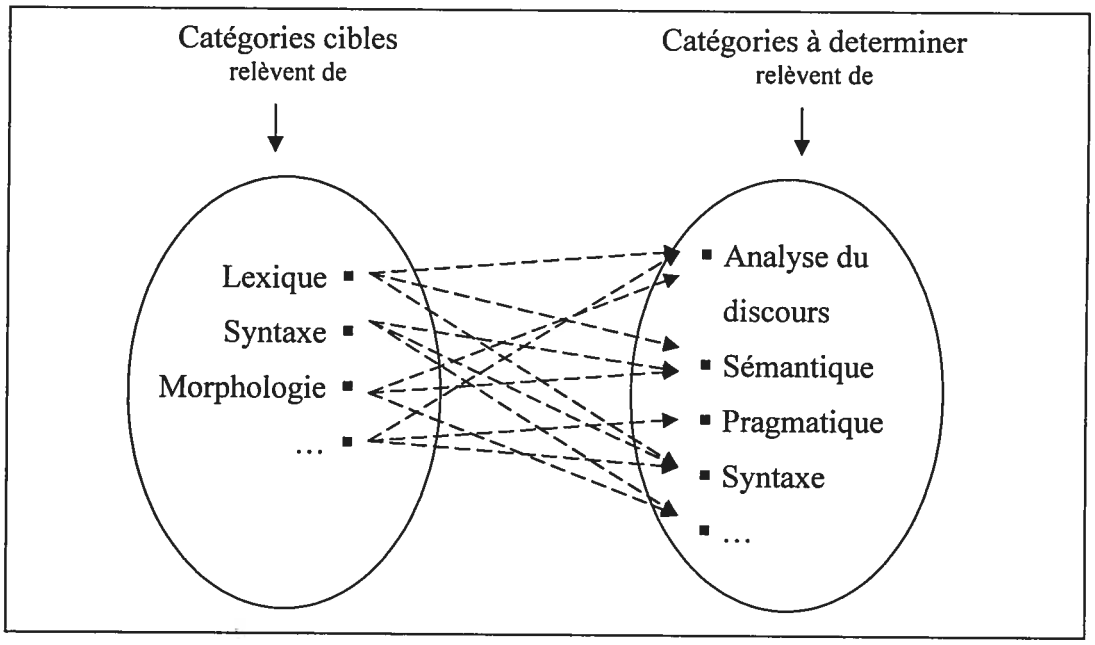

Figure 2 Schématisation du problème de la classification des démonstratifs

Tout d'abord, la division la plus évidente est celle basée sur la catégorie syntaxique : pronoms, déterminants, adverbes. Ensuite, à l'intérieur de la catégorie des pronoms, on retrouve, d'une part, ceux qui possèdent un genre masculin ou féminin et, d'autre part, les autres, qu'on qualifie de «neutres». Enfin, pour les

pronoms et les déterminants, on peut ranger d'un côté ceux à qui sont adjoints une particule adverbiale «-ci» ou «-là», et de l'autre, ceux qui n'en ont pas. Il faudra ajouter à cela les cas qui s'éloignent du comportement «par défaut» de ces formes, ou qui en forment une sous-classe au comportement particulièrement régulier. Cela peut être dû à ce qu'ils sont utilisés dans le cadre d'une locution ou d'une expression particulière, ou d'une construction syntaxique spécifique, comme dans certains emplois du pronom CE_1.[15]

3.2.2.1 LES PRONOMS DÉMONSTRATIFS DANS LA COHÉSION TEXTUELLE

Tout d'abord, les pronoms démonstratifs. Parmi ceux-ci, la première distinction à effectuer est entre les pronoms «neutre» et les pronoms de genre masculin ou féminin.

Les pronoms démonstratifs neutres : CE_1, CECI, CELA, ÇA. Le genre neutre n'existe pas pour la classe ouverte des substantifs en français contemporain. Cependant, certains pronoms peuvent être qualifiés de neutres parce qu'ils ne portent pas d'information sur le genre et qu'ils sont utilisés de préférence pour désigner des entités syntaxiques sans genre, comme des syntagmes non nominaux, ou des entités conceptuellement sans genre. Il ne faut pas les confondre avec les formes ambiguës où il est impossible de distinguer entre genre masculin et féminin, comme dans des formes plurielles telles que le déterminant «ces». Un cas similaire, hors du groupe des démonstratifs, serait le pronom clitique LE_2, comme dans la phrase : «Il faut **le** faire.» (Qu'on peut paraphraser comme «Il faut faire **quelque chose**.») À la différence, bien sûr, qu'il existe dans ce cas un LE_1 quasiment identique, mais de genre masculin.[16] Notez que l'on voit ici le parallèle entre le neutre grammatical et le «neutre» conceptuel. Ce dernier correspond à une «chose» quelconque, indéfinie. C'est l'objet par excellence, par opposition au sujet; ce dernier étant plus naturellement décrit comme masculin ou féminin.

---

[15] **CE_1** est un nom de lexie. Il s'agit du pronom démonstratif neutre atone. Prière de se référer à la «Liste des abréviations, des sigles et des symboles», pour l'explication de la notation des noms de lexie en général. Pour l'explication de la notation des lexies en CE, se référer au premier paragraphe du chapitre 5, page 41.

[16] **LE_1** et **LE_2** sont des noms de lexie. Ce sont tous les deux des pronoms personnels. Le premier est masculin et le deuxième est neutre. Prière de se référer à la «Liste des abréviations, des sigles et des symboles», pour l'explication de la notation des noms de lexie en général.

Corblin (1995:91-93) décrit les pronoms démonstratifs neutres comme étant «à contenu indistinct», comme étant des «formes déclassifiantes». Leur valeur neutre est un signe «qu'on est hors du système du Nom, système dans lequel toute unité doit prendre une de ces deux valeurs [(masculin ou féminin)], constitutive de son identité.»

Selon Riegel et al. (1999:206), CE est une forme neutre atone qui a trois usages : sujet clitique du verbe «être», support non-animé d'une proposition relative ou d'une proposition subordonnée interrogative portant sur le complément d'objet direct (COD). Quant à CECI, CELA et ÇA, malgré la graphie sans trait d'union, Riegel et al. (1999:206-207) les classe parmi les formes composées (ÇA étant, à l'origine, une forme contractée de CELA). Cela se comprend si l'on fait l'analogie avec les autres formes démonstratives, comme dans les exemples 14 à 16.[17]

Ex. 14 : ***Ce chapeau-ci*** *me plaît.*
Ex. 15 : ***Celui-ci*** *me plaît.*
Ex. 16 : ***Ceci*** *me plaît.*

Toujours selon Riegel et al. (1999:206-207), CECI, CELA et ÇA sont utilisés pour «désigner déictiquement des référents non-catégorisés», parfois même de «décatégoriser péjorativement un référent». En tant qu'anaphoriques, ils peuvent reprendre des propositions ou des segments textuels plus étendus, autrement dit des entités textuelles sans genre, ni nombre. Ils peuvent aussi reprendre un SN, souvent à interprétation générique, en en neutralisant le genre et le nombre.

Pour Corblin (1995:91), CECI, CELA et ÇA s'emploient plus naturellement avec des antécédents qui ne sont pas des «groupes nominaux au sens étroit», c'est-à-dire «pourvu[s] d'une (et d'une seule) tête lexicale *N*». En effet, un référent mentionné à l'aide d'un «groupe nominal au sens strict» se voit ainsi attribuer une étiquette (le nom *N*) et donc une catégorie. Cela revient donc à formuler en termes syntaxiques la notion sémantique de catégorisation. Les syntagmes nominaux dont la tête est un pronom, les coordinations de syntagmes nominaux et les autres types de syntagmes non nominaux, n'étant pas classifiés explicitement, seraient donc plus susceptibles d'être repris par un démonstratif de forme neutre composée.

---

[17] Les exemples 14 à 16 ont été inventés par nous.

Les pronoms démonstratifs non neutres se répartissent en deux groupes distincts : les formes simples (CELLE, CELUI) et les formes composées (CELLE-CI, CELLE-LÀ, CELUI-CI, CELUI-LÀ). Contrairement aux formes neutres, toutes sont variables et se déclinent au singulier et au pluriel.

Le comportement des formes simples est très différent de celui des formes composées. Selon Riegel et al. (1999:205), les formes simples «reprennent le contenu lexical et le genre d'un nom antécédent (ou de la forme lexicale associée à un référent présent dans la situation), mais en modifient le nombre et les déterminations à de nouvelles fins référentielles». Donc, contrairement à certaines formes d'anaphores pronominales, l'accord en nombre n'est absolument pas pertinent pour l'identification de l'antécédent de ce type de démonstratif. De plus, ce type de lien anaphorique fait appel à une identité des sens lexicaux, mais pas à une identité référentielle. Par contre, même si le pronom démonstratif non neutre simple en lui-même n'est pas coréférent avec son antécédent, l'ensemble de l'expression démonstrative en entier, composée du pronom et de son ou ses modificateurs, peut, parallèlement, participer à une relation anaphorique coréférentielle avec un antécédent identique ou différent.

Les formes simples non neutres du pronom démonstratif sont toujours accompagnées d'au moins un modificateur, qui sert à préciser la nouvelle référence à laquelle est maintenant associée le contenu lexical repris. Selon Riegel et al. (1999:205-206), ce modificateur peut-être une proposition relative, un complément propositionnel ou un participe avec sa complémentation.

Riegel et al. (1999:206) cite un autre usage pour CELUI et CELLE, celui de «support animé» à une relative périphrastique. Exceptionnellement, dans ce cas, il n'y a pas d'anaphore de type identité lexicale. Le sens du pronom démonstratif est alors simplement 'la personne' ou 'l'homme' pour CELUI, 'la femme' pour CELLE.

Corblin (1995:89-91) se sert de «l'opposition du saturé et du non saturé», qu'il considère comme «une notion primitive», pour décrire la différence entre les pronoms neutres CECI, CELA et ÇA, d'une part, et les pronoms non neutres IL, CELUI-CI et CELUI-LÀ, d'autre part. Selon lui, les premiers ont une signification

«saturée» mais «indistincte», alors que les seconds ont une signification «non saturée». Dans les deux cas, selon Corblin (1995:89), il y a absence de «Nom», ou d'«unité lexicale N». Toutefois, les pronoms masculins et féminins nécessitent que leur signification soit saturée par l'association à une «unité nominale N_i [...] empruntée au contexte» (Corblin, 1995:89). Quant aux pronoms neutres, ils conservent un «contenu nominal indistinct», même s'ils sont interprétés en rapport à un antécédent nominal (Corblin, 1995:89).

Selon Riegel et al. (1999:206), les «formes composées variables» du pronom démonstratif, c'est-à-dire CELUI-CI, CELUI-LÀ, CELLE-CI et CELLE-LÀ, obéissent aux mêmes principes que le déterminant démonstratif sur le plan des usages déictiques et anaphoriques. Contrairement aux «formes simples masculines ou féminines», elles ne sont jamais accompagnées d'un modificateur. Nous sommes porté à croire que c'est dans ce cas la particule adverbiale (-CI ou -LÀ) qui remplace le modificateur dans sa fonction référentielle. Autrement dit, on revient à la règle générale du démonstratif, qui serait de fixer la référence par proximité ou par saillance, selon Corblin (1995:78). En cela, ce type de pronom démonstratif rejoindrait effectivement le comportement du déterminant démonstratif.

Plus spécifiquement, Riegel et al. (1999:206) note que ces pronoms démonstratifs peuvent êtres utilisés pour extraire d'un antécédent qui désigne un ensemble, un membre de ce dernier, et vice versa. L'exemple 17, tiré de Riegel et al. (1999:206) illustre ce phénomène. En outre, ces démonstratifs connaissent un «emploi contrastif» dans lequel «les formes en *-ci* sont censées renvoyer à ce qui est le plus proche dans l'espace référentiel ou dans le texte et la forme en *-là* à ce qui est le plus éloigné.» Toutefois, selon Riegel et al (1999:206), cette opposition tend à disparaître avec la tombée en désuétude des formes en -CI.

Ex. 17 : *Vos livres ne sont pas chers, je prends celui-ci / ceux-là.*

3.2.2.2 LES DÉTERMINANTS DÉMONSTRATIFS DANS LA COHÉSION TEXTUELLE

Le déterminant démonstratif est classé par Riegel et al. (1999:152,156-157) parmi les déterminants définis, avec l'article défini et le déterminant possessif. Il est variable en genre et en nombre et il possède des «formes composées discontinues» en –ci et en –là. Il rencontre deux types d'usages : déictique, dans laquelle il «désigne

un référent présent dans la situation de discours ou accessible à partir d'elle»; anaphorique, où il «identifie un référent déjà évoqué au moyen d'une description identique ou différente». Riegel et al. (1999:156) mentionne que le déterminant démonstratif peut dans certains cas être utilisé dans les mêmes contextes que l'article défini, particulièrement dans les cas d'anaphore.

Corblin (1995:51), en comparant le démonstratif au défini, observe que le syntagme nominal pourvu d'un déterminant démonstratif utilise plutôt des critères de saillance et de proximité pour repérer son référent : «*Ce N*, en revanche, désigne nécessairement *a*, repéré par proximité, et le classifie comme *un N particulier*.»[18] Le référent désigné est contrasté par rapport aux autres référents possibles de la classe des *N*. Comme la proximité prime sur le contenu lexical quand un déterminant démonstratif est utilisé, le substantif qui lui est associé peut alors servir à donner une nouvelle classification au référent. Toutefois, notons que cela n'est pas obligatoire et que la description peut demeurer constante, tout comme ce serait le cas avec l'article défini. D'où la concurrence qui peut survenir dans l'usage de ces deux formes.

3.2.2.3 LES ADVERBES DÉMONSTRATIFS DANS LA COHÉSION TEXTUELLE

La théorie ne nous fournit que très peu d'information sur l'usage réel des adverbes démonstratifs (ICI et LÀ). Il est certain qu'ils peuvent être utilisés de manière déictique, pour désigner un lieu ou un endroit dans la situation d'énonciation. Cependant, dans un texte informatif écrit, tel qu'un article de journal, ce n'est peut-être pas sa fonction la plus fréquente, sauf peut-être lorsqu'il s'agit de parler du lieu où a été effectué un reportage. De toutes manières, il est possible que le lieu désigné ait déjà été mentionné dans l'article. Il faut également considérer que l'adverbe démonstratif puisse être utilisé de manière méta-textuelle, pour désigner un «endroit» dans le texte en cours.

Riegel et al. (1999:615) note : «La reprise par anaphore ne concerne pas uniquement les expressions nominales. [...] De même un adverbe de lieu comme *là*

---

[18] Dans cette citation, *N* est une variable qui représente un substantif et *a* est une variable qui représente un antécédent.

peut renvoyer à une localisation déjà mentionnée.» À titre d'exemple, il cite un élégant passage de Beaudelaire :

> Ex. 18 : *Il est une contrée qui te ressemble, où tout est beau, riche, tranquille et honnête, où la vie est douce à respirer, où le bonheur est marié au silence. C'est* ***là*** *qu'il faut aller vivre, c'est* ***là*** *qu'il faut aller mourir!*

3.3 Notions, termes et concepts retenus

Chez certains auteurs, le terme *reprise directe* désigne le même phénomène que le terme *reprise fidèle*. Manuélian (2003a:2) rapporte qu'il est employé quand la «tête nominale de l'antécédent est identique à celle de l'anaphore» (notre *anaphorique*). Nous n'avons pas d'objection à qualifier de *fidèle* une telle anaphore. Cependant, nous irons plutôt dans le sens de Tutin (1992) pour ce qui est des épithètes *direct* et *indirect* appliqués aux phénomènes endophoriques. Autrement dit, quand il y a présence d'un *antécédent sémantique* au sens de Tutin (1992:68), le lien endophorique sera qualifié d'indirect. L'anaphore associative, quant à elle, sera pour nous un cas particulier d'anaphore indirecte, où il y a présence d'un antécédent sémantique, mais absence de coréférent dans le texte.

Nous ne ferons pas de distinction, dans le cadre de cette étude, entre les notions d'exophore et de déixis. Certains auteurs y voient une distinction importante, principalement sur le plan cognitif (voir, par exemple, Cornish, 1999:112-146). Cependant, comme notre perspective est avant tout textuelle, étant donné que nous visons un apport à l'analyse automatique de textes, ces nuances sont pour nous de peu d'utilité. Ces deux termes seront donc utilisés comme des synonymes tout au long de cet exposé.

Chapitre 4 Hypothèses

4.1 Élaboration des hypothèses

Avant de pouvoir élaborer les hypothèses qui seraient testés sur corpus, il fut nécessaire de faire un premier sondage des faits empiriques. En effet, les ouvrages théoriques, dans le but de tirer des généralisations et d'isoler un phénomène particulier, aboutissent très souvent à des conclusions trop vagues pour être vérifiées directement. D'autre part, les articles de revues savantes qui rapportent les résultats d'une étude particulière, par leur spécificité, ne se laissent pas facilement appliquer à d'autres situations ou applications. En outre, leur caractère très technique les rend malheureusement hermétiques, non seulement au profane, mais également au spécialiste d'une branche voisine de la recherche linguistique. Donc, même après avoir mit autant d'ordre que possible dans tout cela, il s'avéra utile d'éclairer certains points obscurs par l'observation directe, bien que partielle, des phénomènes en question ; à savoir, des occurrences attestées d'expressions démonstratives dans des textes écrits, dans leurs utilisation endophorique, déictique ou autre.

Ces premières observations furent effectuées en analysant les articles disponibles gratuitement dans les numéros 330 (6 articles) et 331 (5 articles) de la revue de vulgarisation scientifique *Pour la science*. Les expressions démonstratives trouvées furent classées selon leur forme et leur comportement. Afin de mieux comprendre ce phénomène en particulier, des exemples supplémentaires des pronoms démonstratifs CECI, CELA et ÇA furent également extraits du Quartier Libre du 22 mars 2006, vol. 13, no. 14.

Nous nous sommes appuyé sur ces observations, ainsi que sur le fruit des recherches antérieures présentées dans le chapitre précédent, pour construire une série d'hypothèses sur le comportement de divers types d'expressions démonstratives. L'objectif était de pouvoir classer les occurrences trouvées dans l'étude de corpus et vérifier sur chacune si telle ou telle hypothèse était vérifiée ou pas.

Les hypothèses suivantes sont formulées dans un contexte bien précis : l'identification de manière probable de la présence et des caractéristique de l'antécédent, s'il y a lieu, de diverses formes d'expression démonstratives, et ce dans le cadre de l'analyse d'articles de journaux écrits en français ou de textes similaires.

Aussi, il est important de comprendre que la résolution des anaphores est une opération qui demande de considérer plusieurs expressions anaphoriques et antécédents potentiels à la fois et de procéder par élimination. Pour cela, il faut considérer les principes syntaxiques qui font qu'à l'intérieur d'une phrase, proposition ou syntagme, certaines expressions ne peuvent être coréférentielles. Ensuite, en résolvant d'abord les liens anaphoriques qui sont, soit, a priori, soit a posteriori, les plus simples, on peut souvent arriver à en déduire les autres. Les dernières incertitudes doivent alors être résolues à l'aide d'heuristiques. Voir à ce sujet les pages 185 à 191 de Corblin (1995).

Le fait de procéder ainsi par élimination implique que chaque petite «victoire» est importante, car elle peut permettre, en contexte, de déduire la solution à un problème beaucoup plus complexe ou flou.

Au sujet des données nécessaires à l'application des hypothèses, il est important de garder en tête les considérations suivantes. Dans le cadre de l'étude de corpus que nous avons effectuée, nous avons manuellement identifié les données nécessaires au fur et à mesure. Par exemple, en vérifiant l'hypothèse concernant les démonstratifs de type 5A, nous avons identifié manuellement le mot-forme dont dépendait chaque antécédent pour vérifier si c'était bien un verbe. Toutefois, dans l'optique d'une application informatique, si on choisit de retenir une méthode de résolution des anaphores qui fait appel à la notion de synonymie, par exemple, il faudra bien sûr déterminer ce qu'on considère comme un synonyme et comment on calculera les synonymes d'un mot-forme donné. Pour cela, il faudra tenir compte des besoins spécifiques à cette application en particulier, des résultats attendus, des ressources disponibles, etc. Autrement dit, il s'agit de choix de conception. Par conséquent, cela sort du cadre de cette recherche. Pour nous, il suffira de savoir qu'il est concevable d'accomplir ces tâches automatiquement.

4.2 Hypothèse générale

Avant de présenter les hypothèses spécifiques que nous avons été amenés à formuler pour certaines classes d'expressions démonstratives, il serait bon de synthétiser l'idée qui les sous-tend toutes. Ceci constituera une première tentative de réponse à notre question de recherche. Nous faisons donc l'hypothèse qu'il est effectivement pertinent, pour la résolution automatique des anaphores, de tenter d'opposer le fonctionnement des expressions démonstratives à celui d'autres moyens linguistiques, ainsi que de chercher à faire de même entre différentes classes d'expressions démonstratives.

