Université de Montréal

# Résolution d'anaphores et identification des chaînes de coréférence selon le type de texte

par :
*Sylvie Boudreau*

Unité académique 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 M.A.
en linguistique

Octobre 2004

Université de Montréal
Faculté des études supérieures

Ce mémoire intitulé :

# Résolution d'anaphores et identification des chaînes de coréférence selon le type de texte

présenté par :
*Sylvie Boudreau*

a été évalué par un jury composé des personnes suivantes :

*Richard Patry*
*président-rapporteur*

*Richard Kittredge*
*directeur de recherche*

*Lyne Da Sylva*
*membre du jury*

Mémoire accepté le 26 octobre 2004

# Résumé et mots-clés

## Résumé

Dans ce mémoire de maîtrise, je me suis intéressée au problème de la partition des expressions référentielles d'un texte en chaînes coréférentielles disjointes. En effet, l'établissement de ces chaînes coréférentielles est souvent une étape nécessaire dans bien des domaines du traitement automatique de la langue (*TAL*). Dans cette étude, j'ai porté une attention particulière à l'identification automatique des chaînes chapeautées par des noms propres grâce à un algorithme simple (c'est-à-dire ne nécessitant pas de parsage syntaxique complet) et partiellement adaptable au type de texte. Pour élaborer cet algorithme, j'ai effectué la comparaison de trois types de textes appartenant à des domaines assez différents et j'ai utilisé le codage *XML* pour représenter les données. Aussi, dans la réalisation de ce travail, j'ai dû m'attaquer à quelques sous-problèmes comme l'identification automatique des syntagmes nominaux (trouver les bonnes bornes gauches et droites) et la sélection des syntagmes qui semblent être les plus importants.

## Mots-clés

résolution d'anaphores, chaîne de coréférence, nom propre, traitement automatique de texte, linguistique du texte, langage de balisage *XML*

# Abstract and Keywords


## Abstract

In this master's thesis, I have considered the problem of partitioning a text's referential expressions into mutually-exclusive coreferential chains. This partitioning is an important step in many applications of natural language processing (*NLP*). In the present work, I have emphasized the automatic identification of chains headed by a proper noun, using a simple algorithm (i.e. which doesn't require a complete syntactic parse) partially adapted to the type of text and domain. To design this algorithm, I have compared three types of texts from vastly differing domains and I have used the *XML* markup language to represent the relevant metadata. I have also had to consider a few subproblems, such as the automatic identification of noun phrases (i.e. finding the appropriate left and right boundaries) and the identification of the "important" phrases of a text.



## Keywords

anaphora resolution, coreference chains, proper name, natural language processing, discourse analysis, *XML*

# Table des matières

**Identification du jury** i

**Résumé et mots-clés** ii

**Abstract and Keywords** iii

**Liste des tableaux** xii

**Conventions d'écriture** xv

**Remerciements** xviii

**Le corps de l'ouvrage** 1

**1 Introduction** 2

1.1 Problématique . . . 3

1.2 Objectifs . . . 6

1.2.1 Facteurs influençant la coréférence . . . 6

1.2.2 Comportement coréférentiel selon le type de texte . . . 7

1.2.3 Schématisation de la coréférence . . . 8

1.2.4 Adaptation au français . . . 8

1.2.5 Minimalisation des ressources . . . 9

1.3 Structure du mémoire . . . 9

1.4 Applications possibles . . . 10

1.4.1 Interrogation et indexation automatique . . . 10

1.4.2 Extraction d'informations et résumé automatique . . . 10

1.4.3 Synthèse de textes . . . . . 11
1.4.4 Traduction automatique . . . . . 11

**2 Notions théoriques** **13**

2.1 Syntagmes référentiels . . . . . 14
2.1.1 Têtes$_r$ . . . . . 14
2.1.1.1 Nom commun . . . . . 14
2.1.1.2 Nom propre . . . . . 15
2.1.1.3 Pronom et ellipse . . . . . 16
2.1.2 Déterminants . . . . . 17
2.1.2.1 Indéfini . . . . . 18
2.1.2.2 Défini . . . . . 18
2.1.2.3 Démonstratif . . . . . 19
2.2 Chaîne coréférentielle . . . . . 19
2.2.1 Relations . . . . . 20
2.2.1.1 Coréférence . . . . . 20
2.2.1.2 Anaphore . . . . . 21
2.2.2 Positions . . . . . 23
2.3 Conclusion . . . . . 25

**3 Travaux précédents** **26**

3.1 Méthodes *knowledge-poor* . . . . . 26
3.2 Historique . . . . . 28
3.2.1 Années 80 . . . . . 29
3.2.2 Début 90 . . . . . 29
3.2.3 Stratégies *knowledge-poor* . . . . . 30
3.2.4 *cogNIAC* . . . . . 31
3.2.5 Chaînes importantes d'un texte . . . . . 32
3.2.6 Réduction du prétraitement . . . . . 32
3.3 Conclusion . . . . . 33

**4 Outils *XML*** 35

4.1 *XML* . . . 36

4.1.1 Classification des mots . . . 38

4.1.2 Identification des syntagmes référentiels . . . 39

4.1.3 Établissement des liens anaphoriques et coréférentiels . . . 39

4.1.4 Autres possibilités . . . 41

4.1.5 Schématisation finale . . . 42

4.2 Autres outils *XML* . . . 43

4.2.1 *XHTML* . . . 43

4.2.2 *CSS* . . . 44

4.2.3 *XPath* . . . 45

4.2.4 *XSLT* . . . 46

4.3 Conclusion . . . 49

**5 Méthodologie** 50

5.1 Sélection du corpus . . . 50

5.1.1 Choix des domaines et des textes . . . 51

5.1.2 Prétraitement . . . 52

5.1.3 Corpus d'entraînement et de test . . . 52

5.1.4 Balisage . . . 52

5.2 Élaboration de l'algorithme . . . 53

5.2.1 Niveau lexical . . . 55

5.2.2 Niveau syntaxique I . . . 55

5.2.3 Niveau syntaxique II . . . 57

5.2.4 Niveau textuel . . . 57

5.3 Mesures d'efficacité de l'algorithme . . . 58

5.3.1 Précision et rappel . . . 58

5.3.2 *B-CUBED* . . . 59

5.4 Conclusion . . . 60

**6 Présentation et analyse des résultats** 62

6.1 Corpus balisé manuellement . . . . . 62

6.1.1 Présentation des textes balisés . . . . . 63

6.1.1.1 Textes en chaînes coréférentielles . . . . . 63

6.1.1.2 Chaînes coréférentielles en contexte . . . . . 63

6.1.2 Fréquence des noms communs . . . . . 64

6.1.3 Comparaison des types de syntagmes référentiels . . . . . 64

6.2 Règles *knowledge-poor* . . . . . 65

6.2.1 Traitements au niveau lexical . . . . . 65

6.2.1.1 Mots grammaticaux . . . . . 66

6.2.1.2 Noms communs associés au domaine . . . . . 67

6.2.1.3 Noms communs référant à la situation du discours . . . . . 67

6.2.1.4 Verbes impersonnels . . . . . 68

6.2.1.5 Noms propres . . . . . 68

6.2.2 Traitements au niveau syntaxique I . . . . . 68

6.2.2.1 Syntagmes dont la $tête_r$ est un pronom . . . . . 69

6.2.2.2 Syntagmes dont la $tête_r$ est un nom propre . . . . . 69

6.2.2.3 Syntagmes dont la $tête_r$ est un nom commun associé au domaine 70

6.2.2.4 Syntagmes comportant un déterminant démonstratif . . . . . 70

6.2.2.5 Syntagmes référentiels enchâssés . . . . . 71

6.2.3 Traitements au niveau syntaxique II . . . . . 71

6.2.3.1 Complément du nom . . . . . 72

6.2.3.2 Complément d'objet indirect . . . . . 72

6.2.3.3 Apposition . . . . . 72

6.2.3.4 Sujet . . . . . 72

6.2.4 Traitements au niveau textuel . . . . . 73

6.2.4.1 Coréférence . . . . . 73

6.2.4.2 Anaphore de reprise . . . . . 74

6.2.4.3 Anaphore lexicale . . . . . 74

6.2.4.4 Anaphore pronominale . . . . . 75

6.3 Mesures d'efficacité . . . . . 75

6.4 Variations selon le type de texte . . . . . 77
6.4.1 Vocabulaire . . . . . 78
6.4.2 Éléments référentiels . . . . . 79
6.4.3 Constructions particulières . . . . . 79
6.4.4 Chaînes coréférentielles . . . . . 80
6.5 Conclusion . . . . . 82

**7 Discussion** 84
7.1 Problèmes rencontrés . . . . . 84
7.2 Points forts de la démarche . . . . . 88
7.3 Points faibles de la démarche . . . . . 89
7.4 Améliorations possibles . . . . . 89
7.4.1 Construction du vocabulaire . . . . . 89
7.4.2 Longueur des chaînes existantes . . . . . 90
7.4.3 Mise en forme du texte . . . . . 90
7.4.4 Nombre de textes et de domaines traités . . . . . 91
7.4.5 Constructions `det nom_commun npr` . . . . . 91
7.5 Pour aller plus loin... . . . . 92
7.5.1 Traiter d'autres cas simples . . . . . 92
7.5.2 Traiter d'autres types de cas coréférentiels . . . . . 93
7.5.3 Utiliser les différences des moyens coréférentiels selon le type de texte . . 93
7.6 Conclusion . . . . . 93

**8 Conclusion** 95
8.1 Facteurs influençant la coréférence . . . . . 95
8.2 Comportements coréférentiels selon le type de texte . . . . . 97
8.3 Schématisation de la coréférence . . . . . 98
8.4 Adaptation au français . . . . . 98
8.5 Minimalisation des ressources . . . . . 98

**Annexes** ii

**A Glossaire** ii

**B Vocabulaire** iv

**C Exemples de textes** xv

C.1 Fusion de comapgnies . . . . . xv

C.2 Critique de film . . . . . xviii

C.3 *HOWTO Linux* . . . . . xix

**D Exemple d'un texte balisé** xx

**E Tableaux des coréférences** xxiii

**F Coréférences en contexte :** xxxi

**G Fréquence des noms communs** xlviii

G.1 Fusions de compagnies . . . . . xlviii

G.2 Critiques de films . . . . . l

G.3 *HOWTO Linux* . . . . . lii

**H Fréquence des syntagmes nominaux dans les chaînes coréférentielles** liv

H.1 Fusions de compagnies . . . . . lv

H.1.1 S.N. appartenant à une chaîne coréférentielle (de longueur > 1) . . . . . lv

H.1.2 Têtes de chaînes coréférentielles . . . . . lvi

H.1.3 Syntagmes anaphoriques . . . . . lvii

H.1.4 Syntagmes coréférentiels . . . . . lviii

H.2 Critiques de films . . . . . lix

H.2.1 S.N. appartenant à une chaîne coréférentielle (de longueur > 1) . . . . . lix

H.2.2 Têtes de chaînes coréférentielles . . . . . lx

H.2.3 Syntagmes anaphoriques . . . . . lxi

H.2.4 Syntagmes coréférentiels . . . . . lxii

H.3 *HOWTO Linux* . . . . . lxiii

H.3.1 S.N. appartenant à une chaîne coréférentielle (de longueur > 1) . . . . . lxiii

H.3.2 Chaînes coréférentielles . . . . . lxiv

H.3.3 Syntagmes anaphoriques . . . . . lxv

H.3.4 Syntagmes coréférentiels . . . . . lxvi

H.4 Tous les domaines . . . . . lxvii
H.4.1 S.N. appartenant à une chaîne coréférentielle (de longueur > 1) . . . . . lxvii
H.4.2 Chaînes coréférentielles . . . . . lxviii
H.4.3 Syntagmes anaphoriques . . . . . lxix
H.4.4 Syntagmes coréférentiels . . . . . lxx

**Bibliographie** **lxxiii**

**Sources des exemples** **lxxv**

# Liste des tableaux


3.1 Poids des antécédents candidats (S. Lappin et H. Leass) . . . . . 30
3.2 Critères d'antécédence (R. Mitkov) . . . . . 31
3.3 Critères d'antécédence (cogNIAC) . . . . . 32

5.1 Composition – Corpus d'entraînement . . . . . 52
5.2 Composition – Corpus de test . . . . . 53

6.1 Mesures d'efficacité – Résolution des pronoms (corpus d'entraînement) . . . . . 76
6.2 Mesures d'efficacité – Résolution des pronoms (corpus de test) . . . . 76

H.1 S.N. appartenant à une chaîne coréférentielle – Fusions de compagnies . . . . . lv
H.2 Têtes de chaînes coréférentielles – Fusions de compagnies . . . . . lvi
H.3 Syntagmes anaphoriques – Fusions de compagnies . . . . . lvii
H.4 Syntagmes coréférentiels – Fusions de compagnies . . . . . lviii
H.5 S.N. appartenant à une chaîne coréférentielle – Critiques de films . lix
H.6 Têtes de chaînes coréférentielles – Critiques de films . . . . . lx
H.7 Syntagmes anaphoriques – Critiques de films . . . . . lxi
H.8 Syntagmes coréférentiels – Critiques de films . . . . . lxii
H.9 S.N. appartenant à une chaîne coréférentielle – HOWTO Linux . . . . lxiii
H.10 Chaînes coréférentielles – HOWTO Linux . . . . . lxiv
H.11 Syntagmes anaphoriques – HOWTO Linux . . . . . lxv
H.12 Syntagmes coréférentiels – HOWTO Linux . . . . . lxvi
H.13 S.N. appartenant à une chaîne coréférentielle – Tous les domaines . lxvii

H.14 CHAÎNES CORÉFÉRENTIELLES – TOUS LES DOMAINES . . . . . lxviii
H.15 SYNTAGMES ANAPHORIQUES – TOUS LES DOMAINES . . . . . lxix
H.16 SYNTAGMES CORÉFÉRENTIELS – TOUS LES DOMAINES . . . . . lxx

# Sigles

*ACL*: Association for Computational Linguistics
*CSS*: Cascading Style Sheet
*DTD*: Document Type Definition
*MUC*: Message Understanding Conferences
*NLP*: Natural Language Processing
*HTML*: HyperText Markup Language
*TAL*: Traitement Automatique de la Langue
*W3C*: World Wide Web Consortium
*XHTML*: Extensible HyperText Markup Language
*XML*: Extensible Markup Language
*XPATH*: *XML* Path Language
*XSLT*: Extensible Stylesheet Language Transformation

# Abréviations[1]

*adj.*: adjectif
*adv.*: adverbe
*anaph.*: anaphore
*cataph.*: cataphore
*C.C.*: complément circonstanciel
*C.nom*: complément du nom
*C.O.D.*: complément d'objet direct
*C.O.I.*: complément d'objet indirect
*coréf.*: coréférence
*dét.*: déterminant
*npr.*: nom propre
*prép.*: préposition
*pro.*: pronom
*S.N.*: syntagme nominal

[1] Certaines de ces abréviations seront parfois écrites en minuscules, sans les points ni les accents. Ces simplifications ont pour but d'alléger la notation (par exemple, à l'intérieur des balises *XML*).