4.3 Liste des hypothèses spécifiques par lexie

Voici donc une liste des types d'expressions démonstratives classés par lexie, accompagnée des hypothèses à tester et de quelques exemples.

4.3.1 Types 1A à 1F : CE_1 (pronom)

1A) CE_1 + relative

Hypothèse :

Le démonstratif n'a pas de rôle particulier dans la cohésion, c'est seulement une construction syntaxique particulière.

Exemples :

- *Cette fois, le public se demande ce que Tesla va faire de ce lac où flotte un bateau miniature de deux mètres de longueur.*[19]
- *Contrairement à ce que l'on pensait, la domestication du cochon a eu lieu non une fois, mais plusieurs fois en divers endroits de l'aire de répartition du sanglier.*[20]

1B) «c'est ... qui», «ce n'est pas ... qui»

Hypothèse :

Clivage. N'est pas une expression anaphorique ou déictique.

Remarques :

---

[19] CARLSON, W. Bernard. «Nikola Tesla, inventeur de rêves», in *Pour la science*, no 330, avril 2005.
[20] SAVATIER, François. «Cochons multiples», in *Pour la science*, no 331, mai 2005.

Mel'čuk (2001:189) classe le clivage dans les procédés de focalisation.

On remarque la montée du «ne … pas» dans le cas d'une proposition négative. En voici un exemple :

Et, encore une fois, ***c'est***	*la Louisiane* ***qui***	*semble avoir souffert le plus.*[21]
(Et [, encore une fois,]	la Louisiane	semble avoir souffert le plus.)
(Et [, encore une fois,]	la Louisiane	ne semble pas avoir souffert le plus.)
(Et, encore une fois, *ce* n'*est* pas	la Louisiane *qui*	semble avoir souffert le plus.)

1C) «c'est le cas de», «ce n'est pas le cas de»

Hypothèse :

Antécédent textuel coréférent direct sans lien lexical ou autre. Il s'agit d'une proposition subordonnée. C'est la première proposition subordonnée rencontrée devant l'expression.

Exemples :

- *On connaît peu d'autres exemples où des agriculteurs soient redevenus chasseurs-cueilleurs, condition supposée originelle des humains ; et quand* ***c'est le cas****, les raisons sont souvent liées à des contraintes imposées par le milieu de vie.*[22]
- *Ce bruit de fond est comparable à de l'agitation thermique, car de multiples discontinuités et structures géologiques diffusent et réémettent les ondes initiales de sorte que l'énergie finit par se répartir dans toutes les directions (comme* ***c'est le cas*** *pour l'énergie thermique).*[23]

1D) «c'est la question»

Hypothèse :

Antécédent textuel coréférent direct sans lien lexical ou autre. Il s'agit d'une proposition interrogative. C'est la première proposition interrogative rencontrée devant l'expression.

Exemple :

- *Par quel prodige ?* ***C'est la question*** *que s'est posée Claude Cyr, de la Faculté de médecine de l'Université de Sherbrooke, au Québec.*[24]

1E) «c'est-à-dire»

---

[21] [La Presse-1] (Voir l'annexe 6 pour la référence complète.)
[22] MASHAAL, Maurice. «Quand des cultivateurs redeviennent chasseurs-cueilleurs», in *Pour la science*, no 330, avril 2005.
[23] SAVATIER, François. «Utile bruit de fond sismique», in *Pour la science*, no 331, mai 2005.
[24] GROUSSON, Mathieu. «Surprenante jeunesse», in *Pour la science*, no 330, avril 2005.

Hypothèse :

Pas de réel rôle anaphorique. Relève plutôt de la cohérence que de la cohésion.

Exemple :

- *La plupart de ces mutations sont sans conséquence parce qu'elles touchent les régions non codantes de l'ADN,* ***c'est-à-dire*** *qui ne codent pas de molécules fonctionnelles.*[25]

1F) Autres cas

Hypothèse :

L'antécédent est la proposition précédente et il y a coréférence. Si le démonstratif est situé dans une proposition subordonnée à une autre, l'antécédent n'est pas cette dernière, mais celle qui précède.

4.3.2 Type 2 : CECI, CELA, ÇA

Hypothèse :

La distance au fragment textuel qui fournit le référent est au plus de deux phrases en arrière. L'antécédent, s'il est coréférent, n'est pas un groupe nominal strict, c'est-à-dire qu'il n'a pas de tête nominale, ou qu'il est une conjonction de plusieurs têtes nominales. Sinon, l'antécédent fournit le référent par association (anaphore associative). Sinon, le référent est implicite et est inférée du contexte qui précède immédiatement.

Exemple :

- *Sur un site d'épuration israélien étudié en 1998, 2,5 milliards de chironomes adultes émergeraient chaque nuit, en pleine saison, par kilomètre carré. Les biologistes ayant évalué à plus de huit pour cent la proportion de chironomes adultes porteurs de la bactérie du choléra,* ***cela*** *ferait 200 millions d'insectes porteurs par kilomètre carré et par nuit.*[26]

Note : Ceci est un cas d'anaphore implicite, où le texte précédent permet, à l'aide d'inférences, de construire une nouvelle connaissance implicite, qui sert de référent à l'expression référentielle «cela». Plus précisément, il faut assembler une opération mathématique : huit pour cent de 2,5 milliards ferait 200 millions.

---

[25] MANGIN, Loïc. «Le parent fântome», in *Pour la science*, no 331, mai 2005.
[26] MASHAAL, Maurice. «Les ailes du choléra», in *Pour la science*, no 331, mai 2005.

4.3.3 Type 3 : CELUI-CI, CELUI-LÀ, CELLE-CI, CELLE-LÀ

Hypothèse :

L'antécédent est le SN de même genre et de même nombre le plus proche qui correspond aux contraintes sémantiques imposées par le gouverneur syntaxique (Verbe ou parfois Nom, probablement) de l'expression démonstrative. Il est situé avant l'expression démonstrative et dépend syntaxiquement d'un verbe. C'est un antécédent coréférent direct.

4.3.4 Types 4A et 4B : CELUI, CELLE

Hypothèse :

Antécédent textuel non coréférent direct sans lien lexical ou autre. Il s'agit d'un substantif de même genre, mais de nombre quelconque. Si un candidat antécédent est lui-même un complément du nom (ou est le complément d'un participe passé qui complète un nom, ou est situé dans un proposition relative), il est plus probable que ce soit ce dernier nom qui soit l'antécédent.

Remarque :

C'est un cas d'identité lexicale.

4A) Avec proposition subordonnée relative

Hypothèse :

L'antécédent est la tête du premier SN rencontré devant l'expression.

Exemples :

- *Leur système de la récompense, **celui qui** libère de la dopamine, l'hormone du plaisir, fonctionne peut-être en « sous-régime ».*[27]
- *La plupart de ces mutations sont sans conséquence parce qu'elles touchent les régions non codantes de l'ADN, c'est-à-dire qui ne codent pas de molécules fonctionnelles. Parmi les autres, **celles qui** procurent un avantage sont très minoritaires par rapport aux mutations néfastes.*[28]

4B) Avec complément du nom ou participe passé

Hypothèse :

---

[27] PÉTRY, Françoise (2005). «De l'ecstasy chez les petits rats», in *Pour la Science*, no 330, avril 2005.

[28] MANGIN, Loïc. «Le parent fântome», in *Pour la science*, no 331, mai 2005.

L'antécédent est la tête d'un SN situé devant l'expression. Il est dans la même phrase et devant l'expression.

Remarque :

On remarque qu'il existe parfois un lien d'hyponymie entre l'antécédent et le SN complément du pronom démonstratif.

Exemples :

- *Les deux systèmes de neuromédiateurs les plus perturbés sont **celui** de la sérotonine et **celui** de la dopamine : la concentration de ces deux médiateurs augmente. ».*[29]
- *D'après les résultats de cette étude génétique, qui s'ajoutent aux indices linguistiques et culturels, les Mlabri sont probablement originaires d'un tout petit groupe d'agriculteurs qui ont, à un moment donné, changé de mode de vie et adopté **celui** de chasseurs-cueilleurs.*[30]
- *En tout cas, l'exemple des Mlabri, s'il se confirme, indique que le mode de vie de chasseur-cueilleur ne précède pas forcément **celui** de cultivateur.*[31]
- *Vieux de quelque 8 500 ans, ces os sont **ceux** d'un cochon déjà typiquement domestique, et non **ceux** d'un sanglier.*[32]
- *Une certaine communauté génétique entre les cochons sauvages$_{s10}$ d'Australie (des cochons marrons introduits par l'homme) et **ceux** d'Asie du Sud-Est désigne aussi un ancien foyer de domestication dans cette région.*[33]

4.3.5 Types 5A à 5C : CE_2 (déterminant)

5A) CE_2 + dernier(e)(s)

Hypothèse :

Antécédent textuel coréférent direct sans lien lexical ou autre. Il s'agit d'un SN de même genre et nombre. Il est le premier SN rencontré qui dépende d'un verbe.

Exemples :

- *Ce scénario rejoint une tradition orale des Tin Prai quant à l'origine des Mlabri, selon laquelle, il y a plusieurs siècles, les villageois Tin Prai ont*

---

[29] PÉTRY, Françoise (2005). «De l'ecstasy chez les petits rats», in *Pour la Science*, no 330, avril 2005.
[30] MASHAAL, Maurice. «Quand des cultivateurs redeviennent chasseurs-cueilleurs», in *Pour la science*, no 330, avril 2005.
[31] Ibidem.
[32] SAVATIER, François. «Cochons multiples», in *Pour la science*, no 331, mai 2005.
[33] Ibidem.

*expulsé deux enfants en les mettant sur un radeau ; **ces derniers** auraient survécu et se seraient réfugiés dans la forêt.*[34]

- *Elle assure le bon fonctionnement d'autres glandes endocrines, par exemple de la glande thyroïde et des glandes surrénales ; elle stimule chez la femme la maturation des follicules ovariens et la production d'estrogènes et, chez l'homme, la spermatogenèse, et elle produit l'hormone de croissance. **Cette dernière** entraîne la croissance des os longs en stimulant l'activité du cartilage et favorise l'augmentation de la masse musculaire, assurant ainsi la croissance.*[35]
- *Pour ce faire, ils ont assimilé le bruit de fond sismique à l'agitation thermique, ce qui leur permet de transposer à la sismique l'un des théorèmes fondamentaux de la thermodynamique. **Ce dernier** relie les fluctuations thermiques d'une grandeur physique observable à la valeur de cette grandeur après une excitation extérieure.*[36]

5B) CE_2 + nom temporel ou nom métatextuel

Hypothèse :

Aucun antécédent textuel. Réfère à l'unité textuelle ou temporelle courante.

Remarque :

Se référer à l'annexe B de Boudreau (2004:iv) pour la liste des mots de ce genre qu'elle a retenus : [jour(s), semaine(s), année(s), chapitre(s), section(s), article(s)]

Puisque nous procédions à une annotation manuelle, nous avons décidé de ne pas nous limiter à une liste fermée de termes. Cependant, nous avons gardé à l'esprit cette liste lors de l'annotation.

5C) Autres cas

Hypothèse :

L'antécédent est le premier SN dont la tête est identique à celle de l'expression démonstrative. Pour cela, retourner en arrière jusqu'à une distance maximale de quatre phrases ou jusqu'au début du paragraphe courant, en prenant la plus longue de ces deux distances.

---

[34] MASHAAL, Maurice. «Quand des cultivateurs redeviennent chasseurs-cueilleurs», in *Pour la science*, no 330, avril 2005.
[35] GROUSSON, Mathieu. «Surprenante jeunesse», in *Pour la science*, no 330, avril 2005.
[36] SAVATIER, François. «Utile bruit de fond sismique», in *Pour la science*, no 331, mai 2005.

Si aucun candidat n'est trouvé de cette manière, en reculant, à partir de l'expression démonstrative, on calcule le «nom de base»[37] qui correspond à chaque syntagme nominal, verbal ou propositionnel rencontré dans la phrase courante, puis dans la phrase précédente, jusqu'à ce qu'on en trouve un qui corresponde à l'expression démonstrative.

Si aucun candidat n'est trouvé, on fait la même chose avec les synonymes, les quasi-synonymes, puis les hyponymes.

Si aucun candidat n'est trouvé, on fait la même chose avec les noms d'actants, excepté que dans ce cas, si un match est trouvé, le syntagme en question ne sera pas coréférent (lien indirect) de l'expression démonstrative. Si l'actant du syntagme en question est explicite, c'est-à-dire qu'il est réalisé par un autre syntagme, c'est ce dernier syntagme qui sera coréférent et on dira qu'il y a anaphore indirecte, sinon, on dira qu'il s'agit d'une anaphore associative.

Si aucun candidat n'est trouvé, (cas de re-classification) l'antécédent sera le premier syntagme devant l'expression démonstrative qui correspond aux contraintes sémantiques imposées par le gouverneur syntaxique de celle-ci, en allant encore une fois au maximum jusqu'au début de la phrase précédente.

Si aucun candidat n'est trouvé, on conclura qu'il s'agit d'un déictique.

Exemples :

a) Répétition de la tête du SN antécédent

- *Pour tenter de répondre à cette question, l'équipe de Sylvie Chalon et Laurent Galineau (INSERM U619, Université François Rabelais à Tours) a étudié les conséquences de l'administration d'ecstasy à des femelles de rat gestantes sur le fœtus à naître. Et sur le rat,* ***cette administration*** *n'est pas sans conséquence.*[38]
- *Ce jour de mai 1899, quand les membres du Club de commerce de Chicago arrivent pour écouter la conférence du célèbre ingénieur électricien Nikola Tesla, ils sont étonnés de découvrir un lac artificiel au milieu de la salle.*

---

[37] Nous faisons référence ici au concept de «*basic level name*» («nom de base», [N.T]) mentionné par Tutin & Viegas (2000:234). Ces dernières font référence à ce sujet à Rosch et al. (1976:382) et Reiter (1991). Par exemple, «mammifère» et «chien» sont tous les deux des hyponymes de «fox terrier», mais il est plus naturel en français courant d'utiliser le terme «chien», pour parler d'un fox terrier, que le terme «mammifère». Le terme «chien» est donc considéré comme un «nom de base». Tutin & Viegas (2002:234) parlent ainsi d'une «*"naturalness" constraint*» («contrainte de naturalité», [N.T]).

[38] PÉTRY, Françoise (2005). «De l'ecstasy chez les petits rats», in *Pour la Science*, no 330, avril 2005.

Chacun sait pourtant que Tesla, l'homme qui a inventé le courant alternatif et qui a amené l'électricité dans les foyers et sur les lieux de travail, est maître dans l'art de la mise en scène. Six ans plus tôt, durant l'exposition colombienne de Chicago, l'ingénieur a ébahi son auditoire en s'infligeant des chocs électriques de 250 000 volts. Cette fois, le public se demande ce que Tesla va faire de ***ce lac*** *où flotte un bateau miniature de deux mètres de longueur.*[39]

- *À la fin de l'adolescence, voulant devenir ingénieur, il intègre l'École polytechnique Joanneum de Graz, en Autriche, et découvre avec enthousiasme la physique dans les cours de Jacob Poeschl en 1876 et 1877. Durant* ***ces cours****, Tesla commence à réfléchir à ce qui sera sa plus remarquable invention : le moteur à courant alternatif.*[40]

b) Nominalisation de l'antécédent

- *Or les rats nés de mère ayant reçu de l'ecstasy ont moins d'attirance pour le sucre. Ils seraient anhédoniques, c'est-à-dire indifférents au plaisir. Leur système de la récompense, celui qui libère de la dopamine, l'hormone du plaisir, fonctionne peut-être en « sous-régime ». Cette* ***anhédonie*** *résulte-t-elle du traitement subi par leur mère ?*[41]

c) Hyponyme de l'antécédent, «nom de base» de l'antécédent

- *Les cerveaux de rats, prélevés à différents stades du développement fœtal et postnatal, ainsi que chez l'adulte, ont révélé quelques anomalies notables. Chez le rat nouveau-né dont la mère n'a pas reçu d'ecstasy, on constate un pic important de la concentration en sérotonine juste après la naissance. En revanche, chez le nouveau-né d'une mère ayant reçu de l'ecstasy, l'augmentation de la concentration en sérotonine est très faible.* ***Ce premier résultat*** *semble indiquer un dysfonctionnement du système sérotoninergique.*[42]
- *Ils ont constaté que la libération de dopamine et de sérotonine était forte chez les animaux dont la mère n'avait pas reçu d'ecstasy, mais faible chez les autres. Selon S. Chalon, les réserves des neuromédiateurs seraient inférieures à la normale ou les neurones qui libèrent* ***ces deux neuromédiateurs*** *fonctionneraient moins bien.*[43]

d) Reclassification, changement de point de vue

- *En dépit de sa démonstration spectaculaire, Tesla ne fit jamais de son bateau télécommandé une arme opérationnelle.* ***Cet échec*** *est caractéristique d'un*

---

[39] CARLSON, W. Bernard. «Nikola Tesla, inventeur de rêves», in *Pour la science*, no 330, avril 2005.
[40] Ibidem.
[41] PÉTRY, Françoise (2005). «De l'ecstasy chez les petits rats», in *Pour la Science*, no 330, avril 2005.
[42] Ibidem.
[43] Ibidem.

trait récurrent qui a marqué sa vie : un profond idéalisme, qui tenait rarement compte des contingences matérielles et financières.[44]

4.3.6 Types 6A et 6B : CE -CI

6A) CE -CI + nom temporel ou nom métatextuel

Hypothèse :

Aucun antécédent textuel. Réfère à l'unité textuelle ou temporelle courante.

Remarque :

Même que pour 5B.

6B) Autres cas

Hypothèse :

Comme pour les «autres cas» de CE_2.

4.3.7 Types 7A et 7B : CE -LÀ

7A) CE - LÀ + nom temporel ou nom métatextuel

Hypothèse :

L'antécédent est la première expression métatextuelle ou temporelle compatible trouvée avant l'expression avec une distance maximale de deux phrases. Sinon, on conclura à l'absence d'antécédent textuel. Réfère alors à l'unité textuelle ou temporelle courante.

Remarque :

Même que pour 5B.

7B) Autres cas

Hypothèse :

Comme pour les «autres cas» de CE_2.

4.3.8 Type 8 : ICI, LÀ

Nous n'avons formulé aucune hypothèse concernant les lexies ICI et LÀ.

---

[44] CARLSON, W. Bernard. «Nikola Tesla, inventeur de rêves», in *Pour la science*, no 330, avril 2005.

Remarque :

Manque d'exemples attestés et d'études théoriques ou empiriques sur le sujet.

4.4 Évaluation des hypothèses

La justesse des hypothèses doit être évaluée de plusieurs manières par rapport aux données du corpus et aux caractéristiques de chaque hypothèse. Il faudra se demander quelle proportion des expressions démonstratives concernées par l'hypothèse y correspond parfaitement, quelle proportion n'y correspond pas du tout, et quelle proportion y correspond un peu. Il faudra aussi discuter de la signification de ces «un peu» et «pas du tout».

Également, il faut évaluer la pertinence de chaque hypothèse pour l'analyse automatique de textes. D'une part, dans quelle proportion l'hypothèse permet-elle de déterminer de manière univoque la présence et l'identité d'un antécédent pour les expressions démonstratives concernées par celle-ci? D'autre part, quelle est la proportion des expressions démonstratives concernées par l'hypothèse? Il faut aussi estimer les ressources nécessaires à identifier qu'un démonstratif appartient au type d'expressions démonstratives visé par l'hypothèse, ainsi que celles nécessaires au traitement de ces expressions démonstratives au moyen de la méthode proposée par cette hypothèse.

Chapitre 5 Méthode

5.1 Définition opérationnelle du démonstratif

Dans le cadre de cette étude de corpus, un démonstratif est tout mot-forme qui figure dans colonne Forme du tableau 5, situé à l'annexe 1. Le tableau 1 indique également quels genre, nombre et lexie nous avons attribués à chaque forme morphographique. Dans le cas des lexies en CE, les indices servent à distinguer le pronom (CE_1) du déterminant (CE_2). Le terme morphographique signifie que la forme est non seulement fléchie (morphologie), mais aussi qu'elle est adaptée en fonction des règles morphographiques du français écrit (l'équivalent des règles morphophonologiques de l'oral). Par exemple, le pronom neutre CE_1 prend la cédille et l'apostrophe devant un mot qui commence par une voyelle graphique basse (*a* et *o*). Par contraste, devant une voyelle, la forme masculine plurielle du déterminant CE_2 ne prend pas l'apostrophe ; on y ajoute plutôt un *t*.

5.2 Données

Nous avons choisi des textes qui pourraient assez bien représenter, à notre avis, la langue générale. Nous ne voulions pas avoir à tenir compte des particularités d'une langue de spécialité, comme celles qu'on retrouve dans les revues scientifiques. De plus, nous voulions travailler sur des textes susceptibles de ressembler à ceux qu'une application d'analyse automatique de texte aurait à traiter. Les articles de journaux nous ont semblé un exemple intéressant de textes disponibles en grand nombre et pour lesquels il serait avantageux de pouvoir extraite automatiquement de l'information, de produire des résumés ou des condensés, de fournir une traduction au besoin, etc. Toutes ces applications ont besoin, pour être performantes, de résoudre certains phénomènes anaphoriques.

Les deux journaux qui furent choisis sont *La Presse* du dimanche 25 septembre 2005 et le *Voir* du 22 septembre 2005. Les articles faisant entre 250 et 1000 mots furent d'abord extraits, ce qui donna 78 articles pour *La Presse* et 45 articles pour *Voir*.

5.3 Outils informatiques

5.3.1 Fichier texte brut

Le texte des journaux fut d'abord converti en fichiers de texte brut. Les fichiers de texte brut sont faciles à traiter, prennent peu d'espace sur le disque et sont compatibles avec quantité de logiciels et de systèmes d'exploitation. Dans beaucoup de systèmes d'exploitation, on les reconnaît à l'extension *.TXT* de leur nom de fichier.

Il existe différents encodages pour les caractères numériques, alphabétiques et autres dans un fichier de texte brut. Le plus populaire et le plus connu est sans doute ASCII. La plupart des autres conventions d'encodage reprennent ASCII en y ajoutant des codes supplémentaires pour représenter, par exemple, les caractères accentués du français ou les symboles de l'écriture chinoise. L'encodage Windows-1252, courant sur les ordinateurs fonctionnant sous le système d'exploitation Windows, fonctionne ainsi. Nous avons également choisi d'utiliser, lorsque c'était pratique, l'encodage UTF-8, une implémentation sur huit bits de la convention d'encodage Unicode. Cet encodage couvre les symboles et caractères accentués dont nous avions besoin et est en passe de constituer un standard international pour l'encodage de données textuelles.

5.3.2 Python

Python est un langage de programmation qui possède de bonnes capacités pour le traitement de texte. De plus, il inclut un module qui permet l'utilisation des expressions régulières pour la recherche et le remplacement de segments textuels. Il n'est pas forcément très rapide, car ce n'est pas un langage entièrement compilé. Cependant, il est relativement facile d'utilisation, de par sa syntaxe claire et lisible, ainsi que sa souplesse face à divers styles de programmation. Il nous a donc apparu un excellent choix pour la rédaction du petit programme d'étiquetage dont nous avions besoin.

Pour une introduction générale à Python, on peut se référer au tutoriel de Guido van Rossum (*Python Tutorial*), disponible sur le site Internet

http://python.org/. Pour plus d'information sur le traitement de texte avec Python, voir le manuel *Text Processing in Python*, par David Mertz. Celui-ci est disponible gratuitement, en format texte brut, sur le site Internet http://gnosis.cx/TPiP/. Pour plus de détails sur l'utilisation des expressions régulières dans Python, se référer à *Regular Expressions HOWTO*, par A.M. Kuchling, dans la documentation de Python (http://www.amk.ca/python/howto/regex/).

5.3.3 XML

Pour ce qui est de l'annotation de notre corpus, nous nous sommes largement inspirés de l'approche adoptée par Boudreau (2004:33-47). Au chapitre 4 de son mémoire, elle présente le langage XML et les raisons qui l'ont poussées à choisir cet outil de structuration de données pour baliser son propre corpus. Nous sommes d'accord avec elle, pour l'avoir expérimenté, que les concordanciers ne sont pas adéquats pour l'étude des phénomènes anaphoriques. En effet, ceux-ci ne permettent l'accès qu'à un extrait relativement petit du contexte d'occurrence de l'expression recherchée.

D'autre part, il est difficile de trouver un outil d'annotation qui corresponde exactement aux besoins d'une étude pointue des phénomènes endophoriques et exophoriques. Des logiciels d'annotation existent, comme MMAX[45], par exemple, mais ils ne sont pas nécessairement faciles d'accès et ne correspondent souvent que partiellement au type d'annotation dont on a besoin. Il ne s'agissait pas uniquement, pour nous, de marquer un syntagme nominal comme étant l'antécédent de tel ou tel pronom. Nous voulions pouvoir annoter et coder chaque occurrence selon la nature du lien anaphorique : coréférence, liens lexico-sémantiques, lien direct ou indirect, etc. Nous voulions aussi pouvoir spécifier la catégorie et la fonction syntaxiques de l'antécédent. XML n'offre certes pas les avantages d'un logiciel *clés en mains*, mais il a le mérite de laisser beaucoup de place à un format de balisage personnalisé. Il suffit alors de formuler précisément les contraintes à suivre dans un fichier DTD

[45] Voir le site Internet http://www.eml-research.de/english/research/nlp/download/mmax.php pour plus de détails sur ce logiciel d'annotation.

(*Document Type Definition*) pour pouvoir ensuite faire valider automatiquement le balisage effectué. Le fichier DTD que nous avons utilisé est présenté à l'annexe 4.

De plus, comme le souligne Boudreau (2004:41-44), XML est accompagné de plusieurs outils qui permettent en fin de compte de traiter un fichier XML, ou un ensemble de fichiers XML, comme une véritable base de données. Il s'agit principalement de XPath et de XQuery. Le premier permet de naviguer dans une arborescence XML et d'y repérer un nœud ou un ensemble de nœuds. Le second permet de rédiger des requêtes plus complexes, comme de traiter plusieurs fichiers XML à la fois ou de donner la cardinalité d'un ensemble de nœud préalablement repéré par une expression XPath. Notez qu'il est la plupart du temps nécessaire de disposer d'un logiciel spécialisé en traitement des fichiers XML pour pouvoir exécuter une requête XQuery complexe. XMLSpy, de la compagnie Altova, est un exemple de ce type de logiciel. Un tel logiciel peut également faciliter le balisage manuel des fichiers XML.

Par exemple, la requête XQuery suivante (encadré 1) parcourt tous les fichiers XML constituant notre échantillons d'articles de *La Presse* à la recherche des expressions démonstratives de type 5C anaphoriques sans lien lexical. Elle en affiche la liste, accompagnés du nom de fichier et du contexte (un paragraphe) dans lequel chaque occurrence répondant à ces critères a été trouvée. Notez que nous avons considéré les expressions anaphoriques avec déterminant démonstratif, mais sans lien lexical (réitération ou hyponymie, par exemple), comme des cas de reclassification quand nous avons voulu dénombrer ces cas.

```
xquery version "1.0";
<liste>
{
for $fichier in ("La Presse_3.xml","La Presse_5.xml","La Presse_25.xml","La
Presse_28.xml","La Presse_32.xml","La Presse_37.xml","La Presse_40.xml","La
Presse_46.xml","La Presse_47.xml","La Presse_49.xml","La Presse_51.xml","La
Presse_55.xml","La Presse_59.xml","La Presse_65.xml","La Presse_67.xml","La
Presse_70.xml","La Presse_73.xml","La Presse_74.xml","La Presse_77.xml")
for $blabla in doc($fichier)//dem[@type="5c"]
for $blob in $blabla/ancestor::anaphore
return
        if (count($blob[@lienlexical]) = 0)
        then
        <occurrence>
        <fichier>{$fichier}</fichier>
        <expression>{data($blabla/parent::syntagme)}</expression>
        {$blabla/ancestor::par}
        </occurence>
        else ()
}
</liste>
```

Encadré 1 Exemple de requête XQuery

Pour plus de clarté, un extrait de l'output d'une telle requête est disponible à l'annexe 2. On peut y voir également l'aspect que peut avoir un fichier XML annoté quand il est visionné en format texte brut. Les différentes couleurs ont été ajoutées automatiquement par le logiciel que nous utilisons pour manipuler les fichiers XML pour augmenter leur lisibilité. Elles permettent simplement de mieux distinguer les balises, ainsi que leurs attributs, du reste du texte. Ces couleurs ne font pas partie des fichiers XML proprement dits, qui sont en fait des fichiers de texte brut, donc dépourvus de couleurs, de caractères en gras ou en italique, etc.

Enfin, on peut varier à l'infini la présentation de nos données à l'aide de CSS, XSL ou XSLT. Les fichiers rédigés selon l'un ou l'autre de ces standards permettent, à l'aide d'un navigateur Web ou d'un autre logiciel ayant la capacité d'afficher des fichiers XML, de faire apparaître tel ou tel aspect des données. Par exemple, on peut décider d'afficher tous les démonstratifs en rouge et de faire souligner les antécédents. En ajoutant un indice qui permette d'identifier les syntagmes et un exposant qui corresponde à l'antécédent, on peut vérifier d'un seul coup d'œil si notre annotation contient des erreurs ou des oublis. Le fichier XSL listé à l'annexe 3

est un exemple de ce qu'on peut faire avec ces outils. Il permet d'afficher un fichier XML contenant un article de journal annoté, selon un format d'affichage permettant de repérer facilement les démonstratifs et les liens anaphoriques. De plus, il affiche en haut de la page quelques statistiques sur le contenu du fichier. Également, dans ce cas-ci, on compte le nombre de démonstratifs de type 1A et on les affiche en rouge dans le texte.

L'encadré 2 présente un extrait d'un article annoté de notre corpus affiché selon les paramètres du fichier XSL de l'annexe 3. Il s'agit du fichier "La Presse_28.xml".

	pronoms	**déterminants**	**adverbes**	**total**
nombre d'expressions démonstratives	10	9	3	22
nombre d'expressions anaphoriques	9	4	1	14
anaphoriques coréférentiels	9	3	1	13

nombre de démonstratifs de type 1a : 1

```
Beautés désespérées
Secrets de banlieue
Sarfati, Sonia; Dumas, Hugo
```

Richesse, beauté, amour, trahison, meurtres : Desperate Housewives$_{a1}$ a tous les ingrédients du soap classique. Comment expliquer alors le succès exceptionnel de **cette** émission suivie chaque semaine par plus de 20 millions d'Américains$_{d1}$[a1] ?

Mary Alice qui raconte une journée tout **ce** qu'il y a de plus banale$_{d2}$ culminant par un événement qui l'est moins, son suicide. Après les funérailles, ses quatre amies passent par la maison de la défunte. Susan et son macaroni raté, grâce auquel elle parvient à... faire du plat au nouveau voisin. Lynette et ses trois petits démons, qu'elle ira repêcher dans la piscine. Bree et sa famille tirée à quatre épingles, avec son panier de muffins faits maison. Gabrielle et son riche mari, qui a pour mission de faire connaître le prix du collier qu'elle étrenne. Le tout se terminant par la découverte d'une lettre qui semble en être une de chantage. $_{p1}$

[...]

Encadré 2 Extrait d'un article annoté affiché par le truchement d'un fichier XSL

Pour plus de détails, on peut se référer au chapitre 4 de Boudreau (2004:33-47), qui présente brièvement XML et résume bien son potentiel pour l'annotation de corpus dans l'étude des anaphores. Également, le site Internet http://www.w3schools.com/ contient quantité de tutoriels et de documents de

référence sur XML et sur les divers outils qui gravitent autour de ce standard de structuration de données.

5.4 Assemblage, traitement et balisage du corpus

Tout d'abord, un échantillon de 20 articles de chaque journal fut sélectionné aléatoirement. Ces articles furent transformés en fichiers de texte brut. La liste de ces 40 articles est donnée à l'annexe 6.

Ensuite, un programme informatique spécialement rédigé à cette fin en Python fut utilisé pour effectuer un premier étiquetage. Il s'agissait lors de ce balisage automatique d'identifier tous les démonstratifs et de les annoter selon leur lexie d'appartenance, leur genre, leur nombre et leur catégorie grammaticale. Dans un souci de simplicité, le programme ne fut pas conçu pour distinguer les homographes, tout au plus pour se tromper le moins souvent possible. Une révision manuelle a donc dû être effectuée pour s'assurer que les formes morphographiques présentant une homographie soient correctement étiquetées. Les données qui servirent à rédiger ce programme sont celles présentées à l'annexe 1. Le programme d'étiquetage des démonstratifs ajoutait également au texte brut le squelette d'un fichier XML. Ce squelette devait ensuite être complété par la phase d'étiquetage manuel.

Ce programme est en réalité constitué de deux fichiers, l'un s'occupant d'encapsuler les informations nécessaires à l'étiquetage des démonstratifs, le second effectuant le travail d'étiquetage proprement dit, ainsi que l'ajout du squelette XML. Ce second fichier est d'ailleurs inspiré d'un fichier ayant une fonction semblable m'ayant été transmis par courriel par Sylvie Boudreau. Je lui suis reconnaissant de ce coup de pouce, tout en conservant bien sûr la responsabilité des changements que j'ai dû y apporter pour l'adapter à mes besoins. Les deux fichiers "demonstratifs.py" et "etiquetage3.py" sont présentés à l'annexe 5.

L'étape suivante, l'étiquetage manuel, avait notamment pour but de baliser les frontières des expressions démonstratives pour chaque démonstratif. Il fallait ensuite assigner à chaque syntagme ainsi délimité une fonction et une catégorie syntaxiques. S'il y avait lieu, le ou les antécédents textuels étaient alors identifiés et

la relation d'anaphore ou de cataphore qualifiée selon divers critères : distance en nombre de phrases, et distance en nombres d'éléments de sa propre catégorie syntaxique, entre l'antécédent et l'expression démonstrative; présence d'un lien lexico-sémantique (fonction lexicale, synonymie, etc.) ou d'un autre type de lien; coréférence; lien direct ou indirect. Pour la distance en nombre phrases, nous avons considéré qu'une frontière de phrase était constituée par une des ponctuations suivantes : [«.»,«:»,«?»,«!»,«;»,«...»]; ou, le cas échéant, par un changement de paragraphe. Nous avons également annoté le genre, le nombre, la fonction et la catégorie syntaxique de l'antécédent. Toutes ces annotations furent effectuées à l'aide de balises XML. Un fichier Document Type Definition (DTD) servit à valider cet étiquetage. Une retranscription est disponible à l'annexe 4. Elle permet de voir les conventions de balisage que nous avons adoptées.

5.4.1 Protocole d'annotation manuelle du corpus

Voici le protocole que nous avons suivi dans l'annotation manuelle du corpus. À titre d'exemple, un extrait du fichier "La Presse_28.xml" est fourni à l'annexe 7.

1) Ouvrir un fichier XML contenant un article dont les démonstratifs ont été annotés automatiquement.
2) Baliser le titre et l'auteur avec les balises *titre* et *auteur*.
3) Baliser les autres informations apparaissant en début et en fin de fichier et ne faisant pas partie de l'article proprement dit avec la balise *nontraite*.
4) Trouver le premier démonstratif balisé (balises *dem*)
5) Ajouter l'attribut *type* à la balise *dem* et entrer comme valeur le type d'expression démonstrative, tel que spécifié dans le chapitre *Hypothèses*.
6) Entourer de balises *syntagme* le syntagme correspondant à l'expression démonstrative dont fait partie le démonstratif trouvé en (4). (La notion d'expression démonstrative est définie dans le chapitre Problématique.)
7) Entrer les valeurs des attributs de la balise *syntagme* :
    a. Attribut *num* : Il s'agit simplement d'un numéro identifiant de manière unique le syntagme dans le fichier XML courant. Pour faciliter la lecture

par l'annotateur, on suggère de mettre comme valeur *d1* pour la première expression démonstrative, *d2* pour la deuxième, etc.

b. Attribut *categorie* : Entrer la catégorie syntaxique du syntagme.

c. Attribut *fonction* : Entrer la fonction syntaxique du syntagme.

d. Attribut *genre* : Entrer le genre du syntagme.

e. Attribut *nombre* : Entrer le nombre du syntagme.

8) Vérifier si l'expression démonstrative est une expression endophorique. Si oui, poursuivre à l'étape (9). Si non, passer à l'étape (14).

9) Repérer le ou les antécédent(s) de l'expression démonstrative.

10) Sélectionner le premier antécédent trouvé.

a. S'il s'agit d'un syntagme, l'entourer de balises *syntagme* et entrer les valeurs de ses attributs comme en (7). À l'étape (7a), pour faciliter la lecture, il est suggéré d'attribuer l'identificateur *a1* pour le premier antécédent trouvé dans l'ensemble du fichier, *a2* pour le second, etc.

b. S'il s'agit d'un nom seul, comme dans les cas d'identité lexicale, utiliser les balises *nom*. Entrer ensuite les valeurs des attributs *num* et *genre* comme en (10a) et (7d) respectivement.

c. S'il s'agit d'un paragraphe en entier, identifier la balise *par* pertinente déjà placée par le programme d'étiquetage automatique. Ajouter simplement un attribut *num* et lui donner une valeur : *p1*, par exemple.

d. S'il s'agit d'une autre expression démonstrative, noter simplement l'identificateur qui lui a déjà été attribué si elle a déjà été balisée. Sinon, appliquer les étapes (6) et (7) à cette expression démonstrative avant de poursuivre.

11) Répéter l'étape (10) pour tous les antécédents trouvés en (9). Quand tous les antécédents ont été traités, passer à l'étape (12).

12) Entourer les balises *syntagme* placées en (6) de balises *anaphore*. (Cette balise inclut les anaphores et les cataphores.)

13) Entrer les valeurs des attributs de la balise *anaphore* :

a. Attribut *antecedent* : Entrer la liste des identificateurs des éléments qui forment les antécédents de l'expression endophorique. Ces éléments peuvent eux-mêmes être des expressions anaphoriques.
b. Attribut *coreference* : Mettre *oui* si au moins un des antécédents est coréférent à l'expression démonstrative. Mettre *non* sinon.
c. Attribut *lien* : Mettre *indirect* s'il s'agit d'une anaphore indirecte ou d'une anaphore associative. Mettre *direct* sinon.
d. Attribut *lienlexical* : S'il y a lieu, entrer la valeur de la fonction lexicale ou des fonctions lexicales qui décrivent le mieux la relation lexico-sémantique entre l'antécédent pertinent et l'expression démonstrative. Si cela est impossible, indiquer le plus clairement et brièvement possible la nature du lien lexico-sémantique. S'il n'y a pas de lien lexico-sémantique, omettre cet attribut.
e. Attribut *lienautre* : S'il y a lieu, indiquer la nature du lien paralinguistique entre l'antécédent pertinent et l'expression démonstrative. S'il n'y en a pas, omettre cet attribut.
f. Attribut *distph* : Entrer la distance en phrases entre l'expression démonstrative et l'antécédent le plus proche. Si la distance est supérieure ou égale à 4, entrer *4etplus*.
g. Attribut *distcat* : Entrer la distance en nombres d'éléments de sa propre catégorie syntaxique entre l'antécédent le plus pertinent et l'expression démonstrative. Par exemple, si l'antécédent est un SN, compter le nombre de SN entre l'antécédent et l'expression démonstrative. Si ce nombre est très grand, il suffit d'omettre cet attribut. La valeur par défaut *beaucoup* sera automatiquement attribuée.
h. Attribut *position* : C'est la position du ou des antécédent(s) par rapport à l'expression démonstrative. Entrer *avant* s'il s'agit d'une anaphore et *après* s'il s'agit d'une cataphore.

14) Trouver le démonstratif suivant et retourner à l'étape (5). S'il ne reste plus de démonstratifs à baliser, passer à l'étape suivante.

15) Fermer le fichier XML en cours. Ouvrir le fichier XML suivant et retourner à l'étape (2). S'il ne reste plus de fichiers XML à traiter, passer à l'étape suivante.
16) Le corpus est maintenant étiqueté.

5.5 Extraction des données et vérification des hypothèses

Enfin, on put procéder au listage et au dénombrement des divers types d'expressions, classées au besoin par distance, catégorie d'antécédent, etc. Cette étape fut facilitée par l'utilisation de XPath et XQuery. Nous avons, d'une part, fait une étude quantitative des données en compilant le résultat des diverses requêtes de comptage dans des fichiers Excel : un pour chaque sous-corpus et un autre pour les effectifs totaux du corpus dans son ensemble. Ce fichier nous a également permis de calculer des proportions et autres statistiques dérivées, en plus de nous aider à produire des graphiques résumant d'une manière plus visuelle nos données. Pour ce qui est de l'étude des données sous leur aspect plus qualitatif, nous avons constitué des listes pour la plupart des types prédéfinis d'expressions démonstratives. Ces listes nous ont servi à analyser plus en détails, dans certains cas, en quoi et pour quelles raisons une hypothèse s'avérait inadéquate, ou à préciser le sens à donner à telle ou telle donnée statistique.

Pour plus de détails concernant le corpus et les données qui en ont été extraites, prière de communiquer avec l'auteur par l'intermédiaire du département de linguistique et de traduction de l'Université de Montréal.