# Conventions d'écriture

La langue est un bel outil pour exprimer les faits et les pensées. Cependant, dans ce mémoire de linguistique, puisque nous désirons justement parler de cette langue, nous devons poser certaines règles métalinguistiques. Voici les conventions d'écritures que nous utiliserons :

### Sigles

Les sigles seront en *MAJUSCULE ITALIQUE*.

> *Pour étudier la langue à ces niveaux structurels, il nous semble que l'utilisation des outils XML soit une meilleure approche.* (1)

### Ellipses

Nous utiliserons le symbole ∅ pour marquer la position de l'ellipse.

> *J'ai trois crayons. Je préfère le ∅ bleu.* (2)

### Têtes$_r$ et Têtes$_c$

Nous utiliserons la notion de tête pour signifier *tête de syntagme référentiel* ou *tête de chaîne coréférentielle*. Pour distinguer ces deux cas, nous noterons ces notions *tête*$_r$ et *tête*$_c$ respectivement.

### Signifiants

Les mots en *italiques* seront utilisés pour marquer les signifiants. Si le texte est déjà en italique, ils seront aussi en gras.

> *Dans les phrases suivantes, l'antécédent de **Il** (au masculin) est **Noah** car il est un homme.* (3)

### Sens

Les sens seront encadrés par les parenthèses en exposant.

*Pour nous, le sens le plus important du nom propre est son sens de dénomination, c'est-à-dire* (*être appelé* $x$). (4)

### MOTS NOUVEAUX OU IMPORTANTS

Les mots nouveaux ainsi que les mots importants seront identifiés en caractères **gras**. Parfois, nous utiliserons aussi le soulignement s'il y a plusieurs niveaux d'importance.

***la fille de Coline Serreau*** (5)

### BOUTS DE CODE INFORMATIQUE

Les bouts de code informatique ou de *XML* seront écrits en `police de machine à ecrire`. Parfois, si le code comprend un contenu linguistique, ce dernier sera en *italique*.

```
<vocabulaire>
   <déterminants>
      <det type="défini" genre="m">le</det>
      <det type="défini">l'</det>
      <det type="défini" genre="f">la</det>
      <det type="défini" nombre="plur">les</det>
      <det type="démonstratif" genre="m">ce</det>
      <det type="démonstratif">c'</det>

                              ...

   </déterminants>
   <pronoms>
      <pro type="personnel" personne="1">je</pro>
      <pro type="personnel" personne="1">j'</pro>
      <pro type="personnel" personne="2">tu</pro>
      <pro type="personnel" genPer="m" personne="3">il</pro>
      <pro type="personnel" genPer="f" personne="3">elle</pro>

                              ...

   </pronoms>
</vocabulaire>
```
(6)

### LIENS ANAPHORIQUES ET CORÉFÉRENTIELS

Les syntagmes référentiels seront étiquetés à l'aide d'indices[2]. Chaque syntagme référentiel recevra un numéro d'identification. Par la suite, les liens coréférentiels et anaphoriques seront

[2]Dans le mémoire mais pas dans les textes balisés en *XML*.

notés avec les indices *coref=id* et *anaph=id* où *id* est l'expression coréférentielle non anaphorique la plus récente.

> ***Une femme***$_1$ *demande à* ***Donna***$_2$ *de* ***lui***$_{3,anaph=1}$ *créer une robe spécialement pour une séance de spiritisme.* ***Celle-ci***$_{4,anaph=1}$ *voudrait entrer en communication avec* ***son défunt mari***$_{5,anaph=1(possession)}$ *afin qu'****il***$_{6,anaph=5}$ *l'*$_{7,anaph=4}$ *aide à retrouver son bracelet de diamants.* (7)

*Je remercie mon directeur de recherche, Richard Kittredge, pour ses suggestions judicieuses et ses commentaires toujours appréciés.*

*Je remercie aussi Martin pour son soutien moral.*

# Le corps de l'ouvrage

# Chapitre 1

# Introduction

*On dit que* $\lim_{(x,y)\to(a,b)} f(x,y) = L$ *si et seulement si pour tout nombre réel positif* $\varepsilon$ *il existe un nombre réel positif* $\delta$*, qui dépend de* $\varepsilon$ *[c.-à-d.* $\delta(\varepsilon)$*], tel que* $0 < \sqrt{(x-a)^2 + (y-b)^2} < \delta$ *implique que* $|f(x,y) - L| < \varepsilon$. (1)

Ce texte est la définition de la limite d'une fonction à deux variables[1]. Les lecteurs les plus perspicaces auront déjà compris que nous ne nous intéresserons pas réellement aux fonctions à trois dimensions, ni même à quoi que ce soit de mathématique car le titre de ce mémoire ne suggère rien de tel.

Cette définition peut toutefois cadrer dans notre introduction par ses aspects linguistiques, surtout par rapport à l'utilisation des **expressions référentielles** (les variables) dans le texte. Puisque les mathématiciens cherchent à tout prix à éviter l'ambiguïté, ils utilisent un mécanisme formel pour introduire un nouveau **référent**, le réutiliser au besoin et décrire les liens de dépendance face à ce référent.

Prenons par exemple un actant de (1) (la variable $\varepsilon$). La première utilisation de cette variable (l'expression *pour tout nombre réel positif* $\varepsilon$) sert à fixer le référent pour le reste de la définition. Ici, la variable a été nommée $\varepsilon$ mais il aurait pu en être autrement, car $\varepsilon$ n'est pas rattachée à un objet réel existant à l'extérieur de la définition. La seule contrainte étant qu'elle préserve «le même nom» par la suite. Ainsi,

*On dit que* $\lim_{(\clubsuit,\heartsuit)\to(\star,\circ)} f(\clubsuit,\heartsuit) = \text{m}$ *si et seulement si pour tout nombre réel positif* $\flat$ *il existe un nombre réel positif* $\sharp$*, qui dépend de* $\flat$ *[c.-à-d.* $\sharp(\flat)$*], tel que* $0 < \sqrt{(\clubsuit-\star)^2 + (\heartsuit-\circ)^2} < \sharp$ *implique que* $|f(\clubsuit,\heartsuit) - \text{m}| < \flat$. (2)

signifie exactement la même chose que (1).

De plus, pour exprimer que l'interprétation d'une variable ($\delta$) dépend d'une autre variable ($\varepsilon$), les mathématiciens utilisent des expressions comme $\delta = \delta(\varepsilon)$. Cette notation indique que la valeur de $\delta$ varie en fonction de celle de $\varepsilon$. La valeur d'une variable peut aussi dépendre de plus d'une variable. C'est notamment le cas de $f(x,y)$ qui dépend de $x$ et de $y$ à la fois.

[1] Tiré de [1], page 52.

Tout comme (1) page précédente, le français[2] utilise aussi des mécanismes pour introduire un nouveau référent dans un texte (une nouvelle variable), le réutiliser (**coréférence**) et marquer les liens de dépendance référentielle d'un élément vis-à-vis d'un autre élément du texte (**anaphore**)[3].

Pour l'illustrer, nous avons marqué avec un indice séquentiel chacune des trente expressions référant à une personne dans le texte suivant :

> *Une femme*$_1$ *demande à* **Donna**$_2$ *de* **lui**$_3$ *créer une robe spécialement pour une séance de spiritisme.* **Celle-ci**$_4$ *voudrait entrer en communication avec* **son défunt mari**$_5$ *afin qu'*$\mathbf{il}_6$ $\mathbf{l'}_7$*aide à retrouver son bracelet de diamants.* **Steve**$_8$ *demande à* **Carly**$_9$ *de* $\mathbf{l'}_{10}$*accompagner au bal de la moisson et* **Cooper**$_{11}$ *en fait de même avec* **Val**$_{12}$*. La situation financière de* **David**$_{13}$ *ne s'arrange pas. On* **lui**$_{14}$ *refuse même une demande de crédit. Au bal,* **Brandon**$_{15}$*,* **Kelly**$_{16}$*,* **Steve**$_{17}$ *et* **Carly**$_{18}$ *sont surpris de voir* **Val**$_{19}$ *et* **Cooper**$_{20}$ *danser ensemble.* **Donna**$_{21}$ *arrive seule au bal et provoque la jalousie de* **Val**$_{22}$ *en flirtant avec* **Noah**$_{23}$*.* **Ce dernier**$_{24}$ *s'approche tout de même de* **Val**$_{25}$ *et provoque maintenant la jalousie de* **Cooper**$_{26}$ *qui décide de s'en aller. Mais lorsque* **Noah**$_{27}$ *voit le bracelet que* **Cooper**$_{28}$ *a offert à* **Val**$_{29}$*,* $\mathbf{il}_{30}$ *devient furieux*[4]*.* (3)

Dans ce texte, le formalisme est quelque peu laissé de côté au profit d'une lecture plus naturelle. Ce texte est toutefois similaire à (1) sur certains points :

- La première occurrence d'un **nom propre** introduit un nouveau référent. Cependant, dans ce cas-ci, il est impossible de substituer un nom propre à un autre nom propre car ce nom propre est déjà rattaché à un personnage. Il est donc probable que les auditeurs de l'émission *Beverly Hills* connaissent Donna, Steve, Carly, Cooper, Val, David, Brandon, Kelly et Noah.
- Une utilisation subséquente du même nom propre suppose un lien de coréférence entre les deux noms propres.
- Des liens de dépendance entre des éléments référentiels existent dans (3). Ils sont toutefois marqués différemment de (1). Dans (3), la dépendance est signalée par l'utilisation de certains éléments linguistiques anaphoriques comme les pronoms (*lui*, *Celle-ci*, *il* et *l'*) et par les déterminants de certains syntagmes nominaux (*son défunt mari* et *Ce dernier*).

## 1.1 Problématique

Formellement, nous pouvons considérer que l'anaphore est un sous-ensemble de la coréférence car un élément anaphorique et son antécédent réfèrent au même objet (ou personne) du monde du discours ; ils sont donc coréférents. Toutefois, le traitement des anaphores est

[2] Comme si (1) n'était pas écrit en français !

[3] Les concepts d'expression référentielle, de coréférence et d'anaphore seront décrits plus formellement au chapitre 2 page 13.

[4] Tiré de [2].

très différent de celui des coréférences et mérite une attention particulière. Nous reparlerons d'ailleurs de ces différences au chapitre suivant (section 2.2.1 page 20).

Nous appellerons l'opération consistant à repérer l'antécédent des éléments anaphoriques d'un texte **résolution des anaphores**. Nous verrons dans ce mémoire que cette opération est une étape importante pour presque tous les domaines du traitement automatique de la langue. Cependant, la résolution des anaphores est une étape assez difficile à traiter de façon automatique.

Revoyons l'exemple (3) page précédente en ajoutant, cette fois-ci, les liens coréférentiels (notés *coref=id* où *id* est l'expression coréférentielle **la plus récente**) et les liens anaphoriques (notés *anaph=id* où *id* est l'expression coréférentielle **non anaphorique** la plus récente) :

> ***Une femme***$_1$ *demande à* ***Donna***$_2$ *de* ***lui***$_{3,anaph=1}$ *créer une robe spécialement pour une séance de spiritisme.* ***Celle-ci***$_{4,anaph=1}$ *voudrait entrer en communication avec* ***son défunt mari***$_{5,anaph=1(possession)}$ *afin qu'****il***$_{6,anaph=5}$ ***l'***$_{7,anaph=4}$ *aide à retrouver son bracelet de diamants.* ***Steve***$_8$ *demande à* ***Carly***$_9$ *de* ***l'***$_{10,anaph=8}$*accompagner au bal de la moisson et* ***Cooper***$_{11}$ *en fait de même avec* ***Val***$_{12}$*. La situation financière de* ***David***$_{13}$ *ne s'arrange pas. On* ***lui***$_{14,anaph=13}$ *refuse même une demande de crédit. Au bal,* ***Brandon***$_{15}$*,* ***Kelly***$_{16}$*,* ***Steve***$_{17,coref=8}$ *et* ***Carly***$_{18,coref=9}$ *sont surpris de voir* ***Val***$_{19,coref=12}$ *et* ***Cooper***$_{20,coref=11}$ *danser ensemble.* ***Donna***$_{21,coref=2}$ *arrive seule au bal et provoque la jalousie de* ***Val***$_{22,coref=19}$ *en flirtant avec* ***Noah***$_{23}$*.* ***Ce dernier***$_{24,anaph=23}$ *s'approche tout de même de* ***Val***$_{25,coref=22}$ *et provoque maintenant la jalousie de* ***Cooper***$_{26,coref=20}$ *qui décide de s'en aller. Mais lorsque* ***Noah***$_{27,coref=23}$ *voit le bracelet que* ***Cooper***$_{28,coref=26}$ *a offert à* ***Val***$_{29,coref=25}$*,* ***il***$_{30,anaph=27}$ *devient furieux.* (4)

Voici quelques points intéressants à propos des relations anaphoriques :

**Position de l'antécédent :** Dans ce texte, l'antécédent des anaphores est un autre élément linguistique situé avant (et pas très loin). Dans les textes en général, les antécédents précèdent presque toujours les éléments anaphoriques. Toutefois, il peut y avoir des cas où l'antécédent se situe après l'élément anaphorique pour créer un effet de style[5].

> *Mais lorsqu'****il***$_{1,cataph=4}$ *voit le bracelet que* ***Cooper***$_2$ *a offert à* ***Val***$_3$*,* ***Noah***$_4$ *devient furieux.* (5)

**Connaissances linguistiques :** Dans les phrases suivantes, l'antécédent de *Il* (au masculin) est *Noah* car il est un homme. Ceci se vérifie en permutant les noms propres de la phrase la précédant :

> ***Donna***$_1$ *arrive seule au bal et provoque la jalousie de* ***Val***$_2$ *en flirtant avec* ***Noah***$_3$*.* ***Il***$_{4,anaph=3}$ *[...]* (6-a)

> ***Donna***$_1$ *arrive seule au bal et provoque la jalousie de* ***Noah***$_2$ *en flirtant avec* ***Val***$_3$*.* ***Il***$_{4,anaph=2}$ *[...]* (6-b)

*Noah*$_1$ *arrive seul au bal et provoque la jalousie de* ***Val***$_2$ *en flirtant avec* ***Donna***$_3$. ***Il***$_{4,anaph=1}$ *[...]* (6-c)

**Connaissances du monde :** Déterminer le sexe des personnages est difficile si l'on ne connaît aucun des personnages de *Beverly Hills*. Savoir que certains noms propres sont typiquement masculins ou typiquement féminins facilite un peu le problème. De plus, en supposant qu'habituellement une femme accompagne un homme à un bal (et vice-versa), un petit exercice de déduction peut aider à établir les sexes manquants.

**Connaissances de la situation d'énonciation :** Le pronom *Celle-ci* est rattaché à *Une femme* car nous déduisons que puisque la femme veut une robe spécialement conçue pour une séance de spiritisme, il est probable que ce soit elle qui veuille entrer en communication avec son défunt mari. Cependant, une situation où Donna voudrait entrer en communication avec son défunt mari n'est pas impossible. On peut imaginer que la femme avec laquelle Donna veut faire affaire pour la séance de spiritisme a perdu sa robe spéciale réservée à cet effet. Sachant ce fait, la lecture des deux premières phrases de (3) page 3 aurait une autre signification.