Chapitre 6 Analyse des résultats

6.1 Aperçu général

Dans le sous-corpus de *La Presse*, on dénombre 289 démonstratifs; dans celui du *Voir*, on en dénombre 290, pour un total de 579 dans tout le corpus. Le nombre moyen de démonstratifs par article dans ces deux échantillons est donc sensiblement le même. Toutefois, on compte 1 cas dans chacun des deux sous-corpus où deux démonstratifs sont associés dans une même expression. Le nombre total d'expressions démonstratives est donc en réalité de 577.

Les expressions démonstratives anaphoriques contenant un déterminant démonstratif ont une distance à l'antécédent, en termes de phrases, plus importante («moyenne»[46] de 1,67 phrases, médiane de 1 phrase) que celles contenant un pronom démonstratif («moyenne» de 0,64 phrase, médiane de 0 phrase). L'antécédent des premières est aussi plus souvent un syntagme nominal (80%) et moins souvent une proposition (13%) que celui des deuxièmes (SN 51% et proposition 30%). Parmi les pronoms démonstratifs, ce sont CECI, CELA et ÇA (type 2) qui illustrent le mieux cette dernière tendance (SN 35% et proposition 53%).

Les cas de type 4B ont beaucoup plus d'antécédents «nom seul» (87%) que ceux de type 4A (45%), qui comptent davantage de syntagmes nominaux comme antécédent (4A 55% et 4B 13%). Les cas de type 4B correspondent ainsi beaucoup plus que ceux de type 4A à l'hypothèse préalablement formulée à leur sujet (4B 77% et 4A 29%).

En ce qui concerne la fonction syntaxique des antécédents syntagmes nominaux, dans presque tous les cas où le nombre d'occurrences est suffisamment grand, la répartition semble ne pas dévier de celle valable pour les syntagmes nominaux en général. On remarque toutefois un plus grand nombre de COD que de SG pour les cas de type 1F, ce qui est peut-être dû à ce que ce les occurrences de ce

---

[46] Nous ne pouvons pas calculer la véritable distance moyenne, car les distances supérieures à 3 ont été simplement notées «4 et plus». Néanmoins, dans le but de comparer les types d'expressions entres elles, nous avons élaboré une formule qui se rapproche de celle de la moyenne. Le calcul de cette fausse moyenne est le suivant : (proportion d'antécédents à la distance 0) × 0 + (proportion d'antécédents à la distance 1) × 1 + … + (proportion d'antécédents à la distance 4 ***et plus***) × 4.

type ont une grande proximité à leur antécédent (56% d'antécédents dans la même phrase).

Les figures 3 et 4, à la page suivante, montrent la répartition des expressions démonstratives de chaque catégorie et de chaque type. Dans le cas des déterminants, on a regroupé les types 5B, 6A et 7A, d'une part, et les types 5C, 6B et 7B, d'autre part. Le lecteur est prié de se référer à l'analyse des résultats de ces types d'expressions démonstratives, présentée plus bas, pour de plus amples explications.

Il est important de tenir compte de la proportion de chaque type d'expression démonstrative. Cette proportion peut en effet servir à cibler les problèmes prioritaires et les solutions les plus avantageuses à implanter dans une application d'analyse automatique de texte. Le cas optimal est celui d'une approche simple et efficace s'appliquant bien à un grand nombre de cas. Toutefois, si une méthode très simple solutionne un sous-ensemble plus réduit d'expressions, mais qui n'était pas traité convenablement jusqu'à présent, on pourra envisager de l'implanter. Au contraire, une méthode coûteuse sur le plan computationnel sera rejetée, à moins qu'on puisse vérifier qu'elle apporterait une amélioration substantielle au taux de résolution d'un grand nombre de cas d'anaphores démonstratives, par exemple.

À titre de comparaison, la figure 5 montre la répartition des démonstratifs du corpus de Baudot (1992), réorganisée selon les types définis dans notre propre étude. Baudot (1992) ne détaillant pas davantage les fréquences du déterminant, le graphique ne montre que le détail de la répartition des pronoms démonstratifs. Le corpus de Baudot (1992:14) est constitué «de 803 échantillons de textes (d'environ 1000 à 1500 mots chacun) répartis en 15 genres littéraires». Les textes datent d'entre «1906 et 1967, la majorité [...] entre 1960 et 1967». «62 %» des textes viennent de «France» et «37 %» du «Canada».

Les données de Baudot (1992) qui ont servi à élaborer la figure 5 sont disponibles à l'annexe 9. Veuillez noter que puisque Baudot (1992) ne fournissait pas le contexte de chaque expression comptabilisée dans son répertoire, nous avons dû nous contenter de regrouper les données de la manière dont ils apparaissent dans la figure 5. Ces regroupements ne demandent pas de connaître le contexte dans lequel apparaissent les démonstratifs en question. Il suffisait de faire correspondre les

entrées de Baudot (1992) avec les lexies correspondantes, telles que ces dernières sont définies à l'annexe 1 de ce mémoire.

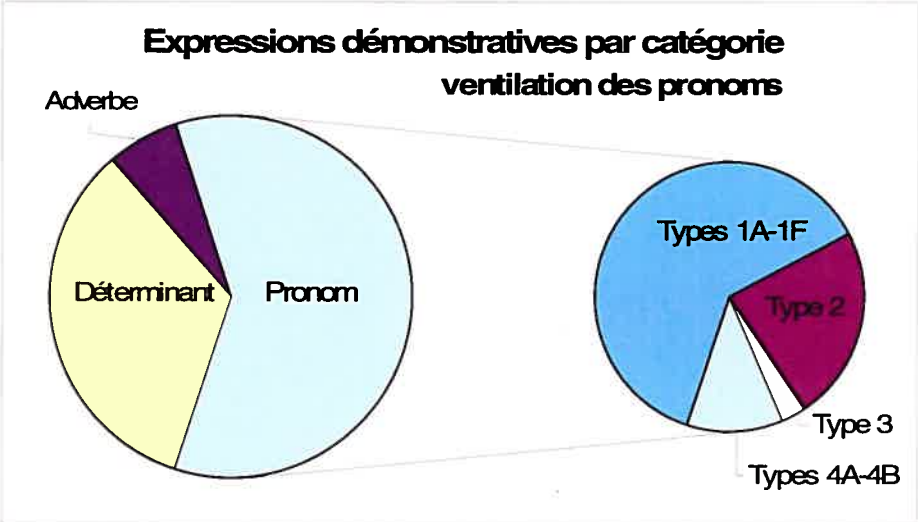

Figure 3 Expressions démonstratives par lexies : pronoms

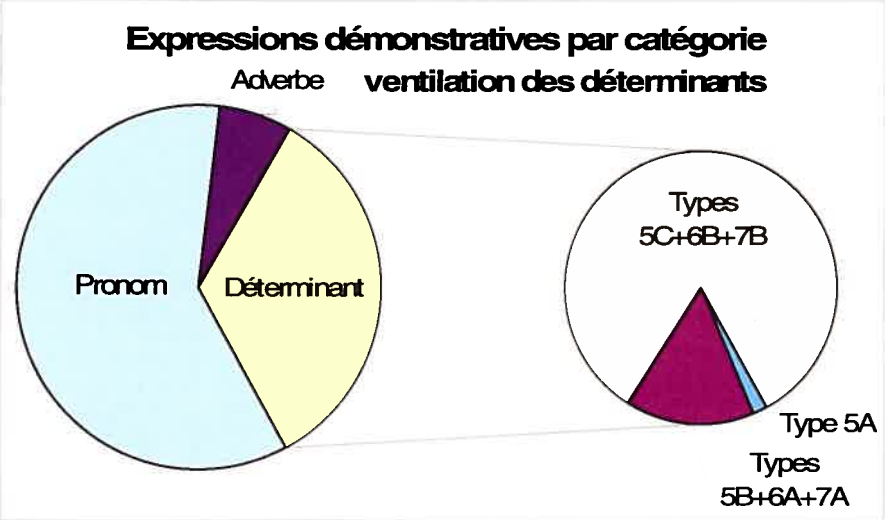

Figure 4 Expressions démonstratives par lexies : déterminants

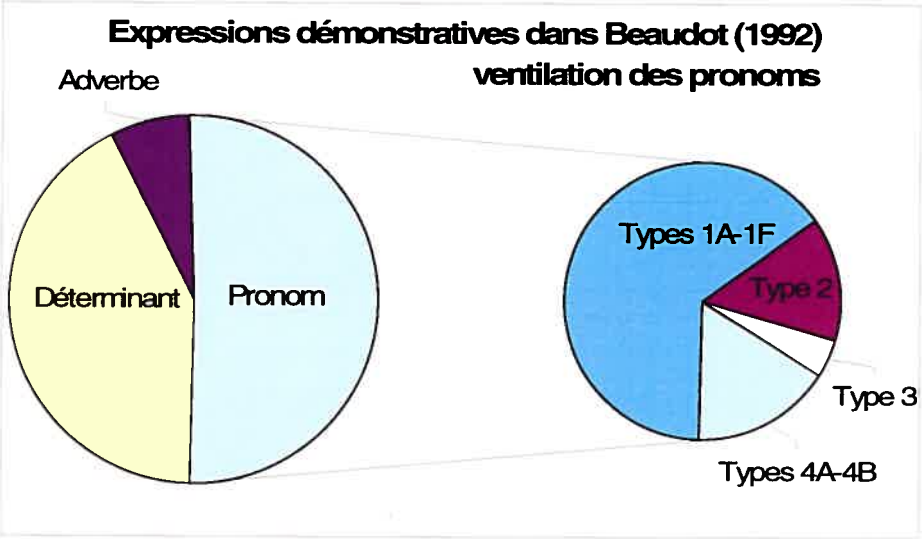

Figure 5 Expressions démonstratives par lexies dans Baudot (1992)

6.2 Vérification des hypothèses

6.2.1 Analyse des résultats pour les expressions de type 1A

Les expressions démonstratives de type 1A sont les expressions composées du pronom CE_1 et d'un complément sous la forme d'une proposition relative. Les 40 expressions de ce type constituent 7 % du total des expressions trouvées. Sur ce nombre, 78 % correspondent à l'hypothèse formulée, c'est-à-dire qu'elles ne constituent ni des endophores, ni des exophores.

> Ex. 1 : *Côté douceurs, à supposer qu'il vous reste de la place, le "pin-du-ri" est, contrairement à* ***ce*** *qu'on pourrait penser*$_{d8}$*, un pouding au pain.* (Voir-5)

Parmi les expressions de ce type qui ne correspondent pas à cette hypothèse, les 7 expressions qui proviennent du sous-corpus de *La Presse* se présentent d'une manière similaire. Elles sont anaphoriques et ont comme antécédent coréférent la proposition qui précède immédiatement et sont précédés d'une virgule. (Sauf un cas, auquel il manque la virgule. Il s'agit probablement là d'une faute de frappe.) On peut penser qu'il s'agit d'une sorte d'apposition. Le genre neutre de ce pronom répond à la nature de l'antécédent propositionnel, qui n'a pas de genre.

> Ex. 2 : *Il faudra donc s'assurer qu'il ne transmettra pas une mauvaise information aux agences d'évaluation de crédit (Équifax et Trans*

Union)$_{a14}$, ce$_{d14}$[a14] qui entacherait le dossier de crédit de la cliente$_{a15}$. (La Presse-51)

6.2.2 Analyse des résultats pour les expressions de type 1B

33 expressions démonstratives ont été identifiées comme ayant une forme de type 1B, c'est-à-dire une forme en «c'est [...] qui/que». Cependant, seulement 28 d'entre elles (85%) sont effectivement des cas de clivage véritables, comme l'exemple 3. Il est en effet possible de le paraphraser comme en 4, simplement en «retirant» la construction en clivage, sans en altérer le sens (sauf pour ce qui est de l'effet de focalisation produit par le clivage).

Ex. 3 : *C'$_{d26}$est sur lui que se fondent tous les espoirs.* (La Presse-49)
Ex. 4 : *Tous les espoirs se fondent sur lui.*

Dans les autres cas, comme dans l'exemple 5, on a affaire à un syntagme nominal attribut du sujet qui comporte une proposition relative. Le pronom démonstratif en position sujet de la proposition principale est alors anaphorique. On peut s'assurer qu'il ne s'agit pas simplement d'un cas de clivage ordinaire en tentant, dans l'exemple 6, de répéter l'opération qui a permit de construire l'exemple 4. On constate alors que l'expression «le genre de pièce» n'est plus apte à repêcher l'antécédent «La Visite de la vielle dame». Il faudrait remplacer l'article défini par un déterminant démonstratif, comme dans l'exemple 7, pour que ce syntagme nominal conserve le même référent.

Ex. 5 : *"La Visite de la vieille dame$_{a3}$ est une pièce que j'aime depuis longtemps et c'$_{d3}$[a3]est le genre de pièce que je veux que l'on présente au Rideau Vert, explique-t-elle d'entrée de jeu.* (Voir-14)
Ex. 6 : *La Visite de la vielle dame est une pièce que j'aime depuis longtemps et je veux que l'on présente le genre de pièce au Rideau Vert, explique-t-elle d'entrée de jeu.*
Ex. 7 : *La Visite de la vielle dame est une pièce que j'aime depuis longtemps et je veux que l'on présente ce genre de pièce au Rideau Vert, explique-t-elle d'entrée de jeu.*

Ceci dit, il paraît très difficile de formuler des principes formels simples permettant de distinguer entre un cas simple de clivage, où le pronom démonstratif n'est pas impliqué dans la cohésion textuelle, et les autres cas, plus complexes, où le pronom démonstratif est anaphorique. Cela semble pour l'instant hors de portée d'un traitement automatique réalisable en pratique.

Au pire des cas, une solution de rechange serait d'envisager de noter les formes les plus fréquentes d'expression ressemblant à un clivage simple, mais qui n'en sont pas. Par exemple, nous avons trouvé deux occurrences d'expressions de type «c'est le **seul** [...] qui» (ex. 8 et 9). Elles présentent toutes deux des pronoms démonstratifs anaphoriques.

> Ex. 8 : *Train est à la fois un garçon- il a à peine 18 ans- et un homme, car il$_{a8}$ doit subvenir à ses propres besoins$_{n6}$. Et à ceux^{n6} des autres$_{d6}$, car de tous les caddies qui peuplent ce livre$_{d7}{}^{d5}$, c'$_{d8}{}^{a8}$est le seul qui sache vivre dans le monde des Blancs.* (La Presse-46)
> Ex. 9 : *C'$_{d14}{}^{a14}$est épuisant, la poésie$_{a14}$. C'$_{d15}{}^{a14}$est le seul genre qui soit plus fatigant à lire qu'à écrire. "* (La Presse-47)

6.2.3 Analyse des résultats pour les expressions de type 1C

Les expressions démonstratives de type 1C sont les expressions de la forme « c'est le cas de » ou « ce n'est pas le cas de ». Moins de un pour cent des expressions démonstratives recensées dans le corpus sont de ce type. Les deux occurrences en question ont été trouvées dans le sous-corpus de *La Presse*. Dans les deux cas, l'hypothèse semble vérifiée, mais aussi peu de données ne nous permettent pas véritablement de trancher. L'exemple 6 présente une des deux occurrences en question. On y constate en effet que l'antécédent coréférent de «ce» est une proposition subordonnée située juste devant lui.

> Ex. 10 : *On lui promet que tout sera rétabli le mois suivant$_{a3}$. Mais ce$_{d3}{}^{a3}$ n'est pas le cas.* (La Presse-51)

6.2.4 Analyse des résultats pour les expressions de type 1D

Aucune expression démonstrative de la forme «C'est la question...» n'a été identifiée dans notre corpus. Il est donc impossible de vérifier l'hypothèse concernant le type d'expression 1D.

6.2.5 Analyse des résultats pour les expressions de type 1E

Les expressions démonstratives de la forme « c'est-à-dire » ont été classées dans le type 1E. Seulement 3 expressions démonstratives sur 579 sont de ce type. Elles ont été trouvées dans le sous-corpus du journal *Voir*. Aucune ne semble aller à l'encontre de l'hypothèse voulant que les expressions de ce type ne soient ni

endophoriques, ni exophoriques. L'exemple 11 illustre cela. Toutefois, encore une fois, un si petit nombre de données ne nous permet pas de nous prononcer avec suffisamment d'assurance sur la validité de notre hypothèse.

> Ex. 11 : *Je suis comme les gens de ma génération, c'$_{d10}$est-à-dire que je ne supporte pas d'écouter un morceau qui dure plus de quatre minutes.* (Voir-27)

6.2.6 Analyse des résultats pour les expressions de type 1F

Le type 1F inclut tous les cas d'expressions démonstratives comprenant un mot-forme appartenant à la lexie CE_1 et qui ne sont pas couverts par les types 1A à 1E. Dès lors, on pouvait prévoir que ce sous-ensemble comprenne toutes sortes de cas différents. Toutefois, nous ne nous attendions pas à autant de variété dans l'usage de CE_1 et, surtout, que ces «autres» usages soient aussi nombreux. En effet, près d'un quart des démonstratifs recensés dans le corpus ont été classé de type 1F. Cela représente presque deux tiers des expressions démonstratives comprenant la lexie CE_1 (types 1A à 1F). Sur ce nombre, moins de un dixième ont un comportement compatible avec l'hypothèse préalablement formulée pour le groupe d'expressions 1F. Cela montre que notre première classification était clairement déficiente pour cette lexie. Toutefois, nous disposons maintenant de beaucoup de données pouvant nous aider à remédier à la situation.

Dans l'hypothèse que nous avions élaborée au sujet des expressions de type 1F, nous supposions qu'il s'agirait d'anaphores et que l'antécédent serait une proposition, située très près devant l'expression. Il s'avère que 69 % des expressions de type 1F ont été étiquetées comme étant des anaphores. Sur ce nombre, 49 % sont situées dans la même phrase que leur antécédent et 38 % dans la phrase suivante. Donc, en ce qui concerne la proximité de l'antécédent, l'hypothèse semble tout de même refléter une part importante des données recueillies. Par contre, seulement 26 % des expressions endophoriques (anaphoriques et cataphoriques) de type 1F ont une proposition comme antécédent ; c'est au contraire les syntagmes nominaux qui prédominent (62 %).

La fonction des antécédents SN des expressions de type 1F semble différer des autres. En effet, on retrouve 15 % de SG et 25 % de COD, alors que les effectifs

totaux contiennent 26 % de SG et 17 % de COD. Le test de Pearson montre que la différence est statistiquement significative : $\chi^2(3 \text{ d.l.}, N = 67) = 16{,}786$, $p < 0{,}001$.

	théorique c	réel o	écart (o – c)	khi² (o – c)² / c
SG	26%*67 = 17	10	-7	2,882
COD	17%*67 = 11	17	+6	3,273
autres fonctions	28%*67 = 19	10	-9	4,263
non classés	29%*67 = 19	30	+11	6,368
total	66	67	1	16,786

Tableau 1 Calcul du χ^2 pour les types d'antécédents de type 1F

Il est possible que le nombre plus important d'antécédents COD soit dû indirectement à ce que ce les occurrences de type 1F ont une grande proximité à leur antécédent (56% d'antécédents dans la même phrase). L'antécédent aurait donc moins besoin d'avoir la saillance d'un SN en position sujet, la proximité jouant un plus grand rôle.

Cependant, le grand nombre d'antécédents non classés rend difficile l'interprétation de ces résultats. Tout en gardant à l'esprit que l'apport des effectifs de SG et de COD au calcul du χ^2 est suffisamment important pour que la différence demeure statistiquement significative[47], il serait bon de se pencher un moment sur ces syntagmes nominaux non classés. Les antécédents non classés sont ceux pour lesquels aucune fonction de la phrase canonique n'a été attribuée. Une hypothèse qui expliquerait le nombre plus faible de SG et le nombre plus grand d'antécédents non classés serait que parmi les expressions de type 1F, on trouve beaucoup de cas où le syntagme nominal qui aurait été le SG de la proposition subit une sorte de dislocation au profit du pronom CE_1. Nous en avons répertorié 17 au total, ce qui représente le quart des expressions de type 1F. L'exemple 12 en est une illustration.

> Ex. 12 : *La transparence*$_{a43}$, *c'*$_{a43}^{a43}$ *est l'absence de frontière entre le privé et le public.* (La Presse-49)

Dans cet exemple, selon notre schème d'annotation, *c'* a été classé *SG*, tandis que son antécédent, *La transparence*, a été classé dans *non classé*.

---

[47] 2,882 + 3,273 = 6,155, ce qui donne déjà, avec 3 d.l., $p < 0{,}1$

La dislocation, un terme employé en syntaxe, est ainsi définie par Neveu (2004:107) :

> *Opération de mise en relief d'un constituant de l'énoncé prenant la forme d'une construction dans laquelle prennent place un pronom régi par un verbe, et une réalisation lexicale dite «disloquée» placée avant ou après le verbe régissant, en relation de coréférence avec le pronom[.]*

La dislocation illustrée par l'exemple 8 a pour fonction de mettre en œuvre un certain type de focalisation. Plus précisément, selon Mel'čuk (2001:185), l'opération syntaxique de dislocation, aussi appelée détachement, est utilisée dans un grand nombre de langues pour réaliser la focalisation du «Sem-T» (i.e. le «thème sémantique»).

Il faut distinguer les cas de dislocation simple, comme en 8, des cas de dislocation avec prolepse. Selon Mel'čuk (2001:186), la prolepse est une opération sémantique, alors que la dislocation relève de la syntaxe profonde. De plus, en français, toute dislocation entraîne la mise en œuvre d'une antéposition (dislocation gauche) ou d'une postposition (dislocation droite) (Mel'čuk, 2001:187-188). Mel'čuk (2001:185) précise encore que ces dernières ont lieu au niveau de la syntaxe de surface.

Notons que Neveu (2004:130) appelle «extrapositions» l'ensemble des constructions disloquée, vocative, apposée, etc. Selon C. Touratier (cité dans Neveu, 2004:130-131), une extraposition antéposée fonctionne comme «support informatif»; en contrepartie, une extraposition postposée sert de «report informatif» à la phrase endocentrique à laquelle elle est adjointe.

Les exemples 13 et 14 illustrent respectivement les dislocations gauche et droite. Dans l'exemple 15 figure une prolepse, qui s'accompagne toujours d'une dislocation gauche en français. L'exemple 16 démontre qu'il est possible d'enchaîner les dislocations l'une à la suite de l'autre, récursivement. Ces exemples sont tous tirés du corpus écrit de Blasco-Dulbecco (1999:323-325).

Ex. 13 : *La vie c'est le bruit.* (Djian, *50 contre 1*, Folio p. 224)
Ex. 14 : *C'est gris et négligé, un mouton.* (Courchay, *Avril est un mois cruel*, Albin Michel, p. 77)
Ex. 15 : *La liberté relative, on y arrive plus facilement encore sans galon.* (Descares, Sous Offs 34)

Ex. 16 : *Cette bagnole, **son** mec c'était Franck Ocker et je me suis approché.* (Djian, *50 contre 1*, Folio p. 120)

Examinons maintenant plus avant les cas que visait notre hypothèse de départ. Par cela, nous voulons dire les expressions démonstratives confirmant cette hypothèse. C'est-à-dire, les cas où une expression démonstrative de type 1F avait effectivement pour antécédent coréférent la proposition précédente, à condition de ne pas se trouver dans une proposition y étant subordonnée.

Parmi toutes les expressions de type 1F, donc, les cas que visait notre hypothèse de départ sont peu nombreux dans le corpus. Cependant, ils sont encore plus rares dans le sous-corpus du *Voir* (5%) que dans celui de *La Presse* (12%). Parallèlement, le *Voir*, au total, contient un nombre inférieur d'expressions démonstratives de type 1F (62 contre 75). Au contraire, on retrouve environ le même nombre de cas de dislocation du SG par CE_l dans chacun des sous-corpus. Peut-être les cas visés par notre hypothèse de départ sont-ils plus fréquents, proportionnellement à toutes les expressions de type 1F, dans les textes plutôt typiques de *La Presse* que dans ceux qui sont caractéristiques du *Voir*.

6.2.7 Analyse des résultats pour les expressions de type 2

Les expressions démonstratives de type 2, c'est-à-dire les pronoms démonstratifs CECI, CELA et ÇA, au nombre de 81, constituent 14 % du total des expressions démonstratives recensées. Il y a en sensiblement le même nombre dans chacun des deux sous-corpus. Par contre, un pourcentage un peu plus élevé d'expressions démonstratives de ce type correspondent à l'hypothèse formulée dans les articles du *Voir* (62 %) que dans ceux de *La Presse* (50 %). Cela donne un taux de succès moyen de 56 %.

Rappelons que le comportement supposé de ces formes était d'être anaphorique. On faisait l'hypothèse que s'il y avait présence d'un antécédent coréférent, ce ne pouvait pas être un «groupe nominal strict». (voir référence à Corblin (1995) dans la section «Problématique», page 24) Sinon, il pouvait s'agir d'une anaphore associative, ou alors d'un cas où le référent est inféré implicitement du contexte précédent. Dans tous les cas, le fragment textuel fournissant l'antécédent devait se trouver à une distance n'excédent pas deux phrases avant l'anaphorique.

Bien que cela décrive une grande partie des occurrences du corpus, beaucoup de contre-exemples ont également été trouvés. Cela n'est pas dû à la distance. Au contraire, 95 % des expressions démonstratives de type 2 pour lesquels on a identifié un antécédent (74 % du total des expressions de ce type) sont à une distance d'une phrase ou moins de celui-ci.

Une possibilité serait de vérifier s'il serait avantageux d'augmenter l'importance accordée aux propositions parmi les candidats antécédents de ce type d'expression. On observe en effet que parmi les expressions démonstratives construites autour d'un pronom démonstratif, les anaphores démonstratives de type 2 sont parmi celles qui récoltent le plus d'antécédents propositionnels : 53 % de leurs antécédents sont des propositions, contre 30 % pour l'ensemble des expressions démonstratives pronominales. Les syntagmes nominaux représentent de leur côté 35 % des antécédents des expressions anaphoriques démonstratives de type 2. Il y a donc 51 % plus d'antécédents propositionnels que d'antécédents syntagmes nominaux chez les expressions endophoriques de type 2. Une meilleure formulation de la méthode de localisation des antécédents des expressions de ce type devrait donc probablement accorder une plus grande importance aux propositions.

6.2.8 Analyse des résultats pour les expressions de type 3

Les expressions démonstratives de type 3 sont composées des occurrences du corpus se rapportant aux lexies suivantes : CELUI-CI, CELUI-LÀ, CELLE-CI, CELLE-LÀ. On en a retrouvé un total de 11 dans les deux sous-corpus, ce qui représente 1,9 % du total des expressions démonstratives recensées. De ce nombre, 55 % vérifient l'hypothèse avancée pour tenter de prédire leur comportement. Nous rapportons ici l'hypothèse formulée dans une section antérieure :

> *L'antécédent est le SN de même genre et de même nombre le plus proche qui correspond aux contraintes sémantiques imposées par le gouverneur syntaxique (Verbe ou parfois Nom, probablement) de l'expression démonstrative. Il est situé avant l'expression démonstrative et dépend syntaxiquement d'un verbe. C'est un antécédent coréférent direct.*

Cette hypothèse laissant beaucoup de place à l'interprétation et les effectifs étant assez réduits, nous essaierons de montrer concrètement ce à quoi ressemble l'utilisation des expressions de ce type. Les exemples suivants serviront à illustrer les

cas où nous avons jugé que l'hypothèse était vérifiée (ex. 17 à 19), ainsi que ceux où elle ne l'était pas (ex. 20 à 22). Nous laissons au lecteur la latitude d'opter pour une analyse différente.

> Ex. 17 : *Vous savez qu'à l'échéance, le 1er décembre 2013, on vous remboursera un montant de 25 $ par action de " Capital Yield Share "$_{a22}$. **Celle-ci**$_{d22}{}^{a22}$ fait l'objet d'une garantie émise par Merrill Lynch et TD Global Finance.* (La Presse-59)
> Ex. 18 : *Les commissaires Marie-Paule Vial et Guy Cogeval ont donc bien raison lorsqu'ils écrivent dans le catalogue que "plus qu'un espace géographique délimité par des frontières issues d'un morcellement administratif, la Provence est avant tout un territoire de l'imaginaire*$_{a11}$*... et **celui-là**$_{d11}{}^{a11}$ ne connaît pas de limite".* (Voir-12)
> Ex. 19 : *Le rappeur annonçait récemment son intention de mettre sur pied Def Jam Left, une étiquette "pro-artistes qui, par ses contrats moins lucratifs, allégerait la pression de performance sur les palmarès aux artistes plus à gauche"*$_{a3\text{-}a2}$ *Dans cette optique*$_{d2}{}^{a2}$*, il espère faire des Roots les premiers signataires sur **celle-ci**$_{d3}{}^{a3}$, confirmant que les deux parties sont présentement en pourparlers.* (Voir-23)

Dans les cas où l'hypothèse choisit le mauvais antécédent, il aurait parfois été possible d'éviter l'erreur. Par exemple, pour corriger le tir dans l'exemple 3.3.5, il suffirait de préciser que le pronom démonstratif ne peut pas dépendre directement de son antécédent (en syntaxe de dépendance) ou que l'antécédent ne peut pas c-commander l'anaphorique (en grammaire syntagmatique, selon Reinhart (1983:30)).

Également, dans l'exemple 3.3.6, en incorporant la méthode de résolution par lots proposée par Corblin (1995:187-191)[48], on aurait pu arriver au bon résultat. La contrainte de disjonction référentielle s'appliquerait ici entre le sujet de la dernière phrase et la substitution (ou ellipse, selon l'analyse) dans son attribut. L'association du syntagme *un événement à ne pas manquer* à l'attribut du sujet de la phrase suivante éliminerait alors la possibilité d'avoir le même antécédent pour le démonstratif *celui-ci* en position sujet.

> Ex. 20 : *" Si le budget annuel exige une entrée de fonds supplémentaire, poursuit notre expert, François*$_{a3}$ *aurait l'opportunité de travailler comme contractuel ", comme **celui-ci**$_{d3b}{}^{a3}$ en a lui-même fait l'hypothèse.* (La Presse-55)
> Ex. 21 : *Glacées ou bien chaudes, les tisanes se boivent à longueur de journée et d'année et s'il existe des stars connues et reconnues dans le monde*

[48] Voir la section 3.1.3.3 de ce mémoire, page 15.

de l'infusion comme la camomille, le tilleul ou la verveine, il existe des centaines de starlettes qui attendent leur jour de gloire$_{a3}$ ou le retour de ***celui-ci***$_{d3}^{a3}$*, comme le cynorrhodon, fruit de l'églantier qui serait riche en vitamine C, ou bien le tussilage, plante vivace à fleurs jaunes, qui calmerait efficacement la toux$_{a4}$.* (Voir-10)
Ex. 22 : *Ce$_{d1}$ n'est pas souvent que l'on vous parle d'un récital de fin de doctorat$_{a2b}$ comme d'un événement à ne pas manquer (qui plus est, gratuit!).* ***Celui-ci***$_{d2}^{a2b}$ *en est un.* (Voir-29)

6.2.9 Analyse des résultats pour les expressions de type 4

Le groupe des expressions démonstratives de type 4 est constitué des pronoms démonstratifs non neutres sans particule adverbiale (les lexies CELUI et CELLES) avec leur complément. Nous avions divisé ce groupe en deux sous-catégories, selon quelques différences qui étaient apparues dans nos recherches empiriques préliminaires quant à leur comportement. Le sous-groupe 4A comprend les pronoms démonstratifs non neutres dont le complément est une proposition relative ; le sous-groupe 4B, ceux qui sont accompagnés d'un complément du nom ou d'un participe passé.

On dénombre 17 expressions de type 4A et 22 de type 4B dans le corpus. Au total, les expressions de type 4A et 4B composent 7 % du total des expression démonstratives du corpus. Il semble à première vue y avoir plus d'expressions de type 4B et moins d'expressions de type 4A dans le sous-corpus de *La Presse* que dans celui du *Voir*. Cependant, les effectifs des sous-corpus sont trop petits pour que nous puissions vérifier la signification statistique de cette différence à l'aide du test de Pearson.

L'hypothèse expérimentale prévoit, pour les expressions de type 4A et 4B, que l'antécédent soit un substantif non coréférent de même genre que le pronom démonstratif, mais de nombre quelconque. On ajoute que si l'on trouve une construction dans laquelle un candidat antécédent dépend (syntaxiquement) directement ou indirectement d'un autre candidat, il est plus probable que ce soit ce dernier qui soit le véritable antécédent. Par exemple, si deux substantifs *ringuette* et *équipe* sont trouvés qui correspondent au critère de genre identique, et que le substantif *ringuette* occupe la fonction de complément du nom auprès du substantif

équipe (*équipe de ringuette*), on devrait préférer le choix du substantif *équipe* comme candidat antécédent.

Pour les expressions de type 4A, l'hypothèse propose que l'antécédent soit la tête du premier SN se trouvant devant l'expression, en tenant compte des autres contraintes formulées plus haut. Pour les expressions de type 4B, l'hypothèse est un peu moins restrictive : l'antécédent est la tête d'un SN se trouvant dans la même phrase et devant l'expression démonstrative.

Cette différence dans la formulation des sous-hypothèses peut possiblement expliquer une partie de l'écart entre les résultats obtenus pour ces deux sous-groupes. En effet, le comportement des expressions de type 4B correspond beaucoup mieux à leurs hypothèses respectives. On obtient un taux de succès de 77% pour l'hypothèse sur les expressions de type 4B, contre 29% seulement pour l'hypothèse sur les cas de type 4A.

D'autre part, comme nous l'avons souligné précédemment, les expressions de type 4B ont beaucoup plus d'antécédents «nom seul» (87%) que ceux de type 4A (45%). Ces derniers comptent en revanche davantage de syntagmes nominaux comme antécédent (4A 55% ; 4B 13%).

6.2.10 Analyse des résultats pour les expressions de type 5A

Les expressions démonstratives de type 5A, c'est-à-dire les expressions en CE DERNIER, représentent 0,5 % du corpus. Toutefois, on les retrouve exclusivement dans les articles provenant du deuxième sous-corpus, qui est un échantillon de l'hebdomadaire culturel *Voir*. Toutes sont exactement conformes à l'hypothèse formulée préalablement. Bien que ces résultats soient encourageants, il nous en faut examiner la signification statistique avant de se prononcer davantage. En effet, l'effectif réduit porte tout de même au doute : il n'est que de 3.

Mais tout d'abord, une illustration du phénomène en question :

> Ex. 23 : *Il y a également EPJ 327, un jeune homme entiché lui aussi de CKZ 114, qui le lui rend bien, puis quelques personnages très secondaires dont l'utilité est de donner voix aux* <u>*prisonniers*</u>$_{a5}$. *Au-dessus de la tête de* <u>***ces derniers***</u>$_{d5}^{a5}$, *il y a une épée de Damoclès.* (Voir-30)

Une hypothèse qui expliquerait que les expressions de ce type se retrouvent dans l'hebdomadaire culturel plutôt que dans le quotidien serait que le premier est plus susceptible de comprendre des passages narratifs brefs où le rédacteur doit mettre en relief les relations entre divers personnages. (cf. Charolles, 1995a:90) Par exemple, le *Voir* est plus susceptible de comprendre des résumés, des critiques ou des compte-rendu de pièces de théâtre, de films ou de romans.

En effet, pour résumer une histoire, le moyen le plus communément utilisé consiste à décrire les grandes lignes des actions des principaux personnages, en soulignant au passage les relations que ceux-ci entretiennent par l'utilisation des lexèmes appropriés (frère, mère, époux, etc.). Or, dans un tel contexte, comme le souligne Charolles (1995a:89), l'utilisation judicieuse de certaines «formes anaphoriques spécialisées dans le contrôle des risques d'ambiguïtés» devient cruciale.

Sur ce plan, une expression anaphorique non standard comme CE DERNIER (non standard au sens où elle n'entre pas bien dans les catégories habituellement considérées par la grammaire traditionnelle, ou même par la linguistique structuraliste, quand il s'agit d'anaphore), si l'on considère son rôle dans la langue écrite, se compare tout de même très bien au pronom démonstratif CELUI-CI. Pour preuve, Charolles (1995a:89) tire de Veland (1989) des données qui permettent de conclure à un rapport de « 1 *ce dernier* pour 1,5 *celui-ci / là* » dans un échantillon de romans contemporains. Dans les données extraites de notre propre corpus, on obtient une proportion de 1 pour 3,7. Ce rapport inférieur pourrait être expliqué par les types de textes étudiés. Dans le sous-corpus du *Voir*, on trouve effectivement un rapport de 1 CE DERNIER pour 2,3 CELUI/CELLE -CI/-LÀ .

En somme, les résultats obtenus pour ce type d'expression démonstrative sont intéressants. Toutefois, ils devront éventuellement être confirmés par la collecte d'un plus grand nombre d'exemples attestés.

6.2.11 Analyse des résultats pour les expressions de type 5B, 6A et 7A

Nous avons décidé de regrouper les trois sous-groupes d'expressions 5B, 6A et 7A. Il y trois raisons pour cela. Premièrement, il y avait trop peu d'occurrences de type 6A ou 7A. On dénombre uniquement 3 expressions de chacun de ces deux types dans l'ensemble du corpus. Par conséquent, il nous fallait procéder à un regroupement de ces effectifs pour pouvoir en tirer des informations statistiquement significatives. Deuxièmement, les hypothèses formulées pour les types 5B, 6A et 7A sont presque les mêmes. Troisièmement, au niveau formel, ces trois sous-groupes n'ont été séparés que parce que nous avions choisi de classer nos sous-hypothèses en les répartissant entre les lexies démonstratives retenues. En réalité, il est aisé de reformuler la description de ces expressions de manière uniforme pour les trois sous-groupes.

On retrouve en tout 30 occurrences du déterminant démonstratif suivi d'un nom temporel ou métatextuel, avec ou sans particule adverbiale démonstrative. Cela représente 5,2 % des expressions démonstratives du corpus. Dans le sous-corpus de *La Presse*, cette proportion est de 3,8 %, tandis qu'elle est de 6,6 % dans celui du *Voir*. Il y a donc 73 % plus de ce type d'expressions dans notre exemplaire de l'hebdomadaire culturel quand celui du quotidien. Le test de Pearson peut nous servir à vérifier la signification statistique de cette différence.

H_0 : Il n'y a pas de différence sensible entre les deux populations représentées par ces deux sous-corpus. Les différences sont dues uniquement au hasard de l'échantillonnage.

H_1 : Il existe une différence réelle entre les deux populations représentées par ces deux sous-corpus. Les différences ne sont pas dues uniquement au hasard de l'échantillonnage.

Effectifs réels : $11 + (289-11) = 289$

$19 + (288-19) = 288$

Effectifs théoriques : $((30/577) \times 289) + (289 - ((30/577) \times 289)) = 289$

$((30/577) \times 288) + (288 - ((30/577) \times 288)) = 288$

	Réel		Théorique		Écart		χ^2	
	$o_{La\ Presse}$	o_{Voir}	$c_{La\ Presse}$	c_{Voir}	$(o - c)$		$(o - c)^2 / c$	
5B+6A+7A	11	19	15	15	-4	4	1,067	1,067
Autres	278	269	274	273	4	-4	0,058	0,059
Total	289	288	289	288	0	0	1,125	1,125
Grand total	577		577		0		2,250	

Tableau 2 Calcul du χ^2 pour la fréquence des démonstratifs de type 5B, 6A et 7A

Degrés de liberté : $n = 1 \times 1 = 1$
Khi carré : $\chi^2 = 2{,}250$
Probabilité de $\chi^2 \geq 2{,}250$ dans le cas d'une distribution aléatoire : $p = 0{,}1336$

Le test de Pearson ne démontre pas clairement l'improbabilité de l'hypothèse nulle. En effet, il semble que la différence entre les deux échantillons ait environ 13 % de chances de pouvoir avoir été causée par un simple hasard. C'est beaucoup trop pour pouvoir rejeter cette explication. Nous ne pouvons donc rien conclure au sujet de la différence entre les effectifs des deux sous-corpus. Toutefois, comme une probabilité de 13 % n'est tout de même pas énorme, il faut admettre que ces données mériteraient une enquête plus poussée. Il faudrait alors puiser parmi les articles non utilisés de chacun des deux journaux, ou d'autres exemplaire de ces mêmes journaux, d'autres occurrences d'expressions démonstratives de type 5B, 6A et 7A. On pourrait alors vérifier si la tendance se confirmerait ou, au contraire, serait noyée dans les données supplémentaires.

Revenons maintenant aux hypothèses avancées pour décrire l'utilisation du déterminant démonstratif suivi d'un substantif temporel ou métatextuel.

Dans le cas où le démonstratif est suivi de la particule adverbiale «-là» (type 7A), il fallait tout d'abord vérifier s'il était possible de trouver un candidat antécédent coréférentiel : une expression temporelle ou métatextuelle compatible située devant l'expression démonstrative. La distance maximale de recherche était fixée à deux phrases en arrière.

Si l'on ne trouvait pas un tel antécédent, ainsi que dans tous les autres cas (types 5B et 6A), on devait conclure qu'il n'y avait pas d'antécédent textuel. L'expression devait alors référer à l'unité textuelle ou temporelle courante, respectivement.

Cette analyse est apparemment incomplète, car on obtient le taux de succès très moyen de 53 %. Ce résultat vaut pour l'ensemble des expressions de type 5B, 6A et 7A.