**Type d'anaphore :** Bien que *Une femme* et *son défunt mari* aient des référents disjoints (dans le monde du discours), nous ne pouvons dire qu'ils sont totalement indépendants l'un de l'autre car une partie de l'interprétation de *son défunt mari* se trouve dans l'interprétation de *Une femme*. L'expression *son défunt mari* peut être paraphrasée par *le défunt mari de la femme*. Il existerait donc plus d'un type de dépendance entre les items référentiels d'un texte. Même si ces cas sont très importants, nous limiterons notre étude sur les liens coréférentiels simples (identité des références).

Pour faire suite à la petite analyse des liens référentiels de (4) page précédente, nous constatons qu'il est parfois difficile de trouver le bon antécédent de certains éléments anaphoriques. La résolution des liens anaphoriques demande parfois des connaissances linguistiques, des connaissances générales à propos du monde ou des connaissances des conditions d'énonciation. Implicitement, cela suggère que pour le traitement automatique des liens anaphoriques d'un texte, il est nécessaire de disposer de règles linguistiques assez complexes, d'un dictionnaire complet et d'une base de connaissances générales.

Malheureusement, les ressources linguistiques actuellement disponibles, particulièrement celles en français, ne sont pas adéquates pour une telle tâche. Les dictionnaires électroniques ne sont pas facilement utilisables par un programme informatique[6]. De plus, les bases de connaissances générales à propos du monde n'existent pas encore. Pour compléter ce manque d'informations, il peut être pratique d'avoir à notre disposition un corpus de textes annotés (de préférence). Ces corpus permettent d'extraire des tendances statistiques utiles dans les méthodes quantitatives. Cependant, les corpus doivent habituellement être construits de toutes pièces car il n'existe que très peu de corpus français accessibles à tous. **Pour toutes ces raisons,**

[5]Ce cas est nommé **cataphore**. Les cataphores sont souvent marquées par l'utilisation de « :» (comme dans *Jean n'aimait qu'une seule femme*$_{1,cataph=2}$ *: Jeanne*$_2$).

[6]D'ailleurs, il n'existe pas d'équivalent français (gratuit) de Wordnet (`http://www.cogsci.princeton.edu/~wn/`).

**il est impossible, à l'heure actuelle, de faire un traitement automatique parfait des liens anaphoriques des textes.**

## 1.2 Objectifs

Dans ce mémoire, nous nous intéresserons particulièrement à la différence des moyens coréférentiels selon le type de texte. Nous nous placerons dans une perspective de traitement automatique de la langue et nous nous pencherons sur le problème de l'identification des liens coréférentiels. Le but général de notre démarche est de dégager, à partir d'une analyse faite sur un certain nombre de textes appartenant à des domaines différents, des généralités et des particularités de l'emploi des expressions référentielles. Le but de notre travail est de faciliter l'établissement des liens coréférentiels de façon automatique. En exploitant les tendances des textes selon le domaine, nous tenterons de raffiner et d'ajuster les règles existantes de résolution des liens coréférentiels dans le but d'en augmenter les performances.

De manière plus précise, nous nous attarderons sur les points suivants :

- Caractériser les facteurs linguistiques influençant la coréférence.
- Dégager les comportements coréférentiels variant selon le type de texte.
- Concevoir une façon de baliser la coréférence dans le but de faciliter le traitement informatique.
- Adapter au français certains travaux permettant la résolution des liens coréférentiels qui ont été faits pour des textes anglais.
- Déterminer les ressources linguistiques minimales nécessaires au traitement automatique des liens coréférentiels.

### 1.2.1 Facteurs influençant la coréférence

Depuis le début de cette introduction, nous avons pu constater que le phénomène de la coréférence, plus particulièrement celui de l'anaphore, est un sujet riche et complexe. Les coréférences mettent en jeux toutes sortes d'éléments linguistiques : les noms propres, les pronoms, les déterminants démonstratifs, les déterminants définis et les déterminants possessifs. De plus, certains facteurs peuvent être utiles pour trouver l'antécédent des éléments anaphoriques : le genre et le nombre, la distance entre les éléments référentiels, la situation d'énonciation, etc.

Une première étape d'analyse consiste à déterminer les facteurs influençant les mécanismes référentiels et coréférentiels se produisant à l'intérieur d'un texte. Nous nous y attarderons dans les premiers chapitres de ce mémoire en étudiant la littérature d'un point de vue théorique et pratique.

### 1.2.2 Comportement coréférentiel selon le type de texte

En comparant les exemples (1) page 2 et (3) page 3, nous avons déjà les premiers indices qui supposent que la coréférence et l'anaphore sont des phénomènes qui dépendent du **type de texte**. Pour s'en convaincre, voici un autre exemple de type de texte : les recettes de cuisine[7].

**Mayonnaise à l'estragon**

*1 oeuf biologique*
*1 à 1$\frac{1}{2}$ c. à table (15 à 22 mL) de vinaigre de cidre de pommes*
*$\frac{1}{2}$ c. à thé (2 mL) de sel de mer*
*1 c. à thé (5 mL) de moutarde en poudre*
*1 c. à thé (5 mL) de persil frais ou séché*
*$\frac{1}{4}$ c. à thé (1 mL) de basilic*
*$\frac{3}{4}$ c. à thé (3 mL) d'estragon*
*$\frac{1}{4}$ c. à thé (1 mL) de poivre*
*1 tasse (250 mL) d'huile végétale pressée à froid*

*Mettre* ***tous les éléments*** *dans* ***le blender*** *(sauf* ***l'huile****), bien mélanger [∅].*
*Ajouter ensuite* ***l'huile*** *en filet (pendant que* ***le blender*** *fonctionne à basse vitesse) [à ∅].*
*Battre [∅] jusqu'à ce que* ***le tout*** *soit homogène. Il est possible que vous ne preniez pas toute* ***l'huile****.*
*Réfrigérer [∅]*[8]*.* (7)

Dans ce texte, les procédés anaphoriques diffèrent de (1) et de (3) :

- Les «∅» entre les crochets ont été ajoutés pour illustrer les «vides» que comportent les recettes de cuisine. Ces vides sont d'ailleurs des éléments anaphoriques importants et typiques de ces textes ; ils représentent habituellement un élément en transformation.
- Les syntagmes comme *le tout* et *le mélange* servent à référer à des entités qu'il serait difficile de nommer en utilisant un concept plus précis.
- L'utilisation des syntagmes définis renvoie à un (ou plusieurs) ingrédient de la recette (*tous les éléments* et *l'huile*) ou à un ustensile associé au domaine de la cuisine (*le blender*).

Bien que les phénomènes coréférentiels aient été étudiés dans quelques types de textes particuliers –notamment les textes littéraires, les textes procéduraux et les fusions de compagnies– il n'existe pas, à notre connaissance, d'étude comparative des phénomènes coréférentiels selon le type de texte français basé sur un corpus important[9]. Nous tenterons, dans un deuxième temps, de départager les phénomènes coréférentiels propres à quelques types de textes particuliers des phénomènes communs à tous les textes en général. Pour ce faire, nous effectuerons

[7] Les anaphores contenues dans les recettes de cuisine ont été étudiés par A. Tutin dans [Tut92].
[8] Tiré de [3].
[9] Voir toutefois les travaux de R. Kittredge [Kit82] pour une étude comparative sommaire.

une comparaison portant sur trois types de textes : les fusions de compagnies, les critiques de films et les *HOWTO Linux* français[10].

### 1.2.3 Schématisation de la coréférence

Les liens coréférentiels sont très nombreux dans les textes en français. Aussi, il arrive fréquemment que plusieurs éléments linguistiques soient coréférentiels entre eux. Tous les éléments linguistiques se référant à un même référent (entité du monde du discours) peuvent être regroupés ensemble dans une **chaîne coréférentielle**. Pour analyser et manipuler les chaînes coréférentielles dans les textes, il importe de concevoir une façon de les représenter.

Bien que la notation introduite en (4) page 4 puisse être acceptable pour un texte très court, elle semble inadéquate pour un texte plus long. La complexité de l'ensemble des relations coréférentielles d'un texte peut expliquer le problème :

- Les chaînes coréférentielles sont souvent très longues et sont réparties sur toute la longueur du texte ; c'est habituellement le cas pour l'entité sujet du texte (ce dont le texte parle).
- Il est aussi fréquent de retrouver des chaînes coréférentielles qui s'entremêlent les unes avec les autres. Par exemple, dans le cas des textes dont le sujet est la fusion de deux compagnies, il est probable que les chaînes coréférentielles référant aux compagnies soient en parallèle dans le texte.

Nous examinerons comment l'utilisation des outils *XML* peut faciliter la manipulation des chaînes coréférentielles. Nous verrons que *XML* peut être utile pour la présentation des chaînes coréférentielles, pour l'interrogation du corpus (analyses de fréquences, tableaux statistiques, etc.) ainsi que pour le traitement informatique des données.

### 1.2.4 Adaptation au français

L'établissement automatique des liens coréférentiels a connu une forte popularité depuis quelques années. Les conférences *MUC* (*Message Understanding Conference*) ont grandement contribué à cet enthousiasme. Le but premier des conférences *MUC* était d'automatiser l'extraction automatique d'informations. Un sous-problème de l'extraction d'informations, abordé dans les conférences *MUC-7*, était l'établissement des liens coréférentiels d'un texte. Lors des épreuves *MUC*, plusieurs algorithmes permettant l'établissement des liens coréférentiels ont été proposés pour des textes en anglais (voir chapitre 3 page 26).

Nous ne connaissons cependant que très peu de travaux portant sur les textes en français[11]. Nous tenterons donc de voir comment adapter les algorithmes faits pour des textes anglais et les appliquer à des textes français.

[10] Il existe une distinction entre sous-langage et type de texte (voir [Kit82]). Cependant, cette distinction n'est pas cruciale pour nos conclusions. Nous utilisons donc le terme *type de texte* de façon assez large sans distinguer si les regroupements constituent un sous-langage.

[11] Nous devons toutefois citer les travaux de F. Trouilleux [Tro01] et de M. Dupont [Dup02].

### 1.2.5 Minimalisation des ressources

La coréférence en général est un phénomène très vaste et, pour le moment, très difficile. Nous avons soulevé, à la section 1.1 page 3, que traiter tous les phénomènes coréférentiels d'un texte est impossible à l'heure actuelle car les ressources linguistiques (tels les dictionnaires, les bases de connaissances et les corpus balisés) sont présentement insuffisantes pour le traitement des anaphores. De plus, nous ignorons encore l'ensemble des principes qui gèrent l'emploi des expressions coréférentielles dans toutes les constructions syntaxiques[12].

Dans les exemples précédents, nous avons d'ailleurs été assez sélectifs dans le choix des entités identifiées : les personnages de (3) page 3 ou les ingrédients et les instruments de (7) page 7. Empiriquement, il semblerait que certaines chaînes coréférentielles puissent être plus faciles à identifier. Ces chaînes pourraient aussi être des chaînes importantes pour comprendre le message du texte. Il nous semble envisageable de limiter notre étude à ces chaînes.

Dans ce mémoire, nous porterons une attention particulière aux chaînes coréférentielles comportant un nom propre. Nous étudierons la possibilité de faire plus avec moins. À l'aide d'un dictionnaire de quelques centaines de mots, nous croyons qu'il est possible de construire un algorithme suffisamment simple qui puisse dégager les chaînes coréférentielles importantes d'un texte de façon assez précise.

## 1.3 Structure du mémoire

Dans le chapitre *Notions théoriques* (chapitre 2) nous étudierons les concepts théoriques liés à la notion de référence. Nous examinerons ensuite, dans le chapitre *Travaux précédents* (chapitre 3) quelques méthodes pratiques visant à résoudre automatiquement les liens coréférentiels de façon parcimonieuse, c'est-à-dire utilisant des ressources linguistiques simples.

Au chapitre *Outils XML* (chapitre 4), nous ferons une brève exploration de quelques langages de balisage liés à *XML*. Par la même occasion, nous développerons des moyens pour schématiser, visualiser et interroger les chaînes coréférentielles d'un texte.

Nous décrirons, au chapitre *Méthodologie* (chapitre 5), les étapes techniques de la préparation du corpus, de l'élaboration de l'algorithme ainsi que les métriques utilisées pour mesurer la pertinence des résultats.

Nous présenterons ensuite, au chapitre *Présentation et analyse des résultats* (chapitre 6), les grandes lignes de notre algorithme permettant d'établir les chaînes importantes d'un texte. Nous y exposerons les points de variation possibles selon le type de texte ainsi que les raccourcis choisis pour simplifier le problème. Nous présenterons aussi quelques résultats finaux et ferons une brève évaluation des résultats.

Enfin, dans le chapitre *Discussion* (chapitre 7), nous examinerons les défis que pose la résolution des liens coréférentiels selon le type de texte. Nous présenterons une évaluation de la

[12] Nous nous limiterons, dans ce mémoire, aux constructions les plus fréquentes.

nature du travail à effectuer pour l'application de notre algorithme à un nouveau type de texte. Nous discuterons finalement de certaines améliorations permettant d'améliorer le rendement de notre algorithme.

## 1.4 Applications possibles

Un tel cheminement serait inutile sans un emploi concret dans le monde réel. De façon générale, notre approche peut servir pour toute application exigeant un peu de «compréhension» du texte source. Au cours de notre exposé, nous garderons en tête que notre projet peut, tant au niveau de la synthèse de textes que de l'analyse de textes, aider à quelques mises en pratique :

- l'interrogation et l'indexation automatique ;
- l'extraction d'informations et la création automatique de résumés ;
- la synthèse de textes ;
- la traduction automatique.

### 1.4.1 Interrogation et indexation automatique

L'identification des chaînes coréférentielles est très importante en recherche d'informations (B. Baldwin [Bal97]). Un principe de base est que si un texte contient plusieurs occurrences d'un terme, ce terme est possiblement en relation avec le sujet du texte.

Un moteur de recherche utilise habituellement ce principe pour isoler les documents correspondants à une requête de l'utilisateur. Pour extraire et ordonnancer les documents (d'une base de documents) correspondant le mieux à une requête, il y a deux stratégies possibles :

- Classer les documents selon le nombre de mots correspondant à la requête.
- Déterminer les documents contenant tous les mots de la requête dans un contexte rapproché.

Il arrive fréquemment qu'un terme important dans un texte soit réalisé par un pronom (ou tout autre élément anaphorique). Il est donc avantageux de résoudre les liens anaphoriques avant de calculer la pertinence des documents de manière à avoir des résultats plus représentatifs de la situation.

### 1.4.2 Extraction d'informations et résumé automatique

L'établissement des chaînes coréférentielles peut s'avérer être un outil très utile à l'extraction d'informations et à la création de résumés automatiques. Il existe des projets (S. Bergler [BWK+03]) qui utilisent directement les mots des chaînes coréférentielles les plus longues pour parvenir à résumer un texte en dix mots. On suppose dans ce cas que les chaînes coréférentielles longues représentent probablement une entité importante.

Pour les résumés plus longs, il est cependant nécessaire d'utiliser un procédé plus complexe. Pour ce faire, il est habituellement nécessaire d'extraire certaines phrases jugées importantes dans le texte. Dans ces cas, une simple agglomération de ces phrases n'est pas suffisante pour construire un résumé cohésif (R. Kittredge [Kit02]) ; les phrases ne semblent pas former un texte suivi car les pronoms présents dans les phrases recueillies ne peuvent pas être résolus correctement. Pour pallier la situation, il est judicieux de résoudre les anaphores du texte avant de sélectionner les phrases importantes.