6.2.12 Analyse des résultats pour les expressions de type 5C, 6B et 7B

Les expressions de type 5C, 6B et 7B sont les syntagmes nominaux dotés d'un déterminant démonstratif et qui n'entrent dans aucunes des catégories précédentes. Rappelons que nous avions tout d'abord classé ces expressions démonstratives dans trois sous-groupes, selon la présence ou l'absence d'une particule adverbiale accompagnant le déterminant. Le type 6B représente les déterminants démonstratifs suivis de la particule adverbiale « -ci » ; le type 7B, ceux suivis de la particule « -là ». En cas d'absence de particule adverbiale démonstrative associée, l'expression démonstrative était classée de type 5C.

Nous avons regroupé les trois sous-groupes d'expressions 5C, 6B et 7B pour trois raisons. Premièrement, il y avait trop peu d'occurrences de type 6B ou 7B. On compte seulement 2 expressions de type 6B et 4 de type 7B dans tout le corpus. Il nous fallait donc procéder à un regroupement de ces effectifs pour pouvoir en tirer des informations statistiquement significatives. Deuxièmement, les hypothèses formulées pour les types 5C, 6B et 7B sont exactement les mêmes. Troisièmement, au niveau formel, ces trois sous-groupes n'ont été séparés que parce que nous avions choisi de classer nos sous-hypothèses en les répartissant entre les lexies démonstratives retenues. En réalité, il est aisé de reformuler la description de ces expressions de manière uniforme pour les trois sous-groupes. Autrement dit, ce sont les syntagmes nominaux dotés d'un déterminant démonstratif (CE_2, CE -CI et CE -LÀ) qui n'entrent dans aucun des autres cas plus spécifiques.

Les expressions de types 5C, 6B et 7B représentent 28 % du total des expressions démonstratives du corpus et 83 % de celles impliquant un déterminant démonstratif. Il y en a davantage dans le sous-corpus tiré du *Voir* (85 occurrences) que dans celui tiré de *La Presse* (76 occurrences).

L'hypothèse que nous avions tout d'abord formulée pour tenter de prédire le comportement de cette grande classe d'expressions était très complexe. En effet, elle comprenait toute une série de cas susceptibles d'être rencontrés dans le corpus. Ces différentes possibilités étaient structurées hiérarchiquement, sous la forme « si… alors… sinon… alors… ». Or, il s'est rapidement avéré que cette manière d'ordonner les utilisations des expressions démonstratives de ce type ne fonctionnait pas. Après plusieurs essais infructueux, nous avons abandonné l'idée de tester cette hypothèse sur toutes les occurrences pertinentes. Pour justifier ce choix, nous invoquerons des raisons pragmatiques : cette sorte de test est long et fastidieux à effectuer manuellement, et il se trouve que le corpus compte 155 occurrences d'expressions de type 5C. Donc, comme il semblait clair après quelques essais que nous devions rejeter cette hypothèse, nous n'avons pas achevé de la tester sur toutes les expressions démonstratives pertinentes du corpus.

Cependant, il est possible de tirer une somme importante d'informations des données recueillies. L'hypothèse initiale prévoyait plusieurs cas de figures, qui correspondent à divers types d'anaphores. Comme il en fut question dans le chapitre « Problématique » de ce mémoire, plusieurs auteurs ont élaboré des typologies des relations anaphoriques, exophoriques, déictiques, de coréférence, etc. Pour les besoins de notre étude, il fut nécessaire de faire une synthèse de ces diverses approches ; ceci dans le but d'en arriver à une classification qui éclaire suffisamment les données recueillies, du point de vue que nous nous proposions d'adopter dans le cadre de cette recherche. Pour ce qui est des expressions démonstratives avec déterminant démonstratif, bien que nous ayons renoncé à vérifier l'hypothèse initiale, celle-ci fut tout de même utile, avec quelques ajustements, pour la classification des données obtenues. Le tableau 3 présente le décompte auquel nous sommes arrivés. Veuillez notez que ce décompte n'a été effectué que pour les expressions de type 5C, puisque les expressions de type 6B et 7B étaient encore, à cette étape, traités séparément. Toutefois, puisque les expressions de type 6B et 7B représentent moins de 4 % des occurrences l'ensemble de ce groupe, les chiffres suivant en sont assez représentatifs.

Type de lien	Sous-corpus *La Presse*	Sous-corpus *Voir*	**Total**
Répétition de la tête du syntagme	10	13	**23**
Nom de base	0	0	**0**
Synonymie et quasi-synonymie	2	10	**12**
Nominalisation (S_0)	3	1	**4**
Adjectivation (A_0)	1	0	**1**
Hyponymie	4	1	**5**
Noms d'actant (FL S_i où i>0, ou $S_{(circ)}$)	3	0	**3**
Autres liens lexicaux	5	4	**11**
Reclassification [sans lien lexical ; peut avoir, ou non, un autre type de lien (i.e. paralinguistique)]	21	39	**60**
Exophore, déixis… (pas d'antécédent textuel)	24	14	**40**
Total	**73**	**82**	**155**

Tableau 3 Type de lien anaphorique pour les expressions de type 5C

Notre hypothèse initiale prévoyait de revenir en arrière jusqu'à une distance maximale de 4 phrases pour les cas de répétition de la tête du syntagme et jusqu'à une distance de 1 phrases pour les autres cas. Le graphique suivant montre la répartition des distances en phrases dans antécédents des expressions endophoriques de types 5C, 6B et 7B. Rappelons qu'une distance de 0 phrases signifie que l'antécédent se trouve dans la phrase courante, c'est-à-dire dans la même phrase que l'expression démonstrative à laquelle il se rapporte. La médiane est de 1 phrase et la «moyenne»[49] est 1,72 phrases.

[49] Voir la note 46 au bas de la première page de ce chapitre.

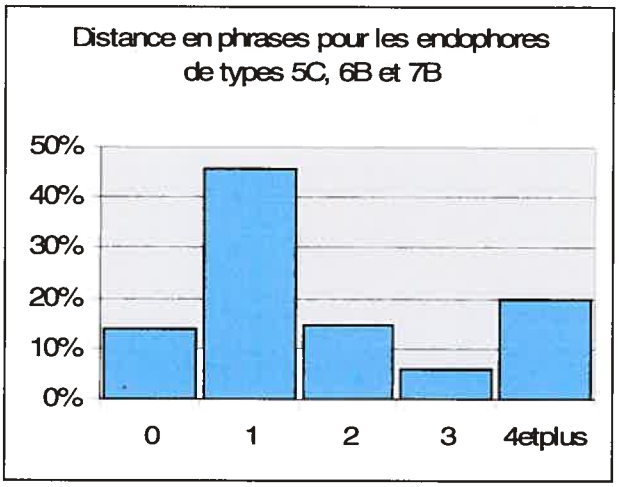

Figure 6 Distance à l'antécédent pour les types 5C, 6B et 7B

Selon les chiffres du tableau 3, les cas de répétition de la tête du syntagme représentent 20 % des cas d'endophore pour les expressions de type 5C. D'autre part, si l'on examine les pourcentages illustrés dans la figure 4, on obtient que 40 % des antécédents sont à une distance de 2 phrases ou plus, pour l'ensemble des expressions de type 5C, 6B et 7B. Or, nous avions prévus nous limiter à une distance de 1 phrase pour tous les cas d'endophore, excepté les cas de répétition de la tête du syntagme. L'hypothèse telle qu'elle était formulée aurait donc raté plusieurs antécédents. Cependant, la solution n'est probablement pas d'augment la distance en terme de phrases. En effet, les données suggèrent que le gain en terme d'antécédents récupérés, versus le coût computationnel du traitement d'une phrase supplémentaire, serait probablement minime au-delà d'une distance de 1 ou 2 phrases. Comme plusieurs auteurs (cf. Mitkov, 2002:39) l'ont déjà suggéré, peut-être faudrait-il plutôt mesurer la distance selon un autre critère. Par exemple, en termes de structure syntaxique ou discursive.

Notons que nous avons identifié un antécédent pour 121 expressions démonstratives de types 5C, 6B et 7B. Cela représente 75 % des expressions démonstratives de ce groupe. Il faut aussi remarquer que seulement 4 antécédents sur 121 sont situés après l'expression démonstrative, ce qui en fait des cas de cataphore. L'anaphore compte ainsi pour 97 % des phénomènes endophoriques reliés aux expressions de types 5C, 6B et 7B. Donc, même si notre hypothèse négligeait les cas

de cataphore, somme toute, cela semble tout à fait adéquat, du moins dans un premier temps.

Les antécédents des expressions endophoriques des types 5C, 6B et 7B sont pour 81 % des syntagmes nominaux et pour 13 % des propositions. Les antécédents syntagmes nominaux ont été classés selon leur fonction syntaxique. Le tableau suivant résume ces données.

Fonction	SG	COD	COI	CdN	Autres	Non classé	Total
Nombre	27	13	9	16	6	24	95
Proportion	28 %	14 %	9 %	17 %	7 %	25 %	100 %

Tableau 4 Fonction des antécédents SN des expressions de types 5C, 6B et 7B

6.2.13 Analyse des résultats pour les expressions de type 8

Les adverbes démonstratifs (les lexies ICI et LÀ) forment le dernier groupe d'expressions démonstratives, le type 8. Étant donné le peu d'indices empiriques et le manque de documentation du rôle anaphorique ou déictique de ces formes, nous avions préféré ne pas formuler d'hypothèse particulière à leur sujet.

Les adverbes démonstratifs représentent 6 % des expressions démonstratives relevées. On en dénombre 19 occurrences dans le sous-corpus *La Presse* et 18 occurrences dans le sous-corpus *Voir*, ce qui donne sensiblement la même proportion du total de chacun. On dénombre 6 cas d'adverbes démonstratifs anaphoriques, ce qui représente 22 % des expressions démonstratives adverbiales. La répartition des distances à l'antécédent pour ces adverbes se rapproche de ce qui est observé pour les pronoms démonstratifs. Les antécédents sont des syntagmes nominaux dans 63 % des cas. Il n'y a pas d'exemple de cataphore impliquant un adverbe démonstratif dans notre corpus.

Parmi les éléments que nous n'avions pas prévus, on trouve deux cas distincts où deux adverbes démonstratifs sont associés dans une même expression. Premièrement, on peut remarquer une occasion dans le sous-corpus de *La Presse* où deux adverbes démonstratifs forment en réalité une seule unité lexicale : ICI ET LÀ. L'exemple 24 présente la phrase où on peut observer ce phénomène. Il semble clair qu'ici, si jusqu'ici nous avons procédé en faisant le décompte par expressions impliquant une lexie donnée, nous devrions réviser la liste des lexies que comprend

le groupe d'expressions de type 8. Il résulterait que nous devrions également retrancher 1 du nombre d'occurrences trouvé, ce qui donnerait 18.

> Ex. 24 : *L'auteur principal (Marc Cherry), fait la bible, le premier et le dernier épisode et en signe quelques autres* ***ici et là***$_{d8et9}$. (La Presse-28)

Le deuxième cas est plus subtil. On serait en effet tenté d'arriver à la même conclusion avec l'expression *d'ici là*. Toutefois, il ne s'agit pas d'une association obligatoire entre les deux adverbes, qui ait un sens particulier en soi, i.e. non compositionnel. En effet, dans les expressions *d'ici 2013* (ex. 26) et *d'ici quelques années* (ex. 27), les expressions *2013* et *quelques années* jouent grosso modo le même rôle dans la construction du sens de l'énoncé que *là* dans *d'ici là*. Ils fournissent chacun à leur manière un point de repère dans le temps (relatif ou absolu, précis ou non) qui permet de construire un intervalle temporel à partir d'*ici*, c'est-à-dire 'maintenant'. Dans le cas de l'exemple 25, le contexte nous permet de déduire que *là* fait référence au 'moment où *il* prendrait sa retraite s'il la précipitait d'un an'.

> Ex. 25 : *En tirant profit des journées de maladie accumulées, il pourrait même précipiter sa retraite*$_{a1}$ *d'un an- " si je suis chanceux et que je ne tombe pas malade* ***d'ici là***$_{d1b}^{a1}{}_{d1a}$ *", précise-t-il.* (La Presse-55)

> Ex. 26 : *Comme l'action se négocie actuellement autour de 17,90 $, le rachat à 25,00 $ représente une hausse de quasiment 40 %* ***d'ici 2013***$_{d23}$. (La Presse-59)

> Ex. 27 : *Ayant occasionnellement collaboré depuis 2003 à quelques pièces et remix, notamment avec Kanye West, il avoue songer de plus en plus à un retour sur un album complet* ***d'ici*** *quelques années*$_{d5}$. (Voir-23)

Chapitre 7 Conclusion

Le but de cette étude était de tenter de réduire la complexité du problème de la résolution des anaphores en cherchant à tirer parti des différences entre le comportement textuel des expressions démonstratives et celui des autres formes linguistiques entrant en jeu dans la cohésion textuelle. Nous avons fait l'hypothèse qu'il serait possible et même utile de considérer les particularités de certaines classes de moyens linguistiques et que plusieurs de ces particularités attendent encore d'être découvertes. Nous nous sommes concentré sur les démonstratifs, en essayant de trouver empiriquement ce qui les caractérise. Nous n'avons pas supposé que l'ensemble des expressions démonstratives formait une classe homogène, mais nous nous sommes plutôt attardé à établir les caractéristiques de l'utilisation d'un grand nombre de types d'expressions démonstratives. De plus, en partant du principe qu'une étude scientifique de la cohésion textuelle fournirait des résultats plus riches si l'on se donnait d'abord une idée précise de ce que l'on recherche, nous avons pris le temps d'élaborer une série d'hypothèses. Parfois sommaires, parfois plus complexes, elles étaient toutes pensées en fonction du but que nous nous étions fixé : procurer des conclusions utiles pour l'analyse automatique de textes. Toutefois, la variété des emplois, quand on l'explore à travers une étude de corpus, est telle que nous n'avons pu qu'entamer sa description.

À travers l'examen des données recueillies dans notre étude de corpus, nous avons pu constater les avantages et les limites de notre approche. Certaines formes linguistiques connaissent un emploi très spécialisé, comme CE DERNIER. Leur mode de fonctionnement est donc plus facile à isoler, mais leur application est par le fait même réduite. Nous croyons tout de même que l'étude de ces emplois spécialisés est importante, car ils sont probablement plus nombreux qu'on ne le soupçonne. D'autres lexies peuvent entrer dans des constructions syntaxiques qui transforment radicalement leur comportement, comme CE_1. Voilà d'ailleurs un des avantages d'une étude de corpus : faire ressortir ces expressions et constructions spécialisées qui, faute d'entrer facilement dans un système structuré, sont souvent oubliées par les grammaires.

Nous avons également fait quelques trouvailles intéressantes concernant des faits plus généraux. Le pronom démonstratif CE_1 semble dévier de la tendance générale en ce qui concerne la fonction syntaxique des ses antécédents nominaux. En effet, on suppose normalement que les SN en position sujet devraient avoir plus de chances d'être l'antécédent d'une expression anaphorique, pour deux raisons. La première est triviale : c'est tout simplement qu'il y a plus de sujets que d'objets directs ou indirects, parce que tous les verbes ont un sujet, mais n'ont pas tous un objet. La deuxième est que dans une langue comme le français ou l'anglais, où l'ordre des constituants est relativement fixe, le sujet est normalement placé en début de phrase et acquière ainsi une grande proéminence. Le locuteur peut d'ailleurs faire varier le SG en fonction de ce critère, en utilisant une construction passive, par exemple. Or, nos données indiquent que CE_1, contrairement aux autres formes recensées dans cette étude et à ce qui est généralement dit à ce sujet dans la littérature, a davantage d'antécédents nominaux COD que SG. La raison en est encore indéterminée et le sujet mériterait d'être exploré de manière plus approfondie. En tous les cas, il serait sûrement important d'en tenir compte dans l'élaboration d'un algorithme de résolution des anaphores. Selon nos données et celles de Baudot (1992), les occurrences de CE_1 forment environ un tiers des occurrences de démonstratifs d'un court texte écrit en français.

De plus, nous avons tenu dans notre étude à considérer les phénomènes de cohésion textuelle «marginaux», notamment : ceux qui n'impliquent pas la coréférence, comme l'identité lexicale et l'anaphore associative; ceux qui impliquent plus que des syntagmes nominaux, comme les anaphores indirectes impliquant un verbe ou les antécédents formés par une proposition; ainsi que le rôle des adverbes démonstratifs dans l'endophore et la déixis. Cela nous a permis de mettre en évidence, par exemple, que seulement 35 % des antécédents attribués aux pronoms CECI, CELA et ÇA sont des syntagmes nominaux. Comme la plupart des algorithmes de résolution des anaphores se limitent à l'heure actuelle à tenter d'identifier les liens de coréférence entre certains SN, il est facile de constater qu'il reste encore beaucoup de problèmes à résoudre.

D'un autre côté, notre étude a permis de confirmer certaines données générales sur les démonstratifs ou les phénomènes de cohésion textuelle en général. Il y a davantage d'occurrences de pronoms démonstratifs que de déterminants démonstratifs. La distance à l'antécédent en termes de phrases est plus petite pour les pronoms démonstratifs que pour les SN avec déterminant démonstratif. Ces distances se comparent grosso modo à celles qu'on observe pour les pronoms personnels et les descriptions définies, respectivement.

En général, toutefois, nous avons pu constater que la tactique de « diviser pour régner » s'applique mal au problème de la résolution des anaphores. Certains démonstratifs s'emploient « à toutes les sauces » et, bien qu'il existe toutes sortes de théories sur la spécificité de leur mode de fonctionnement, l'analyse se heurte à la multiplicité des facteurs mis en cause. Les expressions démonstratives impliquant un déterminant démonstratif, par exemple, résistent à un examen séquentiel des différents cas de figure possibles : il faudrait tous les considérer simultanément. La théorie dit que le déterminant démonstratif, contrairement à l'article défini, n'a pas besoin de respecter le contenu nominal qui l'accompagne pour désigner sa référence. C'est vrai, à strictement parler, mais cela ne reflète pas l'ensemble des données. Même s'ils ne sont pas majoritaires, on trouve un grand nombre de cas de réitération et de synonymie. Or, ces cas seraient probablement plus faciles à traiter automatiquement que les cas de reclassification, si seulement on pouvait les reconnaître *a priori*. Notre étude n'a cependant pas permis de trouver une solution à ce problème.

Pour conclure, nous avons montré que dans plusieurs cas, il semble nécessaire de tenir compte des particularités des expressions démonstratives pour faire avancer les méthodes d'analyse automatique de textes. Toutefois, il reste encore beaucoup d'avenues à explorer dans ce domaine.

Références

APOTHÉLOZ, Denis & REICHLER-BÉGUELIN, Marie-José (1999). «Interpretations and functions of demonstrative NPs in indirect anaphora», dans *Journal of Pragmatics*, 31, 363-397.

BAUDOT, Jean (1992). *Fréquences d'utilisation des mots en français écrit contemporain.* Les Presses de l'Université de Montréal, Montréal.

BÉLANGER, Pascale (2003). *Exploration des procédés de condensation pour le résumé de texte grâce à l'application des formalismes de la Théorie Sens-Texte.* Mémoire de maîtrise, Université de Montréal, Département de linguistique et traduction, Montréal.

BLASCO-DULBECCO, Mylène (1999). *Les dislocations en français contemporain.* Éditions Champion, Paris.

BOTLEY, Simon & McENERY, Tony (2000). «Discourse anaphora: The need for synthesis», dans BOTLEY, Simon & McENERY, Anthony Mark (éd.). *Corpus-based and computational approaches to discourse anaphora*, 1-42. John Benjamins Publishing, Amsterdam/Philadelphia.

BOUDREAU, Sylvie (2004). *Résolution d'anaphores et identification des chaînes de coréférence selon le type de texte.* Mémoire de maîtrise, Université de Montréal, Département de linguistique et traduction, Montréal.

CARTER, David (1987). *Interpreting anaphors in natural language texts.* Halsted Press, New York/Toronto.

CHAROLLES, Michel (1995a). «Comment repêcher les derniers ? Analyse des expressions anaphoriques en *ce dernier*», dans *Pratiques: théorie, pratique, pédagogie*, 85, 89-112.

CORBLIN, Francis. (1987). *Indéfini, défini et démonstratif : constructions linguistiques de la référence.* Droz, Genève.

CORBLIN, Francis. (1995). *Les formes de reprise dans le discours : anaphores et chaînes de référence.* Presses universitaires de Rennes, Rennes.

CORNISH, Francis (1986). *Anaphoric relations in English and French : a discourse perspective.* Croom Helm, Beckenham/Surry Hills/Dover.

CORNISH, Francis (1999). *Anaphora, discourse, and understanding : evidence from English and French.* Clarendon Press, New York/Oxford.

De BEAUGRANDE, Robert (1997). *New foundations for a science of text and discourse : cognition, communication, and the freedom of access to knowledge and society.* Ablex Pub. Corp., Norwood, N.J.

GROSZ, Barbara J., JOSHI, Aravind K. & WEINSTEIN, Scott (1983). «Providing a unified account of definite noun phrases in discourse», dans *Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics (ACL'83)*, 44-50.

Cambridge, Massachusetts.