### 1.4.3 Synthèse de textes

Dans le cas de la synthèse de textes, les problèmes anaphoriques se posent dans le sens inverse de ce que nous faisons ici : sachant que deux éléments sont coréférentiels dans le texte, il faut choisir la façon dont le second élément sera réalisé. En particulier, si la distance entre les deux coréférents est petite, il sera habituellement plus naturel de choisir un pronom ou une autre réalisation anaphorique.

Malgré cette inversion des procédés, les facteurs et les mécanismes référentiels de la synthèse de textes sont assez similaires à ceux impliqués dans la résolution automatique des liens anaphoriques.

Une étude portant sur la génération automatique des liens anaphoriques des textes de recettes de cuisine [KTKL96] relève quelques critères importants à considérer pour le choix d'une réalisation anaphorique :

- la non-ambiguïté ;
- la focalisation ;
- la distance ;
- les critères lexico-grammaticaux (fonctions syntaxiques) ;
- les critères lexico-sémantiques (la nature des éléments coréférentiels) ;
- les contraintes conceptuelles.

Nous constaterons que ces critères sont très semblables à ceux qui s'appliquent à la résolution des liens anaphoriques.

### 1.4.4 Traduction automatique

Puisque le système de pronoms n'est pas identique d'une langue à une autre, déterminer les antécédents des pronoms est essentiel pour arriver à une bonne traduction. On peut penser aux pronoms français et anglais qui ne portent pas exactement les mêmes traits syntaxiques et sémantiques. Si le français distingue habituellement le genre et le nombre, l'anglais préfère distinguer le nombre et l'animation (animé ou inanimé).

Par exemple, le pronom *l'* en position *C.O.D* peut souvent être traduit par *it*. Cependant, il peut aussi correspondre à *he* ou *her* dans certains contextes. M. Salkoff ([Sal99] page 162) expose certaines problématiques de la traduction du pronom *le* :

> *«When the object of a verb is a noun phrase, and the latter can refer either to Nh ('human') or Nc ('concrete'), then* **le** *is ambiguous between* ***him/it****,* **la** *between* ***her/it****, and* **l'** *between her/him/it. In the last case,* **l'** *may be partially disambiguated by the gender of the following participle :»*

Et donne comme exemple :

*Je l'ai (frappé + observé) → I (struck + observed) him/it* (8-a)

*Je l'ai (désignée + poussée) → I (designated + pushed) her/it* (8-b)

# Chapitre 2

# Notions théoriques

*«Meschtroumpfs les jurés ! On dit : Ne vous schtroumpfez pas aux apparences et ne schtroumpfez jamais les schtroumpfs sur la mine ! Qui schtroumpfe bien, châtie bien ! Dura schtroumpf, sed schtroumpf ! Et il faut schtroumpfer le bon schtroumpf de l'ivraie ! Et l'ivraie, c'est la Schtroumpfette*[1] *!»*

***– Schtroumpf à lunettes***

Depuis Frege et jusqu'à nos jours, nous nous sommes demandé comment nous pouvions rejoindre le **monde extra-linguistique** (parler de l'univers réel ou fictif) avec les simples éléments linguistiques (les mots) qui sont à notre disposition. Les réponses à cette question ne sont certes pas simples, nous n'avons qu'à penser au vocabulaire engendré par la question pour nous en donner une vague idée : *référer, dénoter, désigner, renvoyer à, connoter, signifier, référence virtuelle, classe conceptuelle,* etc.

**Définition 2.1 (référence)**
*La référence est la fonction par laquelle un signe linguistique renvoie à un objet du monde extra-linguistique, réel ou imaginaire.*

Cette définition (tirée de J. Dubois [DGG+94]) contient l'élément le plus important de la référence : le renvoi à un objet du monde extra-linguistique. Quoique la notion de référence ait beaucoup intéressé les philosophes, elle peut aussi être abordée d'un point de vue linguistique (voir les travaux de G. Kleiber [Kle81], de qui nous tirons une bonne partie de la théorie de cette section). Nous nous intéresserons d'ailleurs à cette notion sous deux aspects :

1. la nature des éléments référentiels ;
2. les liens possibles entre les éléments référentiels.

[1] Tiré de [4].

## 2.1 Syntagmes référentiels

Intuitivement, il semble que les syntagmes nominaux occupent une place de choix dans le phénomène de la référence. En étudiant chacun des éléments qui constituent le syntagme nominal, nous pouvons entrevoir pourquoi il en est ainsi.

Nous verrons que le syntagme nominal détermine un objet unique (un référent unique) grâce aux éléments qui le composent. Les têtes syntaxiques des syntagmes référentiels (**têtes$_r$**[2]) – typiquement les noms communs– permettent d'identifier une classe d'objets partageant une propriété importante : celle de se nommer $x$. Cette propriété est parfois connue sous le nom de **dénotation**. Les déterminants, quant à eux, isolent un objet unique de la classe référentielle représentée par $x$. Le déterminant permet donc de **déterminer** ou de **désigner** un objet (dont le nom est $x$) unique.

### 2.1.1 Têtes$_r$

Selon G. Kleiber [Kle81], tout item lexical (c'est-à-dire tout nom commun, adjectif ou verbe) possède une certaine aptitude à la référence. Cet item présuppose qu'il existe un concept qui lui est rattaché. C'est ainsi que nous réussissons à sortir du monde linguistique avec les mots ; les mots lexicaux évoquent des concepts de l'univers du discours. Par exemple, pour comprendre la phrase :

*Le cheval galope dans l'enclos.* (1)

le lecteur à besoin de certaines connaissances extra-linguistiques, notamment de savoir ce que sont ⟨cheval⟩, ⟨enclos⟩, et ⟨galoper⟩.

La présupposition d'existence se produit même lorsque le lecteur rencontre un mot dont il ne connaît pas la signification. Dans ce cas, le lien référentiel avec la situation du discours se fait quand même, même s'il est difficile de déterminer avec quel objet exactement. La citation en exergue de ce chapitre illustre un cas particulier où le mot *schtroumpf* signifie à peu près n'importe quoi. Nous remarquons que le sens du texte n'en est toutefois pas trop affecté.

Nous considérons qu'une tête$_r$ peut être de trois natures :

- un nom commun ;
- un nom propre ;
- un pronom ou une ellipse.

#### 2.1.1.1 Nom commun

Bien que tous les items lexicaux soient référentiels, il nous semble que les noms communs le soient plus que les autres. Nous expliquons ce fait grâce à la notion de **classe référentielle**[3].

[2] Nous distinguerons les têtes syntaxiques des syntagmes référentiels des têtes de chaînes coréférentielles. Ces deux notions seront notées tête$_r$ et tête$_c$ respectivement.

[3] Notion empruntée à G. Kleiber [Kle81].

**DÉFINITION 2.2 (CLASSE RÉFÉRENTIELLE)**
*La classe référentielle est l'ensemble conceptuel de tous les particuliers réunis sous l'étiquette de l'item lexical.*

Par exemple, le mot *cheval* évoque l'ensemble de tous les chevaux, le mot *riz* évoque l'ensemble de tous les grains de riz et le mot *blanc* évoque l'ensemble de tout ce qui est blanc. Nous notons que les éléments du dernier ensemble (représentés par un adjectif) sont des objets potentiellement disparates, la seule contrainte étant qu'ils soient blancs. Nous pouvons donc y retrouver des chemises, des crayons ainsi que des grains de riz.

Les classes référentielles représentées par des noms communs sont les plus homogènes. Ceci explique pour nous le statut particulier des noms communs dans la référence.

#### 2.1.1.2 Nom propre

Typiquement, le nom propre est référentiel. Nous devons donc l'inclure dans notre étude sur la référence. Quoiqu'il partage quelques similitudes avec le nom commun, le nom propre mérite certainement un traitement différent. Quelques ouvrages se sont penchés sur le statut particulier du nom propre. Nous retenons les travaux de G. Kleiber [Kle81], K. Jonasson [Jon94] et Gary-Prieur [GP94].

Pour beaucoup d'auteurs, les noms propres sont justement dépourvus de contenu lexical ; ils ne sont donc pas des noms communs. Les noms propres ont d'ailleurs des comportements particuliers : ils se composent difficilement avec des déterminants, ils débutent habituellement par une majuscule, ils sont pauvres morphologiquement et n'aiment pas beaucoup les modificateurs :

*Pierre a terminé son examen.* (2-a)

* *Le Pierre a terminé son examen.* (2-b)

*Les Pierre ont terminé leurs examens.* (2-c)

* *Les Pierres ont terminé leurs examens.* (2-d)

* *Pierre gentil a terminé son examen.* (2-e)

De plus, il est plutôt difficile de construire la définition d'un nom propre donné (comme *Pierre*). Au mieux, dans un dictionnaire des noms propres, il est possible de décrire l'entrée vedette associée à un nom propre en donnant les caractéristiques de l'entité (personne, pays, événement, etc.) qui lui sont attachées. Ce n'est cependant pas ce que nous pouvons appeler une définition.

Malgré le fait que les noms propres n'aient pas à proprement parler de contenu lexical, ils peuvent avoir plusieurs sens[4]. Le sens le plus important pour nous est celui de dénomination, c'est-à-dire ‹être appelé $x$›[5]. Ce sens provient d'un acte de baptême initial illustré ainsi[6] :

[4]Dans notre étude, nous ne parlerons pas des autres sens du nom propre, voir K. Jonasson [Jon94] et Gary-Prieur [GP94] pour plus de détails.
[5]Voir G. Kleiber [Kle81].
[6]Voir S. Kripke [Kri80] page 91.

> *«Someone, let's say, a baby, is born ; his parents call him by a certain name. They talk about him to their friends. Other people meet him. Through various sorts of talk the name is spread from link to link as if by a chain.»*

Nous nommons ainsi les personnes, les villes, les rues, les immeubles, les pays, les régions, les entités géographiques (eau, montagne), etc. Ce sont plus ou moins les conventions socio-culturelles qui déterminent qui ou quoi peut porter un nom propre. Ces entités reçoivent un nom habituellement par économie ; il est plus pratique de leur donner un nom unique que de décrire chaque fois l'entité ou l'individu dont nous voulons parler.

Nous avons déjà soulevé la place importante de la classe référentielle pour les syntagmes nominaux. La classe identifiée par l'item lexical sert, en quelque sorte, d'intermédiaire entre le référent et l'expression linguistique. Pour le nom propre, un tel intermédiaire n'est pas nécessaire. Le nom propre peut directement déterminer le référent car il présuppose, à lui seul, l'existence et l'unicité du référent. Les noms propres peuvent se passer de la classe référentielle pour l'identification du référent grâce au **désignateur rigide**, c'est-à-dire une association **directe** et **durable** avec le référent. Le désignateur rigide est une notion qui a été élaborée par Kripke[7] :

> *«'[...] the number of planets might have been different from what it in fact is*[8]*. It doesn't make any sense, though, to say that nine might have been different from what it in fact is'. [...] Let's call something a rigid designator if in every possible world it designates the same object, a nonrigid or accidental designator if that is not the case. [...] proper names are rigid designators [...]»*

#### 2.1.1.3 Pronom et ellipse

Les pronoms et les ellipses sont des moyens permettant de réutiliser un référent précédemment introduit dans le discours. Ces deux types de syntagmes nécessitent une opération de résolution anaphorique (voir F. Corblin [Cor95]). Les syntagmes chapeautés par les pronoms et les ellipses sont totalement dépendants de leur antécédent car l'antécédent contient une grande partie de l'interprétation lexicale du syntagme. Cette dépendance garantit par le fait même l'unicité référentielle des syntagmes pronominaux et des ellipses.

##### PRONOM

Les pronoms personnels de troisième personne (*il, elle, ils, elles, on, le, la, se, en, les, lui, leur, lui, eux*) ainsi que les pronoms démonstratifs (*celui, celle, ceux, celles*) sont des éléments de cette catégorie. Une démarcation nette doit être faite entre les pronoms de première et de deuxième personne et ceux de troisième personne. Les premiers semblent être associés à la situation d'énonciation (le locuteur et l'interlocuteur). Ces pronoms ne sont habituellement pas considérés référentiels[9].

[7] Voir S. Kripke [Kri80] pages 48 et 49.
[8] D'ailleurs, avec la découverte récente de Sedna, il y a maintenant dix planètes dans le système solaire !
[9] Ou du moins, il n'y a pas d'antécédent car le référent est à l'extérieur du texte.

Les pronoms sont très dépendants du contexte et ne peuvent que très rarement fonctionner sans un antécédent. Ils contiennent peu d'informations lexicales ; souvent le genre ou le nombre seulement. Ces traits doivent cependant être compatibles avec le mot auquel ils sont rattachés.

ELLIPSE

Ce qui caractérise l'ellipse n'est pas la présence de certains mots, mais plutôt leur absence. Ces groupes nominaux nécessitent la fixation d'une tête lexicale par emprunt au contexte. Voici des exemples typiques[10] :

*J'ai trois crayons. Je préfère le ∅ bleu.*

*(le crayon bleu)* (3-a)

*J'ai trois crayons. Je préfère le ∅ premier.*

*(le premier des trois crayons)* (3-b)

Trouver l'antécédent associé à une ellipse nécessite toujours au moins deux étapes :

1. la fixation de la tête lexicale ;
2. l'interprétation du groupe nominal entier (indéfini, défini, démonstratif, etc.)[11]

### 2.1.2 Déterminants

Que se passe-t-il de particulier à l'intérieur d'un syntagme si les items nominaux réfèrent déjà ? *un cheval* et *le cheval* ne sont-ils pas aussi référentiels que *cheval* ? Lorsque nous utilisons un syntagme nominal, il y a une présupposition existentielle d'un référent **unique** satisfaisant au syntagme nominal entier. Dans le cas des syntagmes nominaux pluriels, il y a présupposition existentielle d'un groupe d'individus unique satisfaisant au syntagme (nous utiliserons habituellement des exemples au singulier pour ne pas alourdir inutilement le texte).

Nous avons déjà vu que la $\text{tête}_r$ (le nom commun) identifie une classe référentielle (comme la classe des chevaux). Les autres éléments du syntagme indiquent comment aller chercher l'individu particulier dans la classe référentielle. En plus des déterminants, les syntagmes nominaux peuvent contenir des modificateurs (adjectifs, compléments, etc.) Ces éléments servent à restreindre le nombre d'éléments possibles de la classe référentielle ; ils complètent l'information lexicale (comme dans *le cheval blessé*).

Ce sont cependant les déterminants qui remplissent le mieux la fonction individualisante des syntagmes nominaux[12]. Chacun des types de déterminant utilise toutefois des méthodes différentes pour le faire. En suivant F. Corblin [Cor87], nous classons les syntagmes nominaux en trois catégories : les indéfinis, les définis, et les démonstratifs[13].

[10] Nous utilisons le symbole ∅ pour marquer la position de l'élément manquant.
[11] Voir section 2.1.2.
[12] Les déterminants peuvent parfois être associés aux quantificateurs universels de la logique.
[13] La présence du déterminant est assez importante dans notre étude. Parfois, nous ferons l'économie de certains mots en remplaçant, par exemple *les syntagmes nominaux déterminés par un déterminant défini* par *les syntagmes définis*.

#### 2.1.2.1 Indéfini

Les syntagmes indéfinis peuvent être reconnus par la présence des déterminants suivants : *un, une des, du, deux, trois, ..., tout, chaque, n'importe quel, certain, plusieurs, aucun, quelques, nul.* Le déterminant indéfini permet d'isoler un objet par une opération d'extraction ou par une opération de dénombrement (répétition de l'extraction) sur la classe référentielle dénotée par le nom commun.

Si nous considérons :

***Un chat gris*** *miaule à la fenêtre.* (4-a)

***Des chats gris*** *miaulent à la fenêtre.* (4-b)

Dans le premier exemple, le déterminant *un* indique que nous parlons du chat pour la première fois et qu'il n'était pas déterminé à l'avance. Le procédé est similaire pour les expressions plurielles.

À la limite, si rien n'interdit la répétition de l'extraction –si le sens de la phrase n'est pas restrictif sur ce point ; ce peut être à cause d'une structure ressemblant à celle d'un proverbe– nous pouvons donner une interprétation générique (ou générale) à l'énoncé linguistique :

***Un chien*** *qui aboie ne mord pas.* (5-a)

***Chaque chien*** *qui aboie ne mord pas.* (5-b)

Nous ne parlons plus ici d'extraction d'un élément dans la classe référentielle, mais plutôt d'un dénombrement de la classe référentielle.

#### 2.1.2.2 Défini

Nous reconnaissons les syntagmes nominaux définis par la présence des déterminants *le, la, l', les,* ou de quelques mots contractés comme *au, aux, du, des.* L'ensemble des déterminants définis n'est pas une classe aussi homogène que celle de l'indéfini. Toujours, selon F. Corblin [Cor87], nous retrouvons cinq types de syntagmes nominaux définis[14] :

**Autonome (ou description définie) :**

***Le chat de la voisine***$_1$ *miaule à la fenêtre ...* (6-a)

**De reprise :**

*...* ***Le chat***$_{2b,anaph=1}$ *veut entrer dans la maison.* (6-b)

**Lexical :**

*...* ***Le félin***$_{2c,anaph=1}$ *veut entrer dans la maison.* (6-c)

**Associatif :**

*...* ***La chanson***$_{3,anaph=1(association)}$ *est peu mélodieuse.* (6-d)

**Générique :**

***Le chat***$_4$ *est un animal qui miaule.* (6-e)

L'emploi du déterminant défini suppose, dès le départ, qu'il existe un $x$ (de l'ensemble de tous les $x$ de la classe référentielle) unique. Pour l'identifier, le lecteur peut avoir recours aux autres référents du texte (défini de reprise, lexical ou associatif). Parfois, l'expression en elle-même contient toutes les informations pour déterminer le référent (défini autonome). Finalement, lorsqu'aucun antécédent ou référent n'est trouvé, ce syntagme reçoit une interprétation générique. Le défini permet toujours d'identifier un objet du monde. Dans le cas du générique, l'objet n'est plus un individu mais l'espèce (la classe) entière.

Il est à noter que les définis autonomes ressemblent un peu aux noms propres en ce sens qu'ils peuvent désigner un individu directement (voir 2.1.1.2 page 15) ; ils peuvent donc être considérés comme des désignateurs **semi-rigides**. Dans ces cas, le lien avec le référent n'est toutefois pas aussi fort qu'avec les noms propres. Ce lien peut être contingent à la situation d'énonciation, au lieu physique, au temps, etc. C'est pourquoi nous considérons que *le premier ministre canadien* n'est pas un nom propre car il peut changer dans le temps, par contre, *le fleuve Saint-Laurent* ou même *la Lune* sont des noms propres.

#### 2.1.2.3 Démonstratif

Les syntagmes nominaux démonstratifs s'identifient par la présence de l'un de ces mots : *ce, cet, cette, c', ces*. C'est par proximité que l'on parvient à identifier l'antécédent associé au syntagme démonstratif car l'antécédent n'est habituellement pas très loin :

***Le chat de la voisine***$_1$ *miaule à la fenêtre.* ***Ce soprano***$_{2,anaph=1}$ *me tombe sur les nerfs !* (7)

Le contenu du syntagme n'est pas essentiel pour déterminer l'objet mais permet parfois de le reclassifier dans une autre catégorie référentielle. Le but de cette reclassification est de présenter l'objet sous un autre jour :

*La nuit dernière, je me suis embarrée sur* ***le balcon***$_1$. ***Ce lit***$_{2,anaph=1(association)}$ *n'était pas très confortable.* (8)

## 2.2 Chaîne coréférentielle

Il n'est pas rare que plusieurs éléments textuels réfèrent au même objet extra-linguistique. Nous appelons ce regroupement d'expressions **chaîne de coréférence**. Particulièrement, nous nous intéressons aux types de relations reliant deux éléments référentiels et à la distribution des différents types de syntagmes nominaux à l'intérieur des chaînes de coréférence.

[14] La notation de F. Corblin inclue quatre types de déterminants définis. Nous croyons important de distinguer cependant les définis de reprise des définis lexicaux.

### 2.2.1 Relations

Le premier élément de la chaîne (selon l'ordre d'occurrence dans le texte) est, bien sûr, la tête de la chaîne coréférentielle ou plus simplement la **tête coréférentielle** (que nous notons $tête_c$). Les autres éléments dans la chaîne sont répartis selon leur dépendance vis-à-vis d'un autre élément référentiel. Il peut s'agir d'éléments coréférentiels ou anaphoriques. Typiquement, la coréférence s'emploie dans le cas des noms propres et l'anaphore s'emploie dans le cas des pronoms.

#### 2.2.1.1 Coréférence

Il y a coréférence lorsque deux signes linguistiques partagent la même référence (chacun des deux étant coréférent à l'autre). Notons qu'une relation de coréférence possède des propriétés mathématiques intéressantes[15] :

**Symétrie :** La coréférence est symétrique puisque la relation n'a pas à proprement parler d'orientation ; c'est-à-dire qu'aucun élément n'est dépendant de l'autre.

**Transitivité :** La relation est aussi transitive car si un élément référentiel $a$ est coréférentiel à un autre élément référentiel $b$ qui lui-même est coréférentiel à un troisième élément référentiel $c$, il s'en suit que $a$ est nécessairement coréférentiel à $c$.

**Réflexion :** Enfin, on peut considérer que la coréférence est une relation réflexive et que tout élément référentiel est coréférentiel à lui-même.

DÉFINITION 2.3 (SYMÉTRIE D'UNE RELATION)
*Soit $A$, un ensemble et $R$ une relation dans $A$, $R$ est symétrique* $\Leftrightarrow [\forall \langle a, b\rangle \in R \Rightarrow \langle b, a\rangle \in R]$. *La relation est non symétrique sinon.*

DÉFINITION 2.4 (TRANSITIVITÉ D'UNE RELATION)
*Soit $A$, un ensemble et $R$ une relation dans $A$, $R$ est transitive* $\Leftrightarrow [\forall \langle a, b\rangle$ *et* $\langle b, c\rangle \in R \Rightarrow \langle a, c\rangle \in R]$. *La relation est non transitive sinon.*

DÉFINITION 2.5 (RÉFLEXIVITÉ D'UNE RELATION)
*Soit $A$, un ensemble et $R$ une relation dans $A$, $R$ est réflexive* $\Leftrightarrow [\forall a \in A, \langle a, a\rangle \in R]$. *La relation est non réflexive sinon.*

Il est à remarquer que puisque les relations coréférentielles sont des relations **symétriques, transitives et réflexives,** elles sont donc des **relations d'équivalence**.

DÉFINITION 2.6 (RELATION D'ÉQUIVALENCE)
*Une relation d'équivalence est une relation qui est réflexive, symétrique et transitive.*

[15] Les définitions mathématiques suivantes sont tirées de B. Partee [PtMW90].

De façon plus concrète, le fait que les relations de coréférence soient des relations d'équivalence implique qu'il est possible de partager les éléments référentiels d'un texte en chaînes coréférentielles (classes d'équivalence) telles que toutes les expressions coréférentielles appartenant à une même chaîne de coréférence soient plus ou moins interchangeables (ou équivalentes) entre elles[16].

#### 2.2.1.2 Anaphore

La relation anaphorique est décrite par J. Dubois [DGG+94] :

> *«[...] l'anaphore est un processus syntaxique consistant à reprendre par un segment, un pronom en particulier, un autre segment du discours, un syntagme nominal antérieur, par exemple.»*

Présentée ainsi, il peut arriver que l'anaphore et la coréférence soient parfois confondues. D'ailleurs, la nuance est souvent laissée de côté ; *coréférence* et *anaphore* deviennent alors des synonymes. Par contre, pour F. Corblin, ces deux notions sont distinctes et aucune intersection entre les deux phénomènes n'est possible ([Cor87] page 10) :

> *«[...] un rapport d'identité éventuel entre deux termes dont les interprétations sont indépendantes ; s'il s'agit de référence, on parlera de co-référence.»*

> *«[...] un rapport de dépendance en vertu duquel B tire nécessairement son interprétation d'une mise en connexion à A, A saturant l'interprétation de B en fixant un de ses termes : on parlera alors d'anaphore.»*

Selon F. Corblin, la dépendance et l'indépendance (du contexte linguistique) sont des facteurs très importants et nécessitent deux études distinctes. En considérant la contrainte de dépendance entre les éléments référentiels, il en découle d'autres types de relations mathématiques entre les éléments antécédents et anaphoriques. Les anaphores sont aussi des relations transitives, mais elles sont **asymétriques** car la relation de dépendance oriente la relation. De plus, elles sont **irréflexives** car un élément référentiel ne peut pas être anaphorique à lui même.

**DÉFINITION 2.7 (ASYMÉTRIE D'UNE RELATION)**
*Soit A, un ensemble et R une relation dans A, R est asymétrique* $\Leftrightarrow [\forall \langle a, b\rangle \in R, \langle b, a\rangle \notin R]$.

**DÉFINITION 2.8 (IRRÉFLEXIVITÉ D'UNE RELATION)**
*Soit A, un ensemble et R une relation dans A, R est irréflexive* $\Leftrightarrow [\forall a \in A, \langle a, a\rangle \notin R]$.

[16] L'expression *chaîne coréférentielle* peut sembler inadéquate. Mathématiquement, il aurait peut-être été préférable d'utiliser l'expression *classe coréférentielle*. Ceci peut être expliqué par le fait que pour le lecteur, les éléments coréférentiels semblent être ordonnés dans le texte. Par ailleurs, nous ferons quelquefois un abus de langage en parlant de l'antécédent d'un élément coréférentiel. Puisque les éléments coréférentiels sont tous équivalents à l'intérieur d'une chaîne coréférentielle, il est faux de dire que le premier élément de la chaîne coréférentielle, par exemple, est l'antécédent de tous les autres éléments de la chaîne. Toutefois, il est parfois pratique de donner un nom à un élément coréférentiel situé avant un autre élément coréférentiel dans le texte. Comme nous le faisons pour l'anaphore, nous nommons cet élément *antécédent*.

**DÉFINITION 2.9 (RELATION D'ORDRE STRICT)**
*Une relation d'ordre strict est une relation qui est transitive, irréflexive et asymétrique.*

Les relations anaphoriques sont donc des **relations d'ordre strict**. Ceci implique qu'il n'est pas nécessairement possible de permuter un antécédent et un élément anaphorique correspondant dans le texte. Nous garderons à l'esprit la distinction entre anaphore et coréférence, mais pour des raisons pratiques et pour être compatible avec les autres travaux que nous étudierons, nous traiterons le plus souvent les deux phénomènes ensemble. Ainsi, pour nous, les anaphores sont un sous-ensemble des coréférences et sont donc aussi des éléments des chaînes coréférentielles.

### TYPES D'ANAPHORES

Nous pouvons tenter de classer les anaphores selon la nature lexico-cohésive, c'est-à-dire selon la nature des mots utilisés pour marquer le lien anaphorique. Une étude des moyens cohésifs a d'abord été faite par M. Halliday et R. Hasan [HH76] (pour l'anglais) et présentée et adaptée au français par R. Patry dans [Pat93]. Nous pouvons donc déterminer trois types d'anaphores[17] :

**Lexicale :** Les anaphores lexicales sont déterminées par les mots lexicaux (noms communs, adjectifs et verbes). Les éléments reliés par un lien lexical peuvent être de quatre types :

- réitération exacte du syntagme nominal[18] ;

  ***Le chat de la voisine***$_1$ *miaule à la fenêtre ...* ***Le chat de la voisine***$_{2a,anaph=1}$ *veut entrer dans la maison.* (9-a)

- réitération partielle du syntagme nominal ;

  *...* ***Le chat***$_{2b,anaph=1}$ *veut entrer dans la maison.* (9-b)

- utilisation d'un synonyme ;

  *...* ***Le minou***$_{2c,anaph=1}$ *veut entrer dans la maison.* (9-c)

- utilisation d'un hyperonyme.

  *...* ***Le félin***$_{2d,anaph=1}$ *veut entrer dans la maison.* (9-d)

**Pronominale :** Les anaphores pronominales sont déterminées par les mots grammaticaux (les pronoms personnels, démonstratifs et possessifs).

***Une femme***$_1$ *demande à Donna de* ***lui***$_{2,anaph=1}$ *créer une robe spécialement pour une séance de spiritisme.* (10)

**Syntaxique :** Les anaphores syntaxiques sont les liens anaphoriques qui relient des éléments coréférentiels à l'intérieur d'une même phrase (*sentence*). Ils s'identifient par les pronoms réfléchis et relatifs. Ces liens sont entièrement déterminés par la syntaxe.

*Ce dernier s'approche tout de même de Val et provoque maintenant la jalousie de* ***Cooper***$_1$ ***qui***$_{anaph=1}$ *décide de s'en aller.* (11-a)

[17] Nous avons adapté la terminologie en remplaçant le terme *cohésion* par *anaphore*. Nous savons que ces termes ne sont pas équivalents mais *anaphore* s'inscrit mieux dans notre étude.

[18] Ce lien peut parfois être considéré comme étant une anaphore de reprise.

*Jean*$_1$ *et Marie*$_2$ *parlent* ***l'un de l'autre***$_{3,anaph=1,2(reciproque)}$. (11-b)

Il y a beaucoup d'études théoriques qui ont été faites sur les facteurs syntaxiques qui déterminent la pronominalisation syntaxique en français (voir [Tel95]). Cependant, la plupart de ces études sont limitées à l'environnement syntaxique d'une seule phrase. Nous nous intéressons, dans ce mémoire, aux contraintes coréférentielles d'un texte en entier. Mis à part les pronoms relatifs, les types d'enchâssements des phrases comme (11-b) sont peu fréquents dans les textes rencontrés. Ces cas demandent aussi une analyse syntaxique complète de la phrase. Pour toutes ces considérations, nous n'inclurons pas les anaphores syntaxiques dans notre processus de résolution des anaphores.

### 2.2.2 Positions

Typiquement, les différents types de syntagmes nominaux n'occupent pas tous la même position à l'intérieur des chaînes coréférentielles. Les indéfinis, les définis autonomes et les noms propres sont à la tête de ces chaînes. Par contre, les définis de reprise, les démonstratifs, les pronoms et les ellipses sont à l'intérieur des chaînes.