GROSZ, Barbara J., JOSHI, Aravind K. & WEINSTEIN, Scott (1995). «Centering: a framework for modelling the local coherence of discourse», dans *Computational Linguistics*, 21 (2) , 203-205.

HALLIDAY, Michael A.K. & HASAN, Ruqaya (1976). *Cohesion in English.* Longman, London.

HOBBS, Jerry R. (1976). *Pronoun Resolution.* Research Report 76-1. New York: Department of Computer Science, City University of New York.

HOBBS, Jerry R. (1978). «Resolving pronoun references», dans *Lingua*, 44, 339-352.

KLEIBER, George (1990). «Adjectif démonstratif et article défini en anaphore fidèle», dans *Recherches Linguistiques*, 11, 169-185.

KUCHLING, A.M. *Regular Expressions HOWTO*, [En ligne]. http://www.amk.ca/python/howto/regex/

LANGACKER, Ronald W. (1969). «On pronominalisation and the chain of command», dans REIBEL, D. & SCHANE, S. (éd.) *Modern studies in English*, 160-186. Englewood Cliffs : Prentice Hall.

MANUÉLIAN, Hélène (2003a). *Une analyse des emplois du démonstratif en corpus.* Conférence TALN 2003, Batz-sur-Mer.

MANUÉLIAN, Hélène (2003b). *Descriptions définies et démonstratives : analyses de corpus pour la génération de textes.* Thèse de doctorat, université de Nancy 2, UFR des sciences du langage, Nancy.

MEL'ČUK, Igor A. (1988). *Dependency syntax : theory and practice.* State University of New York Press, Albany.

MEL'ČUK, Igor A. (2001). *Communicative Organisation in Natural Language : the semantic-communicative structure of sentences.* John Benjamins Publishing, Amsterdam/Philadelphia.

MEL'ČUK, Igor A. (2003) «Levels of Dependency in Linguistic Description: Concepts and Problems», dans V. Agel, L. Eichinnger, H.-W. Eroms, P. Hellwig, H. J. Herringer, H. Lobin (éd.). *Dependency and Valency. An International Handbook of Contemporary Research,* vol. 1, 188-229. W. de Gruyter, Berlin/New York.

MEL'ČUK, Igor A., CLAS, André & POLGUÈRE, Alain (1995). *Introduction à la lexicologie explicative et combinatoire.* Éditions Duculot, Louvain-la-Neuve.

MERTZ, David (2006). *Text Processing in Python*, [En ligne]. http://gnosis.cx/TPiP/.

MITKOV, Ruslan (2002). *Anaphora Resolution.* Longman, London/Toronto.

MULLER, Charles (1992). *Initiation aux méthodes de la statistique linguistique.* Éditions Champion, Paris.

NEVEU, Franck (2004). *Dictionnaire des sciences du langage.* Armand-Colin, Paris.

REINHART, Tanya (1983). *Anaphora and semantic interpretation*. Croom Helm, London.

REITER, Ehud (1991). «A New Model of Lexical Choice for Nouns.», dans *Computational Intelligence*, 7, 240-251.

RIEGEL, Martin, PELLAT, Jean-Christophe & RIOUL, René (1999). *Grammaire méthodique du français*. Presses universitaires de France, Paris.

ROSCH, Eleanor, MERVIS, Carolyn B., GRAY, Wayne D., JOHNSON, David M. & BOYES-BRAEN, Penny (1976). «Basic Objects in Natural Categories.», dans *Cognitive Psychology*, 8, 382-439.

TESNIÈRE, Lucien (1976, c1959). *Éléments de syntaxe structurale*, 2e édition revue et corrigée, 3e tirage. Klincksieck, Paris.

TUTIN, Agnès & VIEGAS, Evelyne (2000). «Generating coreferential anaphoric definite NPs», dans BOTLEY, Simon & McENERY, Anthony Mark (éd.). *Corpus-based and computational approaches to discourse anaphora*, 227-248. John Benjamins Publishing, Amsterdam/Philadelphia.

TUTIN, Agnès. (1992). *Étude des anaphores grammaticales et lexicales pour la génération automatique de textes de procédures*. Thèse de doctorat, Université de Montréal, Département de linguistique et traduction, Montréal.

Annexe 1 Formes morphographiques démonstratives

Informations morphosyntaxiques et lexicales en fonction de la forme morphographique					
Forme	Catégorie grammaticale	Lexie	Forme fléchie	Genre	Nombre
c’	Pronom	CE_1	ce	N	n/a
ç’	Pronom	CE_1	ce	N	n/a
ça	Pronom	ÇA	ça	N	n/a
ce	Pronom	CE_1	ce	N	n/a
ce	Déterminant	CE_2	ce	M	S
ce -ci	Déterminant	CE -CI	ce -ci	M	S
ce -là	Déterminant	CE -LÀ	ce -là	M	S
ceci	Pronom	CECI	ceci	N	S
cela	Pronom	CELA	cela	N	S
celle	Pronom	CELLE	celle	F	S
celle-ci	Pronom	CELLE-CI	celle-ci	F	S
celle-là	Pronom	CELLE-LÀ	celle-là	F	S
celles	Pronom	CELLE	celles	F	P
celles-ci	Pronom	CELLE-CI	celles-ci	F	P
celles-là	Pronom	CELLE-LÀ	celles-là	F	P
celui	Pronom	CELUI	celui	M	S
celui-ci	Pronom	CELUI-CI	celui-ci	M	S
celui-là	Pronom	CELUI-LÀ	celui-là	M	S
ces	Déterminant	CE_2	ces	M	P
ces	Déterminant	CE_2	ces	F	P
ces -ci	Déterminant	CE -CI	ces -ci	M	P
ces -ci	Déterminant	CE -CI	ces -ci	F	P
ces -là	Déterminant	CE -LÀ	ces -là	M	P
ces -là	Déterminant	CE -LÀ	ces -là	F	P
cet	Déterminant	CE_2	ce	M	S
cet -ci	Déterminant	CE -CI	ce -ci	M	S
cet -là	Déterminant	CE -LÀ	ce -là	M	S
cette	Déterminant	CE_2	cette	F	S
cette -ci	Déterminant	CE -CI	cette -ci	F	S
cette -là	Déterminant	CE -LÀ	cette -là	F	S
ceux	Pronom	CELUI	ceux	M	P
ceux-ci	Pronom	CELUI-CI	ceux-ci	M	P
ceux-là	Pronom	CELUI-LÀ	ceux-là	M	P
ici	Adverbe	ICI	ici	n/a	n/a
là	Adverbe	LÀ	là	n/a	n/a

Tableau 5 Formes morphographiques démonstratives

Annexe 2 Exemple de résultat d'une requête XQuery

```
<liste>
<occurrence>
<fichier>La Presse_28.xml</fichier>
<expression>cette émission suivie chaque semaine par plus de 20 millions
d'Américains</expression>
<par>Richesse, beauté, amour, trahison, meurtres : <syntagme num="a1"
categorie="nominal" fonction="SG" genre="N">Desperate Housewives</syntagme> a
tous les ingrédients du soap classique. Comment expliquer alors le succès exceptionnel
de <anaphore antecedent="a1" distph="1" distcat="2" coreference="oui" lien="direct"
position="avant"><syntagme num="d1" categorie="nominal" fonction="CN"
genre="N"><dem forme="cette" categorie="déterminant" lexie="CE2" flexion="cette"
genre="F" nombre="S" type="5c">cette</dem> émission suivie chaque semaine par
plus de 20 millions d'Américains</syntagme></anaphore> ?
</par>
</occurrence>
<occurrence>
<fichier>La Presse_28.xml</fichier>
<expression>Ce premier épisode</expression>
<par>"<anaphore antecedent="p1" distph="1" distcat="0" coreference="oui"
lien="direct" position="avant"><syntagme num="d3" categorie="nominal"
fonction="SG" genre="N"><dem forme="ce" categorie="déterminant" lexie="CE2"
flexion="ce" genre="M" nombre="S" type="5c">Ce</dem> premier
épisode</syntagme></anaphore> est canon ", résume Isabelle Langlois, auteure de la
série Rumeurs, vantée pour la qualité de son écriture. "Tout est posé rapidement et
efficacement, le ton est donné. <syntagme num="d4" categorie="nominal"
genre="N"><dem forme="cette" categorie="déterminant" lexie="CE2" flexion="cette"
genre="F" nombre="S" type="5c">Cette</dem> espèce d'ironie, d'humour dans la
présentation de <syntagme num="d5" categorie="nominal" fonction="CN"
genre="N"><dem forme="ces -là" categorie="déterminant" lexie="CE -LÀ"
flexion="ces -là" genre="F" nombre="P" type="7b">ces</dem> femmes<part
dem="ces">-là</part>... bon, un peu trop belles pour être vraies, quand
même</syntagme></syntagme>", poursuit-elle. Mais si elle a mordu au départ à
l'hameçon, elle a suivi l'émission avec moins d'assiduité par la suite - conflit d'horaire :
<syntagme num="a6" categorie="nominal" genre="N">Beautés
désespérées</syntagme> contre Tout le monde en parle - mais avec l'intention de se
rattraper grâce au DVD. "<anaphore antecedent="a6" distph="1" distcat="3"
coreference="oui" lien="direct" position="avant"><syntagme num="d6"
categorie="nominal" fonction="SG" genre="N"><dem forme="c'" categorie="pronom"
lexie="CE1" flexion="ce" genre="N" type="1f">C'</dem></syntagme></anaphore>est
demeuré très, très bon, mais j'ai trouvé que, petit à petit, le ton mordant, le cynisme,
l'ironie, se sont au peu estompés et que le trait, très dur au début, a commencé à manquer
de précision."
</par>
</occurrence>
[...]
</liste>
```

Annexe 3 Exemple de fichier XSL

```
<?xml version="1.0" encoding="UTF-8"?>
<xsl:stylesheet version="1.0" xmlns:xsl="http://www.w3.org/1999/XSL/Transform">
<xsl:template match="/">
        <html>
        <body>
                <table border="1">
                        <tr>
                                <th></th>
                                <th>pronoms</th>
                                <th>déterminants</th>
                                <th>adverbes</th>
                                <th>total</th>
                        </tr>
                        <tr>
                                <td>nombre d'expressions démonstratives</td>
                                <td>
                                <xsl:value-of
select="count(//syntagme/descendant::dem[@categorie=&quot;pronom&quot;])"/>
                                </td>
                                <td>
                                <xsl:value-of
select="count(//syntagme/descendant::dem[@categorie=&quot;déterminant&quot;])"/>
                                </td>
                                <td>
                                <xsl:value-of
select="count(//syntagme/descendant::dem[@categorie=&quot;adverbe&quot;])"/>
                                </td>
                                <td><xsl:value-of
select="count(//syntagme/descendant::dem)"/></td>
                        </tr>
                        <tr>
                                <td>nombre d'expressions anaphoriques</td>
                                <td>
                                <xsl:value-of
select="count(//anaphore/descendant::dem[@categorie=&quot;pronom&quot;])"/>
                                </td>
                                <td>
                                <xsl:value-of
select="count(//anaphore/descendant::dem[@categorie=&quot;déterminant&quot;])"/>
                                </td>
                                <td>
                                <xsl:value-of
select="count(//anaphore/descendant::dem[@categorie=&quot;adverbe&quot;])"/>
                                </td>
                                <td><xsl:value-of select="count(//anaphore)"/></td>
                        </tr>
                        <tr>
                                <td>anaphoriques coréférentiels</td>
                                <td>
                                <xsl:value-of
select="count(//anaphore[@coreference=&quot;oui&quot; and
descendant::dem[@categorie=&quot;pronom&quot;]])"/>
```

```
                                        </td>
                                        <td>
                                        <xsl:value-of
select="count(//anaphore[@coreference=&quot;oui&quot; and
descendant::dem[@categorie=&quot;déterminant&quot;]])"/>
                                        </td>
                                        <td>
                                        <xsl:value-of
select="count(//anaphore[@coreference=&quot;oui&quot; and
descendant::dem[@categorie=&quot;adverbe&quot;]])"/>
                                        </td>
                                        <td><xsl:value-of
select="count(//anaphore[@coreference=&quot;oui&quot;])"/></td>
                                    </tr>
                    </table>
                    <p style="color:blue">nombre de démonstratifs de type 1a :
                            <xsl:value-of select="count(//dem[@type=&quot;1a&quot;])"/>
                    </p>
          <xsl:apply-templates/>
          </body>
          </html>
</xsl:template>
<xsl:template match="dem[@type=&quot;1a&quot;]">
          <span style="font-weight:bold; color:red">
          <xsl:value-of select="."/></span>
</xsl:template>
<xsl:template match="dem|part">
          <span style="font-weight:bold">
          <xsl:value-of select="."/></span>
          <span style="vertical-align: super"><xsl:value-of select="@dem"/></span>
</xsl:template>
<xsl:template match="syntagme|nom">
          <span style="text-decoration: underline">
          <xsl:apply-templates/></span>
          <span style="vertical-align: sub"><xsl:value-of select="@num"/></span>
</xsl:template>
<xsl:template match="anaphore">
          <xsl:apply-templates/>
          <span style="vertical-align: super"><xsl:value-of select="@antecedent"/></span>
</xsl:template>
<xsl:template match="texte/par">
          <p><xsl:apply-templates/>
                    <span style="vertical-align: sub"><xsl:value-of select="@num"/></span>
          </p>
</xsl:template>
<xsl:template match="titre">
          <pre><xsl:apply-templates/></pre>
</xsl:template>
<xsl:template match="auteur">
          <pre><xsl:value-of select="."/></pre>
</xsl:template>
<xsl:template match="nontraite">
</xsl:template>
</xsl:stylesheet>
```

Annexe 4 Fichier de validation DTD

```
<?xml version="1.0" encoding="windows-1252"?>
<!ELEMENT texte ANY>

<!ELEMENT nontraite ANY>

<!-- [illegible] -->
<!ELEMENT dem (#PCDATA)>
<!ATTLIST dem forme CDATA #REQUIRED>
<!ATTLIST dem categorie (pronom|déterminant|adverbe) #REQUIRED>
<!ATTLIST dem lexie CDATA #REQUIRED>
<!ATTLIST dem flexion CDATA #REQUIRED>
<!ATTLIST dem genre (M|F|N) #IMPLIED>
<!ATTLIST dem nombre (S|P) #IMPLIED>
<!-- [illegible] -->
<!ATTLIST dem type (1a|1b|1c|1d|1e|1f|2|3|4a|4b|5a|5b|5c|6a|6b|7a|7b|8|nc) #REQUIRED>

<!-- [illegible] -->
<!ELEMENT part (#PCDATA)>
<!ATTLIST part dem CDATA #REQUIRED>

<!-- [illegible]
[illegible] -->
<!ELEMENT syntagme ANY>
<!ATTLIST syntagme num ID #REQUIRED>
<!-- [illegible]
[illegible]
[illegible] -->
<!ATTLIST syntagme categorie
(nominal|prépositionnel|adjectival|adverbial|verbal|propositionnel) "propositionnel">
<!-- [illegible]
-->
<!ATTLIST syntagme fonction (SG|COD|COI|CC|AtS|AtO|CAg|Apt|Aph|CN|CAj|CAv|CI|CP)
#IMPLIED>
<!ATTLIST syntagme genre (M|F|N) "N">
<!ATTLIST syntagme nombre (S|P) #IMPLIED>

<!ELEMENT nom (#PCDATA)>
<!ATTLIST nom num ID #REQUIRED>
<!ATTLIST nom genre (M|F) #REQUIRED>

<!-- [illegible] -->
<!ELEMENT anaphore (syntagme|dem)>
<!ATTLIST anaphore antecedent IDREFS #REQUIRED>
<!ATTLIST anaphore coreference (oui|non) "oui">
<!ATTLIST anaphore lien (direct|indirect) "direct">
<!ATTLIST anaphore lienlexical CDATA #IMPLIED>
```

```
<!ATTLIST anaphore lienautre CDATA #IMPLIED>
<!-- distance en nombre de phrases entre l'antécédent et l'expression démonstrative -
->
<!ATTLIST anaphore distph (0|1|2|3|4etplus) #REQUIRED>
<!-- [illegible]
  [illegible] -->
<!ATTLIST anaphore distcat CDATA "beaucoup">
<!-- l'antécédent est avant ou après l'expression anaphorique -->
<!ATTLIST anaphore position (avant|après) "avant">

<!-- titre de l'article -->
<!ELEMENT titre ANY>

<!-- auteur de l'article -->
<!ELEMENT auteur ANY>

<!-- paragraphe -->
<!ELEMENT par ANY>
<!ATTLIST par num ID #IMPLIED>
```

Annexe 5 Programme d'annotation des démonstratifs

Fichier demonstratif.py

```
# -*- coding: cp1252 -*-

import re

class Forme(object):
    def __init__(self, forme_morphographique, categorie_grammaticale,
                 lexie, forme_flechie, genre = None, nombre = None):
        self.forme_morphographique = forme_morphographique
        self.categorie_grammaticale = categorie_grammaticale
        self.lexie = lexie
        self.forme_flechie = forme_flechie
        self.genre = genre
        self.nombre = nombre
        self.expression_reguliere = self.expreg()
    def __str__(self):
        return "Forme morphographique : «" \
               +" ".join(self.forme_morphographique) + "»\t" \
               +"Catégorie grammaticale : " + self.categorie_grammaticale \
               +"\t" + "Lexie : " + self.lexie + "\t" \
               +"Forme fléchie : «" + self.forme_flechie + "»\t" \
               +"Genre : " + str(self.genre) + "\t" \
               +"Nombre : " + str(self.nombre)
    def __repr__(self):
        return "Forme("+repr(self.forme_morphographique)+',' \
               +repr(self.categorie_grammaticale)+','+repr(self.lexie) \
               +','+repr(self.forme_flechie)+','+repr(self.genre)+',' \
               +repr(self.nombre)+')'
    def expreg(self):
        forme = self.forme_morphographique[0]
        expression = r'\b(?P<dem>' + forme + r')(?!</dem>)(?!")(?! -..")'
        if not forme.endswith("'"):
            expression += r'\b(?!-)'
        if len(self.forme_morphographique)==2:
            particule = self.forme_morphographique[1]
            expression += r'(?P<texte>.*?\w*?)(?P<part>' + particule \
                          + r')(?!</part>)(?!</dem>)(?!")\b'
        return re.compile(expression, re.LOCALE|re.IGNORECASE)

c = Forme(("c'",),"pronom","CE1","ce","N")
c_ced = Forme(("ç'",),"pronom","CE1","ce","N")
ca = Forme(("ça",),"pronom","ÇA", "ça", "N", "S")
ce1 = Forme(("ce",),"pronom","CE1","ce","N")
ce2 = Forme(("ce",),"déterminant","CE2","ce","M","S")
ce_ci = Forme(("ce","-ci"),"déterminant","CE -CI","ce -ci","M","S")
ce_la = Forme(("ce","-là"),"déterminant","CE -LÀ","ce -là","M","S")
ceci = Forme(("ceci",),"pronom","CECI","ceci","N","S")
cela = Forme(("cela",),"pronom","CELA","cela","N","S")
celle = Forme(("celle",),"pronom","CELLE","celle","F","S")
celle_ci = Forme(("celle-ci",),"pronom","CELLE-CI","celle-ci","F","S")
celle_la = Forme(("celle-là",),"pronom","CELLE-LÀ","celle-là","F","S")
celles = Forme(("celles",),"pronom","CELLE","celles","F","P")
celles_ci = Forme(("celles-ci",),"pronom","CELLE-CI","celles-ci","F","P")
celles_la = Forme(("celles-là",),"pronom","CELLE-LÀ","celles-là","F","P")
celui = Forme(("celui",),"pronom","CELUI","celui","M","S")
celui_ci = Forme(("celui-ci",),"pronom","CELUI-CI","celui-ci","M","S")
celui_la = Forme(("celui-là",),"pronom","CELUI-LÀ","celui-là","M","S")
```

```
cesM = Forme(("ces",),"déterminant","CE2","ces","M","P")
cesF = Forme(("ces",),"déterminant","CE2","ces","F","P")
ces_ciM = Forme(("ces","-ci"),"déterminant","CE -CI","ces -ci","M","P")
ces_ciF = Forme(("ces","-ci"),"déterminant","CE -CI","ces -ci","F","P")
ces_laM = Forme(("ces","-là"),"déterminant","CE -LÀ","ces -là","M","P")
ces_laF = Forme(("ces","-là"),"déterminant","CE -LÀ","ces -là","F","P")
cet = Forme(("cet",),"déterminant","CE2","ce","M","S")
cet_ci = Forme(("cet","-ci"),"déterminant","CE -CI","ce -ci","M","S")
cet_la = Forme(("cet","-là"),"déterminant","CE -LÀ","ce -là","M","S")
cette = Forme(("cette",),"déterminant","CE2","cette","F","S")
cette_ci = Forme(("cette","-ci"),"déterminant","CE -CI","cette -
ci","F","S")
cette_la = Forme(("cette","-là"),"déterminant","CE -LÀ","cette -
là","F","S")
ceux = Forme(("ceux",),"pronom","CELUI","ceux","M","P")
ceux_ci = Forme(("ceux-ci",),"pronom","CELUI-CI","ceux-ci","M","P")
ceux_la = Forme(("ceux-là",),"pronom","CELUI-LÀ","ceux-là","M","P")
ici = Forme(("ici",),"adverbe","ICI","ici")
la = Forme(("là",),"adverbe","LÀ","là")

demonstratifs = (c,c_ced,ca,ceci,cela,celle_ci,
                 celle_la,celle,celles_ci,celles_la,celles,celui_ci,

celui_la,celui,ce_ci,ce_la,ce2,ce1,ces_ciM,ces_ciF,ces_laM,ces_laF,cesM,ces
F,
                 cet_ci,cet_la,cet,cette_ci,cette_la,cette,ceux_ci,
                 ceux_la,ceux,ici,la)

def trouverForme(strForme, tupleFormes=demonstratifs):
    listResultat = []
    for x in tupleFormes:
        if x.forme_morphographique[0] == strForme:
            listResultat.append(x)
    return listResultat
```