Nous pouvons désormais revoir les différents types de syntagmes nominaux (voir la section 2.1 page 14), selon leur position à l'intérieur des chaînes coréférentielles[19].

**Indéfini :** Les syntagmes nominaux indéfinis sont presque toujours des têtes de chaînes coréférentielles et sont toujours indépendants du contexte. En ce sens, nous considérons que le groupe nominal indéfini «n'a pas de mémoire».

**Nom propre :** Les noms propres peuvent difficilement reprendre autre chose qu'un nom propre. Ils peuvent être à la tête d'une chaîne de coréférence. Dans ces cas, une description définie peut accompagner le nom propre.

> ***Abitibi-Consolidated, la papetière résultant de la fusion des compagnies Abitibi-Price et Stone Consolidated,*** *installera son siège social dans l'édifice Sun Life à Montréal*[20]. (12)

**Défini autonome :** Les syntagmes définis autonomes s'apparentent un peu aux noms propres en ce qui concerne l'indépendance et la position dans les chaînes coréférentielles. La distinction entre les deux types de syntagmes se situe au niveau de la rigidité du désignateur ; puisque les syntagmes définis autonomes peuvent varier selon le temps, le lieu, etc., ils ne sont pas des désignateurs rigides.

**Défini (non autonome) :** Les syntagmes définis non autonomes (de reprise, lexical ou associatif) peuvent occuper à peu près toutes les positions dans les chaînes coréférentielles. Cependant, ils nécessitent souvent la présence d'un autre élément linguistique dans le contexte pour fixer la référence. Ils peuvent donc plus difficilement être à la tête d'une chaîne de coréférence, sauf possiblement pour les cas génériques. Ils sont normalement les syntagmes les plus courants à l'intérieur des chaînes coréférentielles.

[19] Tiré de F. Corblin [Cor95].
[20] Première phrase d'un article tiré de [5].

*Le chat de la voisine*$_1$ *miaule à la fenêtre.* ***Le félin***$_{2b,anaph=1}$ *veut entrer dans la maison.* (13)

Il est à noter que la distinction entre le défini autonome et non autonome se base sur un jugement du lecteur humain (qui comprend le texte) dans le contexte global. Les jugements sont basés sur la position des syntagmes définis dans le texte et sur leur ordre relatif. Il est donc un peu circulaire de parler de leur position comme propriété indépendante. Dans la pratique, les syntagmes définis non autonomes ne comportent habituellement pas de modificateurs (contrairement aux syntagmes définis autonomes). Ainsi, les syntagmes définis «simples» (c'est-à-dire ne comportant qu'un nom et un déterminant) sont plus souvent des syntagmes définis non autonomes et les syntagmes définis «complexes» (comportant des modificateurs) sont plus souvent des syntagmes définis autonomes.

**Démonstratif :** Les syntagmes démonstratifs sont plutôt rares dans les textes dans les corpus étudiés. Ils ont une forte tendance à se placer à la deuxième position dans les chaînes de coréférence (exemple (14)). Ils indiquent dans ce cas une certaine réintroduction de l'objet ou suggèrent que l'objet est important. Les autres positions du syntagme démonstratif marquent une impression de rupture en indiquant un changement de point de vue (exemple (15)) ou servent à ramener à la mémoire un objet qui a pu être oublié. Le déterminant démonstratif est anaphorique car le recours au contexte est essentiel pour fixer la référence.

*La réalisatrice n'a en effet pu trouver mieux comme prétexte que de ramener Sylvia en France en compagnie d'***un nouveau mari américain (Ken Samuels)**$_1$ *et des deux grands fils de* ***ce dernier***$_{2,anaph=1}$ *; l'un blond, sportif et arrogant (Grégoire Lavollay-Porter) ; l'autre timide et gauche (James Thierrée)*[21]. (14)

*[...] Bien sûr, la donnée de base de* ***La Turbulence des fluides*** *repose sur une impossibilité physique. À preuve du contraire, la Lune exerce toujours son attraction, même dans le film de Manon Briand. Mais il est de ces licences poétiques qui peuvent se révéler fécondes. Le scénario de* ***cette Turbulence*** *qui repose justement sur une correspondance entre les lois de la nature et celles de l'âme, exige du spectateur ce consentement préalable à une entorse possible aux lois de Newton*[22]. (15)

**Pronom et ellipse :** Il va sans dire que ces groupes nominaux sont anaphoriques. Puisque leur contenu lexical est peu informatif, il faut aller chercher l'antécédent dans un contexte très proche.

## 2.3 Conclusion

De ce qu'on vient de voir, il en ressort clairement quelques axes sur lesquels il faut attacher de l'importance pour l'identification automatique des éléments référentiels d'un texte et

[21] Tiré de [6].
[22] Tiré de [7].

du regroupement de ces éléments en chaînes de coréférence. Ce sont essentiellement ces facteurs qui seront utilisés pour notre élaboration de critères formels permettant d'établir les liens coréférentiels et anaphoriques (voir section 6.2 page 65).

Nous considérons que les expressions référentielles sont les syntagmes nominaux (incluant les pronoms et les noms propres). Les expressions référentielles étant identifiées, l'étape suivante pour l'identification des chaînes coréférentielles d'un texte est le partitionnement de syntagmes référentiels en classes d'équivalences. Nous avons déterminé que ces chaînes peuvent être constituées d'éléments coréférentiels ou anaphoriques.

Certaines expressions référentielles sont plus susceptibles d'être les têtes des chaînes de coréférence : les expressions indéfinies, les expressions définies autonomes et les noms propres. Il peut toutefois arriver que deux éléments pouvant être des $\text{têtes}_c$ puissent être reliés dans une relation coréférentielle. Dans ce cas, la forme de l'expression est le meilleur indice pour établir ce lien.

Pour les éléments dépendants des chaînes (les expressions anaphoriques), les indices sur l'appartenance sont fournis principalement par les déterminants des expressions référentielles ; chacun indiquant plus ou moins une fenêtre contextuelle susceptible de contenir l'antécédent. La fenêtre contextuelle engendrée par un pronom ou un déterminant démonstratif devrait être, en principe, plus petite que celle engendrée par un déterminant défini.

# Chapitre 3

# Travaux précédents

Résoudre automatiquement les chaînes coréférentielles pose certains problèmes qui peuvent paraître insurmontables. Le solutionnement purement linguistique des coréférences nécessite des connaissances du monde, du domaine, de la situation d'énonciation ainsi qu'une analyse syntaxique et sémantique complète (voir section 1.1 page 3). D'un point de vue pratique, les modèles linguistiques élaborés demandent beaucoup d'efforts computationnels et manuels et donnent bien souvent de piètres résultats. Pourtant, dans la plupart des cas, le sujet parlant réussit à identifier les actants d'un texte et les liens coréférentiels sans même y penser.

Fort heureusement, il y a moyen de court-circuiter certains problèmes du solutionnement automatique des anaphores en adhérant à l'idéologie des stratégies *knowledge-poor*[1]. Les stratégies *knowledge-poor* misent sur le fait qu'il est possible de réduire la complexité des phénomènes linguistiques et utilisent des règles simples pour faire une approximation de la situation. Ces stratégies donneraient, dans certains cas, de meilleurs résultats que les méthodes linguistiques.

Dans ce chapitre, nous verrons comment les méthodes *knowledge-poor* peuvent être utilisées dans le traitement automatique de la langue. Nous le démontrerons à l'aide d'un exemple concret concernant le solutionnement des pronoms anaphoriques. Nous présenterons ensuite un bref historique de la résolution automatique des anaphores. Enfin, à partir de ce qui aura été observé dans le reste de ce chapitre, nous tenterons de déterminer les facteurs importants influençant les liens coréférentiels d'un texte.

## 3.1 Méthodes *knowledge-poor*

Normalement, l'établissement automatique des chaînes coréférentielles s'effectue en deux étapes :

1. l'identification des éléments référentiels d'un texte ;

[1] Nous ne connaissons pas l'équivalent français du terme *knowledge-poor*. Nous avons donc décidé d'utiliser le terme anglais.

2. le partitionnement de ces éléments en chaînes coréférentielles.

Le but ultime du traitement des coréférences est de déterminer tous les éléments référentiels d'un texte ; qu'ils soient explicites (directement dans le texte) ou implicites (par exemple, l'utilisation de l'expression *le fils* introduit implicitement le père ou la mère de ce fils). De plus, l'établissement des liens coréférentiels devrait inclure non seulement les liens d'identité entre les références, mais aussi toutes les autres relations possibles (les liens de possession, d'association, de partie-tout, etc.)

***Martin***$_1$ *aimerait bien retrouver* ***son chien***$_{2,anaph=1(possession)}$. (1-a)

*Il y a* ***une jolie fenêtre***$_1$ *dans le salon mais* ***les rideaux***$_{2,anaph=1(association)}$ *sont affreux.* (1-b)

*Il y a* ***une lampe***$_1$ *dans le salon mais* ***l'ampoule***$_{2,anaph=1(partie-tout)}$ *est brûlée.* (1-c)

Pourtant, pour accomplir certaines tâches du traitement automatique de la langue, la résolution parfaite et complète des liens anaphoriques n'est pas nécessaire. Par exemple, en recherche d'informations ou dans la création de résumés automatiques, il arrive que ce ne soit que les phrases les plus importantes du texte qui soient traitées. Et si nous pouvions faire un compromis : obtenir quelques-unes des chaînes coréférentielles (pas nécessairement complètes) en utilisant quelques règles lexicographiques et syntaxiques simples ?

Nous reparlerons plus en détail, dans le chapitre *Présentation et analyse des résultats* (section 6.2 page 65), des règles et des simplifications que nous avons choisies pour le solutionnement automatique des chaînes coréférentielles d'un texte. Toutefois, pour illustrer le cheminement possible pour construire une règle *knowledge-poor*, nous examinerons un cas particulier s'appliquant à notre recherche.

Par exemple, pour résoudre l'antécédent d'un pronom, nous pouvons essayer de choisir, parmi les antécédents possibles, un candidat «pas trop loin» et compatible en genre et en nombre. Ainsi, pour résoudre l'anaphore de la phrase suivante :

*Mais lorsque* ***Noah***$_1$ *voit le bracelet de* ***Val***$_2$, ***il***$_{3,anaph=1}$ *devient furieux.* (2)

nous pouvons utiliser une règle comme celle-ci :

RÈGLE 3.1 (PREMIER ANTÉCÉDENT COMPATIBLE)
*L'antécédent d'un pronom est le premier syntagme référentiel compatible en genre et en nombre situé avant ce pronom dans le texte (c'est-à-dire dans le sens inverse de la lecture)*[2].

Bien entendu, la règle 3.1 n'est pas linguistiquement correcte et donne parfois de faux résultats. Par exemple, en appliquant cette règle à une phrase comme (3), nous obtenons que l'antécédent de *il* est *Cooper*.

* *Mais lorsque* ***Noah***$_1$ *voit le bracelet que* ***Cooper***$_2$ *a offert à* ***Val***$_3$, ***il***$_{4,anaph=2?}$ *devient furieux.* (3)

[2]Une contrainte supplémentaire concernant la distance devrait être considérée. En effet, il est assez rare de retrouver l'antécédent à plus de deux phrases de distance de ce pronom. Nous reparlerons de ceci à la section 6.2 page 65.

Pour corriger l'erreur, nous pouvons transformer notre règle de sorte qu'elle privilégie les antécédents sujets des propositions non enchâssées :

RÈGLE 3.2 (PREMIER ANTÉCÉDENT SUJET COMPATIBLE)
*L'antécédent d'un pronom est le premier syntagme référentiel* ***sujet, non enchâssé,*** *compatible en genre et en nombre et situé avant ce pronom dans le texte. Si aucun antécédent n'est trouvé, la règle 3.1 est appliquée.*

Ainsi, nous obtenons :

*Mais lorsque* $\textbf{Noah}_1$ *voit le bracelet que* $\textbf{Cooper}_2$ *a offert à* $\textbf{Val}_3$, $\textbf{il}_{4,anaph=1}$ *devient furieux.* (4)

Évidemment, il est encore possible de trouver un contre-exemple pour contredire l'efficacité de la règle que nous venons de mettre au point. Nous pouvons, vaillamment, continuer notre petit jeu d'essais et erreurs jusqu'à l'obtention d'une règle acceptable. Cependant, il faut savoir quand et où s'arrêter.

* *Mais lorsque* $\textbf{Noah}_1$ ***reprend*** *le bracelet que* $\textbf{Cooper}_2$ *a offert à* $\textbf{Val}_3$, $\textbf{il}_{4,anaph=1?}$ *devient furieux.* (5)

Bien qu'il soit envisageable d'améliorer encore la règle 3.2, il faut déterminer si l'effort en vaut le coût. Il est probable que pour traiter les phrases comme (5), il soit nécessaire de considérer la nature du verbe et de faire une analyse sémantique de la phrase. Ceci complexifierait énormément le traitement automatique de la phrase (5). Dans un cas comme celui-ci, nous pouvons considérer que la règle 3.2 donne de bons résultats la plupart du temps et accepter que quelquefois, nous obtenions des mauvaises réponses[3].

Dans ce court exemple, nous avons essayé de saisir l'état d'esprit des méthodes *knowledge-poor*. Nous avons constaté que les règles doivent rester simples et être assez générales pour s'appliquer à un grand nombre de phénomènes. Autant que possible, les règles doivent aussi éviter le recours aux dictionnaires, aux analyses syntaxiques et sémantiques et aux connaissances du monde. Enfin, ces règles doivent donner des résultats acceptables. Les règles *knowledge-poor* sont donc, en quelque sorte, un compromis entre **simplicité, généralité** et **précision**.

Bien que les méthodes *knowledge-poor* soient utilisées pour le traitement automatique des textes anglais, la vérification de ces méthodes est peu expérimentée dans les textes français. Nous tenterons donc, dans le reste de ce mémoire, de valider et d'ajuster ces méthodes dans le but de les appliquer à un corpus constitué de textes français.

## 3.2 Historique

Dans cet historique, nous mettrons en relief les points qui semblent être importants pour notre travail, notamment la simplification du prétraitement et la réduction des connaissances

[3]Dans les textes étudiés, la construction anaphorique utilisée dans (5) est beaucoup moins fréquente que celle utilisée dans (4). C'est pourquoi, dans ce cas, l'obtention d'un mauvais solutionnement anaphorique n'est peut-être pas dramatique.

linguistiques nécessaires. Nous identifierons aussi quelques éléments pratiques qui nous aideront à développer notre propre algorithme. Notons toutefois que la plupart des méthodes présentées ici ont été développées pour résoudre les liens anaphoriques des textes anglais.

### 3.2.1 Années 80

En 1978, J. Hobbs propose un algorithme permettant de trouver l'antécédent des pronoms d'un texte. En prétraitement, l'algorithme suppose une analyse syntaxique de surface parfaite –y compris la désambiguïsation syntaxique.