Fichier etiquettage3.py

```
# -*- coding: cp1252 -*-

import sys, os, re
import demonstratifs

squeletteXML = """<?xml version="1.0" encoding="%s"?>
<?xml-stylesheet type="text/xsl" href="%s"?>
<!DOCTYPE texte SYSTEM "%s">
<texte>%s</texte>"""

ENCODING = "windows-1252"
STYLESHEET = "texte.xsl"
DTD = "texte.dtd"
TXT = r"C:\Documents and Settings\Administrateur\Mes documents\Mémoire
(maîtrise)\Corpus\Textes 1\Échantillon 1\Texte brut"
ETIQUETTES = r"C:\Documents and Settings\Administrateur\Mes
documents\Mémoire (maîtrise)\Corpus\Textes 1\Échantillon 1\Étiquettage
automatique"

def etiquetterForme(objetMatch, dem):
   remplacement = r'<dem forme="' + " ".join(dem.forme_morphographique) \
                  + r'" categorie="' + dem.categorie_grammaticale \
                  + r'" lexie="' + dem.lexie \
                  + r'" flexion="' + dem.forme_flechie
   if dem.genre != None:
```

```
      remplacement += r'" genre="' + dem.genre
   if dem.nombre != None:
      remplacement += r'" nombre="' + dem.nombre
   remplacement += r'">' + objetMatch.group('dem') + r'</dem>'
   if len(dem.forme_morphographique)>1:
      remplacement += objetMatch.group('texte') + r'<part dem="' \
                    + objetMatch.group('dem') + r'">' \
                    + objetMatch.group('part') + r'</part>'
   return remplacement

def etiquetterLigne(ligne):
    for dem in demonstratifs.demonstratifs:
      fonction = lambda objMatch, dem=dem : etiquetterForme(objMatch, dem)
      remplacement = dem.expression_reguliere.subn(fonction,ligne)
      if remplacement[1] > 0:
            log.write(str(remplacement[0]) + "\t" + str(remplacement[1]) +
"\n")
            ligne = remplacement[0]
    return ligne

def etiquetterFichier(nomFichier):
    fichier = open(nomFichier, 'r')
    texteEtiquettes = ""
    for ligne in fichier:
      if len(ligne.strip()) > 0 :
         texteEtiquettes += "<par>" + etiquetterLigne(ligne) + "</par>"
    fichier.close()
    return texteEtiquettes

log = open(ETIQUETTES+"\\log.txt", 'w')

for f in os.listdir(TXT):
   print f
   if f.endswith(".txt"):
      nomF = f[:-4] # le nom sans l'extension
      contenu = etiquetterFichier(TXT + "\\" + f)
      xml = squeletteXML % (ENCODING, STYLESHEET, DTD, contenu)
      fo = open(ETIQUETTES + "\\" + nomF + ".xml", "w")
      fo.write(xml)
      fo.close()

log.close()
sys.exit()
```

Annexe 6 Liste des articles composant le corpus

Comme il est spécifié dans le chapitre 5, les deux journaux composant notre corpus sont l'édition de *La Presse* du dimanche 25 septembre 2005 et celle du *Voir* du 22 septembre 2005. Les articles faisant entre 250 et 1000 mots sont au nombre de 78 pour *La Presse* et de 45 pour le *Voir*. Pour chaque journal, 20 articles ont été sélectionnés aléatoirement parmi ceux qui répondaient au critère de longueur pour former chacun des deux sous-corpus.

Sous-corpus La Presse

Voici les 20 numéros d'articles qui ont été sélectionnés aléatoirement pour le sous-corpus *La Presse* :

[3, 5, 10, 25, 28, 32, 37, 40, 46, 47, 49, 51, 55, 59, 65, 67, 70, 73, 74, 77]

Voici la liste des titres et auteurs de ces 20 articles :

1) [La Presse-3] PÉLOQUIN, Tristan. *Galveston s'en tire plutôt bien.*
2) [La Presse-5] COUSINEAU, Sophie. *La bulle française de Péquin.*
3) [La Presse-10] AFP. *La guerre en Irak mobilise plus les pacifistes aux É.-U. qu'en Europe.*
4) [La Presse-25] GEISER, Christian. *S'équiper pour la randonnée.*
5) [La Presse-28] SARFATI, Sonia; DUMAS, Hugo. *Secrets de Banlieue.*
6) [La Presse-32] BRUNET, Alain. *Digimart: L'"autre" distribution numérique.*
7) [La Presse-37] HARRIS, Ron; AP. *Pong, Asteroïds, voilà les jeux rétro!*
8) [La Presse-40] APOSTOLSKA, Aline (collaboration spéciale). *Festival mondial des arts pour la jeunesse.*
9) [La Presse-46] HOMEL, David (collaboration spéciale). *Train d'enfer.*
10) [La Presse-47] MARTEL, Réginald. *Bossalo, un Barcelo décevant.*
11) [La Presse-49] LAFERRIÈRE, Dany. *L'autre grand tabou.*
12) [La Presse-51] GRAMMOND, Stéphanie. *Bell et bien compliqué.*
13) [La Presse-55] TISON, Marc. *Retraite simultanée... et équitable.*
14) [La Presse-59] GIRARD, Michel. *Hypothéquer la maison pour investir?*

15) [La Presse-65] LADOUCEUR, Pierre. *Benoit est heureux d'avoir mis "un pied" dans la porte".*

16) [La Presse-67] LABBÉ, Richard. *Encore les arbitres...*

17) [La Presse-70] ALLARD, Sophie. *Des partisans extrêmes de l'Impact.*

18) [La Presse-73] AFP; PC. *Alonso touche du doigt la couronne mondiale.*

19) [La Presse-74] AP. *Tout sera décidé aujourd'hui.*

20) [La Presse-77] LABBÉ, Richard. *NHL 06 pas mal, mais...*

Sous-corpus Voir

Voici les 20 numéros d'articles qui ont été sélectionnés aléatoirement pour le sous-corpus *Voir* :

[3, 5, 9, 10, 12, 14, 15, 20, 21, 22, 23, 27, 29, 30, 31, 34, 36, 37, 40, 43]

Voici la liste des titres et auteurs de ces 20 articles :

1) [Voir-3] HOCHEREAU, Alain. *L'Ancienne-Aluminerie-de-Shawinigan.*
2) [Voir-5] GAGNÉ, Paul. *Kalalu.*
3) [Voir-9] DJOGO, Martina. *Betty's Bazaar déménage.*
4) [Voir-10] FIORENZA, Évelyne. *Tisanes pour relaxer.*
5) [Voir-12] MAVRIKAKIS, Nicolas. *Quinze Césanne.*
6) [Voir-14] BÉRUBÉ, Jade. *Rideau Vert.*
7) [Voir-15] SAINT-PIERRE, Christian. *Evelyne Rompré.*
8) [Voir-20] ROBILLARD LAVEAUX, Olivier. *Les Sainte-Catherines chez Fat Wreck Chords.*
9) [Voir-21] HÉBERT, Francis. *Chloé Sainte-Marie.*
10) [Voir-22] POITRAS, Marie-Hélène; DEFOY, Michel; MARTEL, Stéphane; ROBILLARD LAVEAUX, Olivier; OUELLET, Patrick. *Pop Montréal.*
11) [Voir-23] WITHENSHAW, Anne-Marie. *Noir Désir juge concevable de se reformer.*
12) [Voir-27] HÉLÈNE POITRAS, Marie. *Je reviendrai à Montréal.*
13) [Voir-29] BEAUCAGE, Réjean. *Marie-Hélène Breault.*
14) [Voir-30] MALAVOY-RACINE, Tristan. *D'Amélie Nothomb.*
15) [Voir-31] PAQUIN, Éric. *Gil Courtemanche.*

16) [Voir-34] MARCY, Normand. *Estelle Clareton.*

17) [Voir-36] MANDOLINI, Carlo. *Save and Burn.*

18) [Voir-37] DUMAIS, Manon. *J'aime les filles.*

19) [Voir-40] DUMAIS, Manon; DEFOY, Michel; LAFOREST, Kevin. *FIFM.*

20) [Voir-43] MAVRIKAKIS, Nicolas. *Diane Borsato.*

Annexe 7 Extrait du fichier "La Presse_28.xml"

Cet extrait du fichier "La Presse_28.xml" est un exemple de ce à quoi ressemblent les fichiers de notre corpus, après la phase d'annotation manuelle.

```
<?xml version="1.0" encoding="windows-1252"?>
<?xml-stylesheet type="text/xsl" href="texte.xsl"?>
<!DOCTYPE texte SYSTEM "texte.dtd">
<texte><nontraite><par>La Presse
</par><par>Arts et spectacles, dimanche 25 septembre 2005, p. ARTS
SPECTACLES2
</par></nontraite><titre><par>Beautés désespérées
</par><par>Secrets de banlieue
</par></titre><auteur><par>Sarfati, Sonia; Dumas, Hugo
</par></auteur><par>Richesse, beauté, amour, trahison, meurtres : <syntagme
num="a1" categorie="nominal" fonction="SG">Desperate Housewives</syntagme>
a tous les ingrédients du soap classique. Comment expliquer alors le succès
exceptionnel de <anaphore antecedent="a1" distph="1" distcat="2"><syntagme
num="d1" categorie="nominal" fonction="CN"><dem forme="cette"
categorie="déterminant" lexie="CE2" flexion="cette" genre="F" nombre="S"
type="5c">cette</dem> émission suivie chaque semaine par plus de 20 millions
d'Américains</syntagme></anaphore> ?
</par><par num="p1">Mary Alice qui raconte une journée <syntagme num="d2"
categorie="nominal" fonction="Apt">tout <!-- quoi inclure dans ce syntagme ? --
><dem forme="ce" categorie="pronom" lexie="CE1" flexion="ce" genre="N"
type="1a">ce</dem> qu'il y a de plus banale</syntagme> culminant par un
événement qui l'est moins, son suicide. Après les funérailles, ses quatre amies
passent par la maison de la défunte. Susan et son macaroni raté, grâce auquel elle
parvient à... faire du plat au nouveau voisin. Lynette et ses trois petits démons, qu'elle
ira repêcher dans la piscine. Bree et sa famille tirée à quatre épingles, avec son panier
de muffins faits maison. Gabrielle et son riche mari, qui a pour mission de faire
connaître le prix du collier qu'elle étrenne. Le tout se terminant par la découverte
d'une lettre qui semble en être une de chantage.
</par><par>"<anaphore antecedent="p1" distph="1" distcat="0"><syntagme
num="d3" categorie="nominal" fonction="SG"><dem forme="ce"
categorie="déterminant" lexie="CE2" flexion="ce" genre="M" nombre="S"
type="5c">Ce</dem> premier épisode</syntagme></anaphore> est canon ", résume
Isabelle Langlois, auteure de la série Rumeurs, vantée pour la qualité de son écriture.
"Tout est posé rapidement et efficacement, le ton est donné. <syntagme num="d4"
categorie="nominal"><dem forme="cette" categorie="déterminant" lexie="CE2"
flexion="cette" genre="F" nombre="S" type="5c">Cette</dem> espèce d'ironie,
d'humour dans la présentation de <syntagme num="d5" categorie="nominal"
fonction="CN"><dem forme="ces -là" categorie="déterminant" lexie="CE -LÀ"
flexion="ces -là" genre="F" nombre="P" type="7b">ces</dem> femmes<part
dem="ces">-là</part>... bon, un peu trop belles pour être vraies, quand
```

même</syntagme></syntagme>", poursuit-elle. Mais si elle a mordu au départ à l'hameçon, elle a suivi l'émission avec moins d'assiduité par la suite - conflit d'horaire : <syntagme num="a6" categorie="nominal">Beautés désespérées</syntagme> contre Tout le monde en parle - mais avec l'intention de se rattraper grâce au DVD. "<anaphore antecedent="a6" distph="1" distcat="3"><syntagme num="d6" categorie="nominal" fonction="SG"><dem forme="c'" categorie="pronom" lexie="CE1" flexion="ce" genre="N" type="1f">C'</dem></syntagme></anaphore>est demeuré très, très bon, mais j'ai trouvé que, petit à petit, le ton mordant, le cynisme, l'ironie, se sont au peu estompés et que le trait, très dur au début, a commencé à manquer de précision."
</par><par>Peut-être, avance-t-elle, parce que <syntagme num="a7" categorie="propositionnel">les scénarios de Desperate Housewives sont écrits par un pool d'auteurs</syntagme>, comme <anaphore antecedent="a7" distph="0" distcat="0"><syntagme num="d7" categorie="nominal" fonction="SG"><dem forme="cela" categorie="pronom" lexie="CELA" flexion="cela" genre="N" nombre="S" type="2">cela</dem></syntagme></anaphore> se fait dans la quasi-totalité des séries américaines. "L'auteur principal (Marc Cherry), fait la bible, le premier et le dernier épisode et en signe quelques autres <syntagme num="d8et9" categorie="adverbial"><dem forme="ici" categorie="adverbe" lexie="ICI" flexion="ici" type="8">ici</dem> et <dem forme="là" categorie="adverbe" lexie="LÀ" flexion="là" type="8">là</dem></syntagme>." Vérification faite, la première saison de la série est en effet le fruit du travail de 17 scénaristes. Marc Cherry a signé les trois premiers et participé à l'écriture des deux suivants, a collaboré à quelques autres et est revenu pour le dernier. Tout en supervisant le travail de ses troupes, naturellement.
</par><par>Le résultat est "bien rythmé, porté par des personnages clairs, véhiculant un ton nouveau", dit-elle pour expliquer le succès de la série. "L'élément de suspense est, lui aussi, accrocheur. Et puis, il y a <!-- préciser le lien lexical: le référent serait quelque chose comme : «la voix de la suicidée» --><anaphore antecedent="a10" coreference="non" lien="indirect" distph="1" distcat="0" position="après" lienlexical="A0(histoire)=narrative;Sinstr(raconter)=voix"><syntagme num="d10" categorie="nominal" fonction="COD"><dem forme="cette" categorie="déterminant" lexie="CE2" flexion="cette" genre="F" nombre="S" type="5c">cette</dem> voix narrative</syntagme></anaphore> : <syntagme num="a10" categorie="propositionnel">faire raconter l'histoire par la suicidée</syntagme>, il fallait y penser." Elle a particulièrement aimé "<syntagme num="d11" categorie="nominal" fonction="COD"><dem forme="ce" categorie="déterminant" lexie="CE2" flexion="ce" genre="M" nombre="S" type="5c">ce</dem> regard posé, dès le départ, sur nos faiblesses et nos manques</syntagme>".
</par>
[...]
</texte>

Annexe 8 Dénombrement et hypothèses

Le tableau suivant récapitule le nombre d'expressions démonstratives trouvées dans le corpus pour chaque type d'expression démonstrative. On y trouve aussi le nombre d'expressions démonstratives vérifiant l'hypothèse se rapportant à chaque type d'expression démonstrative, tel que stipulé dans le chapitre Hypothèses.

Type d'expression	Quantité		Hypothèse vérifiée	
	Nombre	Proportion	Nombre	Proportion
1A	40	7 %	31	78 %
1B	33	6 %	28	85 %
1C	2	0,3 %	2	0 %
1D	0	0 %	0	0 %
1E	3	0,5 %	3	100 %
1F	137	24 %	12	9 %
Sous-total 1	215	37 %	76	35 %
2	81	14 %	45	56 %
3	11	1,9 %	6	55 %
4A	17	3 %	5	29 %
4B	22	4 %	17	77 %
Sous-total 4	39	7 %	22	56 %
5A	3	0,5 %	3	100 %
5B	24	4 %	16	67 %
5C	155	27 %	-	-
Sous-total 5	182	32 %	19	10 %
6A	3	0,5 %	0	0 %
6B	2	0,3 %	0	0 %
Sous-total 6	5	0,9 %	0	0 %
7A	3	0,5 %	0	0 %
7B	4	0,7 %	0	0 %
Sous-total 7	7	1,2 %	0	0 %
8	37	6 %	-	-
Total	577	100%	168	29%

Tableau 6 Dénombrement et vérification des hypothèses

Annexe 9 Données de Baudot (1992)

Le tableau ci-dessous donne la liste des entrées du répertoire de Baudot (1992) qui ont servi à établir le graphique de la figure 5 du chapitre 6. Les entrées du répertoire sont rapportées avec exactement la même notation que celle utilisée par Baudot (1992).[50]

Section de la figure 5		Lexie(s)	Entrée(s) de Baudot (1992)	Page dans Baudot (1992)
Adverbe		ICI LÀ	499=**ici**=adv 994=**là**=adv total : 1493	page 55
Déterminant		CE_2 CE -CI CE -LÀ	8810=**ce**=adj dém*	page 54
Pronom	Types 1A à 1F	CE_1	6686=**ce**=pr dém* *(types 1A à 1D et type 1F)* 221=**c'est-à-dire**=loc conj* *(type 1E)* total : 6686	page 54 page 55
	Type 2	CECI CELA ÇA	120=**ceci**=pr dém 878=**cela**=pr dém 527=**ça**=pr dém total : 1525	page 54 page 50
	Type 3	CELUI-CI CELUI-LÀ CELLE-CI CELLE-LÀ	366=**celui-ci**=pr dém* 72=**celui-là**=pr dém* total : 438	page 54
	Types 4A et 4B	CELUI CELLE	1754=**celui**=pr dém*	page 54
	Total pronoms		10624	
Total			20927	

Tableau 7 Données de Baudot (1992)

Veuillez noter que puisque Baudot (1992) ne fournissait pas le contexte de chaque expression comptabilisée dans son répertoire, nous avons dû nous contenter de regrouper les données de la manière dont ils apparaissent dans la figure 5. Ces regroupements ne demandent pas de connaître le contexte dans lequel apparaissent

[50] Les abréviations suivantes sont utilisées par Baudot (1992) : adv : adverbe, adj dém : adjectif démonstratif, pr dém : pronom démonstratif, loc conj : locution de conjonction. L'astérisque marque des entrées qui sont détaillées dans une autre section de Baudot (1992), tel qu'expliqué plus bas.

les démonstratifs en question. Il suffit de faire correspondre les entrées de Baudot (1992) avec les lexies correspondantes, telles que ces dernières sont définies à l'annexe 1 de ce mémoire.

Les entrées apparaissant dans le tableau ci-dessus sont tirées de la section «Liste alphabétique» de Baudot (1992). Baudot (1992) comprend également une section «Lexique», dans laquelle on trouve le détail de certaines entrées de la section «Liste alphabétique». Ces entrées sont marquées par un astérisque (*) dans le tableau 7, tout comme dans la section «Liste alphabétique» de Baudot (1992). Le tableau 8 donne le détail de ces entrées, tel qu'il figure dans la section «Lexique» de Baudot (1992:396).

Entrée de la section «Liste alphabétique»	Entrées correspondantes dans la section «Lexique»
8810=**ce**=adj dém*	2778=**ce**=adj dém 694=**cet**=adj dém 3089=**cette**=adj dém 2249=**ces**=adj dém
6686=**ce**=pr dém*	3175=**ce**=pr dém 3511=**c'**=pr dém
1754=**celui**=pr dém*	548=**celui**=pr dém 508=**ceux**=pr dém 516=**celle**=pr dém 182=**celles**=pr dém
366=**celui-ci**=pr dém*	150=**celui-ci**=pr dém 61=**ceux-ci**=pr dém 120=**celle-ci**=pr dém 35=**celles-ci**=pr dém
72=**celui-là**=pr dém*	25=**celui-là**=pr dém 25=**ceux-là**=pr dém 20=**celle-là**=pr dém 2=**celles-là**=pr dém
221=**c'est-à-dire**=loc conj*	219=**c'est-à-dire**=loc conj 2=**c.-à-d.**=abr

Tableau 8 Détail des entrées marquées d'un astérisque dans Baudot (1992)