La procédure consiste à parcourir la phrase, selon un ordre prédéterminé et à rechercher un antécédent compatible. À partir du noeud pronom à résoudre l'algorithme remonte (récursivement) dans l'arbre syntaxique vers un noeud syntagme nominal (l'arbre syntaxique est traversé selon un parcours de gauche à droite et en largeur d'abord) jusqu'à l'obtention d'une configuration compatible[4]. Implicitement, le parcours de l'arbre syntaxique favorise (dans l'ordre) les antécédents sujets, les antécédents objets puis les antécédents objets d'une préposition. Au besoin, l'antécédent peut être trouvé dans une phrase précédant la phrase courante. Ce procédé offre l'avantage de résoudre les anaphores syntaxiques dont les pronoms relatifs (*who*) et les pronoms réfléchis (*himself*). Les pronoms impersonnels et les pronoms référant à une proposition sont éliminés à l'avance.

Cet algorithme très populaire a été utilisé jusqu'au milieu des années 90. Cependant, il est assez difficile d'utiliser cet algorithme sur des textes réels car l'obtention d'une analyse syntaxique parfaite de façon automatique est encore aujourd'hui un problème de taille.

### 3.2.2 Début 90

Au milieu des années 90, la résolution des anaphores nécessite encore une analyse syntaxique complète et traite tous les pronoms, y compris les anaphores syntaxiques et les pronoms réfléchis. L'article le plus marquant de cette époque fut celui de S. Lappin et H. Leass [LL94].

Dans ce document, les anaphores syntaxiques sont résolues à l'aide de quelques règles tirées de la théorie du gouvernement liage et les pronoms réfléchis sont identifiés essentiellement grâce à la nature de certains verbes[5]. Pour ce qui est des autres pronoms, le repère de l'antécédent s'effectue grâce à des critères de compatibilité (genre, nombre et personne). Par contre, ce qui était implicite pour l'algorithme de Hobbs (l'avantage accordé à certaines fonctions syntaxiques) est devenu explicite.

L'algorithme procède ainsi :

1. Déterminer la liste des antécédents possibles et leur donner un pointage nul (par défaut).

[4]Les détails de l'algorithme se trouvent dans l'article [Hob78] de J.Hobbs.
[5]Nous laisserons de côté les détails de la résolution de ces pronoms.

2. Regrouper les antécédents en classes d'équivalence ; c'est-à-dire que les éléments partageant déjà un lien coréférentiel (trouvé à une itération ultérieure) sont classés dans un même sous-groupe.
3. Pondérer les sous-groupes d'antécédents possibles selon la fonction syntaxique, le niveau d'imbrication et la distance[6] de chacun des éléments qui le composent (voir Tableau 3.1). Le poids du groupe en entien est obtenu en additionnant le poids de chaque élément qui le compose.
4. Choisir le sous-groupe possédant le meilleur poids. Il sera l'antécédent du pronom.

Par exemple, un sous-groupe d'antécédents important (en nombre) situé près du pronom sera souvent privilégié par rapport aux autres sous-groupes d'antécédents possibles. De plus, un sous-groupe comportant des sujets sera aussi favorisé.

TAB. 3.1 – POIDS DES ANTÉCÉDENTS CANDIDATS (S. LAPPIN ET H. LEASS)

| |
|---|
| – L'antécédent est dans la phrase courante (+100). |
| – L'antécédent est un sujet (+80). |
| – L'antécédent est à l'intérieur d'une construction existentielle (+70). |
| – L'antécédent est un objet direct (+50). |
| – L'antécédent est un objet indirect (+40). |
| – L'antécédent n'est pas enchâssée dans un autre syntagme nominal (+80). |
| – L'antécédent n'est pas enchâssée dans un syntagme propositionnel adverbial (+50). |

### 3.2.3 Stratégies *knowledge-poor*

Vers la fin des années 90, un intérêt est porté sur la résolution d'anaphore exigeant peu d'informations syntaxiques et sémantiques. Les travaux de R. Mitkov [Mit98] sont parmi les plus connus. L'idée générale ressemble à ce qui a été présenté par [LL94]. Cependant, la pondération des antécédents est assez différente.

L'algorithme de Mitkov peut être résumé ainsi :

1. Étiqueter les parties du discours et les syntagmes nominaux du texte.
2. Lire le texte linéairement et traiter chaque expression anaphorique rencontrée.
3. Évaluer tous les candidats susceptibles d'être l'antécédent du pronom dans une fenêtre de trois phrases (précédant le syntagme à résoudre). Ces candidats reçoivent un pointage selon un certain nombre de critères préétablis (le Tableau 3.2 page suivante résume les critères favorisant le choix des antécédents[7]).
4. Choisir l'antécédent obtenant le meilleur pointage.

[6] Il y a un facteur d'éloignement qui divise le poids en deux pour chaque phrase de distance par rapport au pronom.
[7] Nous n'indiquons pas ici la pondération des critères.

TAB. 3.2 – CRITÈRES D'ANTÉCÉDENCE (R. MITKOV)

| |
|---|
| – L'antécédent est compatible en genre et en nombre. |
| – L'antécédent est un syntagme défini. |
| – L'antécédent est le thème de la phrase précédente. |
| – L'antécédent est un syntagme nominal suivant un verbe indicateur (*discuss, present, illustrate, identify, summarise, examine, describe, define, show, check, develop, review, report, outline, consider, investigate, explore, assess, analyse, synthesise, study, survey, deal, cover*). |
| – L'anaphore est une réitération lexicale de l'antécédent. |
| – L'anaphore est une expression qui reprend une partie du titre de la section. |
| – L'antécédent n'est pas un complément indirect (*insérer le disque dans le lecteur*). |
| – L'antécédent partage le même contexte que le pronom. |
| – L'anaphore est dans une construction intrinsèquement anaphorique. |
| – La distance entre l'anaphore et l'antécédent est petite. |
| – L'antécédent est un terme appartenant au sujet (domaine) du texte. |

### 3.2.4 *cogNIAC*

Les travaux de B. Baldwin [Bal97] tracent une nouvelle tendance dans le monde de la résolution des anaphores. La prémisse sur laquelle repose le raisonnement est qu'il existe une classe de pronoms dont le solutionnement ne nécessite pas de connaissances générales. Donc, en tentant de résoudre seulement quelques-unes des anaphores pronominales (les plus faciles), il est possible de simplifier le problème et d'obtenir une meilleure précision dans les résultats. En contrepartie, les cas jugés difficiles ne seront pas résolus (causant possiblement un moins bon rappel[8]).

L'algorithme, nommé *cogNIAC*, nécessite en prétraitement l'identification des phrases, des parties du discours et un simple balisage des syntagmes nominaux. Quelques traits syntaxiques et sémantiques tels le genre et le nombre sont aussi annotés. La procédure pour déterminer l'antécédent d'un pronom se déroule comme suit :

Pour tous les antécédents compatibles en traits sémantiques, il suffit d'évaluer leur pertinence selon la liste de critères du Tableau 3.3 page suivante (dans l'ordre). Si un pronom ne trouve pas d'antécédent répondant aux critères de sélection, il n'est tout simplement pas résolu.

> *Il*$_1$ *abandonne tout aussi aisément un job qu'une femme, sans vraiment comprendre les reproches qu'on lui fait.* *Il*$_{2,anaph=1}$ *est extraterrestre, c'est-à-dire qu'il*$_{3,anaph=1}$ *ne touche pas le terre-à-terre, ce qui explique probablement sa forte sensualité et son penchant charnel, qui sont sa forme de contact avec ces êtres dont il*$_{4,anaph=1}$ *ne comprend pas les aspirations*[9]. (6)

[8] Les notions de précision et de rappel seront introduites dans le chapitre *Méthodologie*, section 5.3 page 58.
[9] Tiré de [8].

TAB. 3.3 – CRITÈRES D'ANTÉCÉDENCE (COGNIAC)

1. S'il existe un seul antécédent possible, il est choisi.
2. Si le pronom est un pronom réfléchi, l'antécédent le plus proche (de la phrase courante) est choisi (par exemple : *Julie demande à* **Sophie**$_1$ *de porter* ***elle-même***$_{2,anaph=1}$ *son sac*).
3. S'il existe un seul antécédent possible dans la phrase courante ou précédente, il est choisi.
4. Si le pronom est possessif et s'il existe un antécédent (unique) qui a la même tête lexicale que le pronom à résoudre, il est choisi (par exemple : **le chat de la voisine**$_1$ ... **son chat**$_{2,anaph=1}$).
5. S'il existe un seul antécédent dans la phrase courante, il est choisi.
6. Si le sujet de la phrase précédente contient un seul antécédent et que l'anaphore est le sujet de la phrase courante, le sujet de la phrase précédente est choisi (voir (6)).

### 3.2.5 Chaînes importantes d'un texte

Dans l'article de S. Bergler [Ber97], on souligne que la résolution de toutes les anaphores demande une très grande compréhension de texte. Ce niveau de compréhension n'est pas encore atteint avec les méthodes actuelles du traitement automatique de la langue. Pour contrer cette lacune, il semble que la résolution doit être partielle. Miser sur les particularités des textes selon le domaine semble être un bon moyen d'augmenter la fiabilité de la résolution des anaphores.

Dans cet article, une étude faite sur des textes du *Wall Street Journal* relate quelques constatations :

- Les chaînes appartenant au sujet du texte se comportent différemment des autres chaînes du texte. Ces chaînes sont possiblement plus longues et possèdent aussi un vocabulaire beaucoup plus restreint. Elles ont pour tête$_c$ (la première occurrence dans le texte) un élément situé dans les premières phrases du texte.
- L'identification des syntagmes nominaux qui se ressemblent est importante pour la résolution des anaphores (réitération lexicale). Les syntagmes partageant un même nom propre sont les cas les plus fréquents et les plus simples à résoudre.

### 3.2.6 Réduction du prétraitement

Bien que le prétraitement nécessaire au traitement automatique des anaphores a grandement été simplifié au cours des années, il reste toutefois quelques difficultés : l'identification des fonctions syntaxiques et des traits syntaxiques et sémantiques. Déterminer la fonction syntaxique d'un syntagme référentiel nécessite une analyse syntaxique. Aussi, pour déterminer certains traits syntaxiques et sémantiques, il est nécessaire de consulter un dictionnaire.

L'article de A. Siddharthan [Sid03] propose plutôt d'utiliser les informations déjà présentes dans le texte pour déduire les fonctions syntaxiques ainsi que les traits syntaxiques et sémantiques :

- Il est possible de déterminer les fonctions syntaxiques de la majorité des éléments référentiels d'un texte à l'aide d'expressions régulières[10]. Les prépositions et les positions des syntagmes face au verbe sont deux bons indices pour déterminer si une expression référentielle est un sujet ou un objet.
- Bien souvent, les traits syntaxiques et sémantiques peuvent être tout simplement déduits ; soit à l'aide des autres éléments du texte (déterminants, accord), des chaînes déjà identifiées, de WordNet ou à l'aide de la sous-catégorisation des verbes (par exemple, *x said* implique que $x$ est un animé).

L'établissement des chaînes coréférentielles peut ainsi bénéficier d'une grande simplification du travail de prétraitement syntaxique et donner des résultats tout aussi robustes.

## 3.3 Conclusion

Nous avons passé en revue quelques algorithmes menant à la résolution des anaphores pronominales et à l'identification des chaînes coréférentielles d'un texte. Nous avons remarqué une tendance, au fil des années, à réduire le nombre d'éléments référentiels à traiter, à diminuer le prétraitement et à simplifier l'analyse syntaxique nécessaire. Ainsi, une méthode donnant des résultats ayant une bonne précision (obtenir le plus possible que de bons résultats), moyennant une couverture un peu moins bonne (par exemple, ne pas traiter les cas difficiles) semble être le but visé.

En analysant les critères de choix des antécédents de chacun des algorithmes, nous pouvons dégager quelques critères qui semblent être communs à plusieurs d'entre eux :

- La compatibilité de certains traits syntaxiques (genre, nombre, personne) et sémantiques (animation) entre les antécédents et les anaphores est très importante.
- Les sujets sont les antécédents les plus fréquents. Viennent ensuite les compléments d'objet direct, les compléments d'objet indirect et les autres compléments.
- La distance entre l'antécédent et l'anaphore n'est jamais très grande ; au plus quelques phrases. Certaines conditions permettent aussi à l'antécédent et au pronom d'être dans la même phrase.
- Les réitérations lexicales sont de bons indicateurs de liens coréférentiels.
- Les référents associés au domaine se comportent de façon différente des autres éléments référentiels du texte. En particulier, le vocabulaire utilisé pour désigner ces entités est beaucoup plus prévisible que pour les autres entités référentielles et est relativement stable

[10] Les expressions régulières sont des outils permettant de manipuler des chaînes de caractères (ou les classes de chaînes de caractères). Par exemple, *un syntagme référentiel devancé par un mot appartenant à la classe des prépositions* peut être exprimé à l'aide d'une expression régulière. Nous reverrons quelques exemples d'expressions régulières aux sections 4.2.3 page 45 et 6.2 page 65.

pour un type de texte donné. Ces référents semblent aussi plus faciles à résoudre et sont les constituants des chaînes coréférentielles les plus longues du texte.

Pour établir les chaînes coréférentielles d'un texte, il semblerait qu'un algorithme utilisant une heuristique basée sur une pondération des facteurs ci-dessus puisse être envisageable. Nous tenterons d'exploiter cette voie au chapitre *Présentation et analyse des résultats* (voir chapitre 6 page 62).

# Chapitre 4

# Outils *XML*

> «WHAT IS RECURSION ? *[...] The concept is very general. (Stories inside stories, movies inside movies, paintings inside painting, Russian dolls inside Russian dolls (even parenthetical comment inside parenthetical comment !)–these are just a few of the charms of recursion*[1]*.)»*
>
> ***– Douglas R. Hofstadter***

L'étude des phénomènes linguistiques s'effectue difficilement sans la consultation des productions écrites de la langue. Bien que la construction d'un corpus soit une étape essentielle de notre recherche, il importe de pouvoir avoir les instruments pour observer, analyser et interroger ce corpus. Un outil souvent utilisé pour une telle tâche est le concordancier car il permet de dégager les occurrences en contexte de certains mots ou de certaines chaînes de mots d'un corpus.

Cependant, l'usage d'un concordancier ne semble pas tout à fait être l'outil idéal dans notre cas. Puisque nous étudions des manifestations du langage aux niveaux de la syntaxe et du texte, le contexte des mots est nettement insuffisant pour bien cerner les dépendances entre des réalisations linguistiques pouvant être distancées de plusieurs phrases. Par exemple, pour étudier le choix du type de déterminant utilisé selon la position du syntagme dans une chaîne coréférentielle, il est habituellement nécessaire de parcourir tout le texte. Pour analyser la langue à ces niveaux structurels, il nous semble que l'utilisation des outils *XML* soit une meilleure approche.

Pour les besoins de notre étude, nous utiliserons *XML* pour baliser manuellement les phénomènes linguistiques coréférentiels. Ensuite, nous utiliserons quelques outils de la famille *XML* pour faciliter diverses études sur le corpus balisé. Les besoins d'analyse de notre étude sont essentiellement de trois natures :

[1] Tiré de [9].

1. Afficher le texte en mettant en relief les éléments importants ; ce peut être un code de couleur, encadrer les syntagmes nominaux, souligner les têtes des syntagmes, etc.
2. Traiter le texte (ou le corpus) comme une base de données et effectuer des requêtes sur l'interaction de certains phénomènes linguistiques. Par exemple :
   (a) Obtenir une liste de tous les syntagmes nominaux contenant un nom propre et un déterminant.
   (b) Déterminer quels sont les syntagmes anaphoriques ayant un syntagme défini comme antécédent.
   (c) Compter la proportion des pronoms personnels par rapport à tous les pronoms d'un texte.
3. Réorganiser un document pour observer les phénomènes ne se présentant pas nécessairement dans l'ordre linéaire du texte.

Dans ce chapitre, nous expliquerons brièvement le fonctionnement de *XML* et de quelques autres outils reliés à ce métalangage (*XHTML*, *CSS*, *XPath* et *XSLT*). Bien que la richesse de ces outils soit très grande, nous n'exposerons que les grandes lignes du fonctionnement de ces langages pour avoir une idée générale de leur potentiel.

## 4.1 *XML*

*XML* (Extensible Markup Language) est un langage de balisage permettant, dans un sens très large, de structurer des documents et des données[2].

Un exemple très simple de langage de balisage, peut-être plus naturel à imaginer, est l'utilisation des signes de ponctuation dans un texte. Le point permet de séparer deux phrases, les virgules permettent de séparer des propositions, etc. De plus, les parenthèses (comme celle-ci)[3] permettent de changer d'optique et de faire des remarques ou des ajouts à propos du contenu textuel situé juste avant. Cependant, ce «langage de balisage» manque quelque peu de formalisme pour un traitement informatique.

La syntaxe de base de *XML*, qui peut sembler un peu moins naturelle que ne l'est la ponctuation, est toutefois très simple :

$$<\textit{entité}\ \textit{attribut}_1 = "\textit{valeur}_1"\ \ \textit{attribut}_2 = "\textit{valeur}_2"\ \ \ldots\ > \textit{données} </\textit{entité}> \quad (1)$$

où la variable *entité* est le nom de la balise, et le couple $\textit{attribut}_i$ et $\textit{valeur}_i$ sont des informations supplémentaires concernant *données*[4]. Notons que *données* peut être à son tour subdivisé en sous-données et contenir d'autres informations balisées. Il est important de constater qu'un document *XML* peut être représenté par un arbre (d'où l'appelation arborescence *XML*).

La nature première de *XML* est de structurer des éléments. Par exemple, *XML* est idéal pour stocker les entrées d'une base de données lexicographique (un dictionnaire)[5] :

[2] Les définitions de cette section sont tirée de [Coy02].
[3] ... ou les notes de bas de page.
[4] Il peut y avoir un nombre arbitraire d'attributs, y compris aucun.
[5] Extrait de la base de donnée lexicographique que nous avons construite et utilisée pour le traitement informatique

```
<vocabulaire>
   <déterminants>
      <det type="défini" genre="m">le</det>
      <det type="défini">l'</det>
      <det type="défini" genre="f">la</det>
      <det type="défini" nb="plur">les</det>
      <det type="demonstratif" genre="m">ce</det>
      <det type="demonstratif">c'</det>

                              ...

   </déterminants>
   <pronoms>
      <pro type="personnel" personne="1">je</pro>
      <pro type="personnel" personne="1">j'</pro>
      <pro type="personnel" personne="2">tu</pro>
      <pro type="personnel" genPer="m" personne="3">il</pro>
      <pro type="personnel" genPer="f" personne="3">elle</pro>

                              ...

   </pronoms>
</vocabulaire>
```
(2)

Au-delà de la base de données, *XML* peut aussi servir à baliser un texte dans le but d'en faire un traitement informatique par la suite. De par sa nature récursive, *XML* est tout à fait adapté pour baliser les structures d'arbre syntaxique.

Illustrons l'emploi de *XML* à l'aide d'un exemple appliqué à notre recherche :

> *Lors d'une escapade un peu longue de la mère, Julien se voit obligé de traîner Elsa avec lui dans une excursion à la montagne*[6]. (3)

Pour les besoins de notre étude, nous désirons baliser les informations suivantes :

1. Déterminer que *Julien* et *Elsa* sont des noms propres, que *mère* est un nom commun, et *lui* un pronom et obtenir de l'information à propos des types de déterminants utilisés (le *la* de *la mère*).
2. Identifier les expressions référentielles du texte (*la mère*, *Julien*, *lui* et *Elsa*) ainsi que leurs fonctions syntaxiques.
3. Établir le lien anaphorique entre *Julien* et *lui*.

Nous verrons, dans le reste de cette section, comment il est possible d'effectuer ces tâches en utilisant *XML*.

des anaphores de ce travail (voir annexe B page iv pour la liste complète).

[6] Tiré de [10].

### 4.1.1 Classification des mots

Identifier que *Julien* et *Elsa* sont des noms propres et que *mère* est un nom commun est assez simple :

```
<npr>Julien</npr>
<npr>Elsa</npr>
<nom>mère</nom>
```
(4)

Le même processus pourrait convenir pour *lui* et *la* :

```
<pro>lui</pro>
<det>la</det>
```
(5)

Cependant, nous aimerions avoir un peu plus d'informations à propos des déterminants utilisés. En effet, nous avons écrit au chapitre *Notions théoriques* (section 2.1.2 page 17) que le type de déterminant et de pronom (indéfini, défini, démonstratif, etc.) a une certaine influence sur la référence. Pour annoter ce renseignement, nous pourrions faire comme suit :

```
<pro_personnel>lui</pro_personnel>
<det_défini>la</det_défini>
```
(6)

Bien que cette notation soit tout à fait valide, on remarque toutefois qu'elle est plutôt lourde. Ainsi, pour indiquer que *lui* est un pronom personnel à la troisième personne du singulier donnerait :

```
<pro_personnel_troisième_pers_sing>
    lui
</pro_personnel_troisième_pers_sing>
```
(7)

Un moyen plus simple et plus économique pour ajouter des précisions à propos des données encadrées est l'utilisation des attributs :

```
<pro type="personnel" personne="3" nb="sing">lui</pro>
<det type="défini">la</det>
```
(8)

Cette façon de faire offre aussi l'avantage d'identifier le pronom *eux* de façon similaire au pronom *lui* :

```
<pro type="personnel" personne="3" nb="plur">eux</pro>
```
(9)

Notre exemple (3) page précédente devient donc :

*Lors d'une escapade un peu longue de*
<det type="defini">*la*</det> <nom>*mère*</nom>,
<npr>*Julien*</npr> *se voit obligé de traîner*
<npr>*Elsa*</npr> *avec*
<pro type="personnel" personne="3" nb="sing">*lui*</pro>
*dans une excursion à la montagne.*
(10)

Il est à noter que les balises sont insérées dans le texte, laissant un mélange de texte et de balises. Si nous détruisons tout ce qui se trouve entre les «<» et les «>», nous retrouvons le texte original. Ceci restera vrai pour tous les exemples de ce chapitre et même pour la schématisation finale utilisée dans le traitement informatique des chaînes coréférentielles.

### 4.1.2 Identification des syntagmes référentiels

Lors de notre étude des syntagmes référentiels (chapitre 2 section 2.1 page 14), nous avons convenu que les têtes des syntagmes nominaux peuvent être des noms communs, des pronoms ou des noms propres. De plus, les syntagmes nominaux peuvent aussi contenir des déterminants et des compléments. Pour transposer ceci dans le métalangage *XML*, nous ajoutons des balises <sn> et </sn> pour encadrer les syntagmes nominaux. Ensuite, nous ajoutons en attribut la fonction syntaxique des entités.

*Lors d'une escapade un peu longue de*
<sn **fonction="cnom"**>
<det type="défini">*la*</det>
<nom>*mère*</nom>
</sn>,
<sn **fonction="sujet"**>
<npr>*Julien*</npr>
</sn> *se voit obligé de traîner*
<sn **fonction="cod"**>
<npr>*Elsa*</npr>
</sn> *avec*
<sn **fonction="cc"**>
<pro type="personnel" personne="3" nb="sing">*lui*</pro>
</sn> *dans une excursion à la montagne.* (11)

Ici, l'indentation de la phrase n'est nullement significative, elle ne sert qu'à aider à la lisibilité du texte.

### 4.1.3 Établissement des liens anaphoriques et coréférentiels

Maintenant que nous savons comment étiqueter les éléments référentiels d'un texte, nous pouvons désormais résoudre les liens anaphoriques et coréférentiels entre les références. Nous illustrons ici le procédé à l'aide du pronom anaphorique *lui* :

**<anaph>**
<sn fonction="cc">
<pro type="personnel" personne="3" nb="sing">*lui*</pro>
</sn>
**</anaph>** (12)

Le lien entre *Julien* et *lui* peut être marqué en nommant de façon unique le syntagme contenant *Julien* et en lui faisant référence dans l'anaphore pronominale *lui*. Puisque, a priori, tous les syntagmes sont susceptibles d'être des antécédents, nous donnerons à chacun des numéros

d'identification (`id`) uniques[7]. Par la suite, nous créons la liaison anaphorique en introduisant dans `<anaph>` l'attribut `idref` et en lui donnant la valeur de l'`id` de son antécédent[8] :

*Lors d'une escapade un peu longue de*
`<sn id="r1" fonction="cnom">`
`<det type="défini">`*la*`</det>`
`<nom>`*mère*`</nom>`
`</sn>`,
`<sn id="r2" fonction="sujet">`
`<npr>`*Julien*`</npr>`
`</sn>` *se voit obligé de traîner*
`<sn id="r3" fonction="cod">`
`<npr>`*Elsa*`</npr>`
`</sn>` *avec*
`<anaph idref="r2">`
`<sn id="r4" fonction="cc">`
`<pro type="personnel" personne="3" nb="sing">`*lui*`</pro>`
`</sn>`
`</anaph>`
*dans une excursion à la montagne.* (13)

De la même façon, nous pouvons identifier les mots coréférentiels en remplaçant la balise `<anaph>` par `<coref>` :

*L'an dernier, les ennuis ont commencé pour* `<sn id="r1" fonction="coi">` `<npr>`*ABB*`</npr>` `</sn>`*. Sous la pression de l' investisseur suisse Martin Ebner, qui avait rassemblé une participation de 11% et avait joint les rangs du conseil d'administration de la compagnie en 1999,* `<coref idref="r1">` `<sn id="r2" fonction="sujet">` `<npr>`*ABB*`</npr>` `</sn>` `</coref>` *racheta environ 2% de ses actions pour faire remonter le titre, privant la compagnie de liquidités au moment où la demande commençait à fléchir*[9]. (14)

Notons que dans cet exemple, nous n'avons balisé que les syntagmes référentiels pertinents à l'exemple. Précisons qu'à partir de maintenant, pour alléger la présentation des exemples choisis, nous n'identifions que les phénomènes qui représentent un certain intérêt. Ainsi, nous ferons l'économie de n'inscrire que les éléments et les attributs importants dans le contexte de présentation. Par exemple, le pronom *lui* se notera habituellement `<pro>`lui`</pro>` plutôt que `<pro type="personnel" personne="3" nb="sing">`lui`</pro>`.

[7] Ces numéros d'identification ne sont pas nécessairement séquentiels dans le texte. En pratique, ils sont habituellement générés automatiquement et peuvent avoir des valeurs aussi peu évocatrices que *d0e216*. L'important est qu'ils doivent être uniques à l'intérieur du texte.

[8] *la mère* est probablement anaphorique à un autre élément du texte, ce lien serait identifié de façon analogue à *lui*.

[9] Tiré de [11].

### 4.1.4 Autres possibilités

Dans l'exemple présenté dans ce chapitre, nous avons pu remarquer que la façon de baliser un texte est plutôt permissive. Puisque *XML* est un métalangage (et non un langage comme le *HTML*), nous avons une grande latitude pour le choix des étiquettes. Par exemple, nous aurions pu donner d'autres noms aux balises choisies ou même structurer les balises d'une autre façon (tout en conservant l'ordre linéaire des mots de la phrase) :

*Lors d'une escapade un peu longue de*
<sn id="r1" fonction="cnom" type"defini">
*la*
<nom>*mère*</nom>
</sn>,
<snpr id="r2" fonction="sujet">
*Julien*
</snpr>
<rien></rien>
<encore_rien></encore_rien>
*se voit obligé de traîner*
<snpr id="r3" fonction="cod">
*Elsa*
</snpr> *avec*
<pro type="personnel" personne="3" nb="sing">
<sn id="r4" fonction="cc">
<anapn>
<caract>*l*</caract>
<caract>*u*</caract>
<caract>*i*</caract>
</anapn>
</sn>
</pro> *dans une excursion à la montagne.* (15)

Bien que ce dernier exemple soit tout à fait valide, il est beaucoup plus difficile à lire que l'exemple (13) page précédente et certainement plus difficile à manipuler. De façon générale, nous devons observer quelques principes :

- S'assurer de repérer les occurrences qui représentent un certain intérêt pour notre étude.
- Ne baliser que les éléments pouvant être utiles.
- Donner des noms significatifs aux entités, aux attributs et aux valeurs des attributs.
- Respecter les structures linguistiques (lexical, syntaxique, textuel). Par exemple, les balises lexicales sont plus près du mot et sont donc incluses à l'intérieur des balises syntaxiques.
- Être consistant dans le balisage.

### 4.1.5 Schématisation finale

Jusqu'ici, dans notre petite introduction à *XML*, nous avons élaboré une stratégie permettant de baliser les références d'un texte. C'est essentiellement cette même schématisation que nous utiliserons pour les textes de notre corpus[10]. Un très court[11] exemple balisé d'un des textes de notre corpus est en annexe (annexe D page xx). Notons seulement quelques ajouts et points techniques :

- Nous étendons l'identification de la catégorie lexicale aux autres catégories possibles (au besoin), soit les verbes, les adverbes, les adjectifs, les prépositions, les pronoms relatifs et les conjonctions (<verbe>, <adv>, <adj>, <prep>, <rel> et <conj> respectivement). De plus, nous identifions les signes de ponctuation (<ponc>).
- Certaines informations sémantiques ou syntaxiques (le genre, le nombre, la personne, etc.) peuvent être ajoutées (au besoin) à l'aide des attributs (genre, nb, personne, etc.)
- Les anaphores et les coréférences sont étiquetées de façon similaire. Nous avons déjà dit au chapitre *Notions théoriques* (section 2.2.1 page 20) que les relations coréférentielles sont symétriques alors que les anaphores sont asymétriques (directionnelles). Nous identifions cependant ces deux phénomènes de la même façon tout en gardant ces différences à l'esprit lors du traitement.
- Si un élément linguistique est anaphorique ou coréférentiel à un ensemble d'antécédents, l'utilisation de l'attribut idrefs peut contenir la liste des antécédents (séparés par des espaces) et remplacer l'attribut idref :

  > *Des millions de spectateurs de par le monde ont posé un regard attendri sur la petite Marie lorsque Sylvia, la mère de cette dernière, bien désespérée à l'époque, l'avait déposée sur le seuil de la porte d'entrée de l'appartement qu'habitait* <sn id="r1">*le papa géniteur*</sn> *avec* <sn id="r2">*deux potes*</sn>.
  >
  > ...
  >
  > <anaph idrefs="r1 r2"><sn>*Les trois pères*</sn></anaph> *ne peuvent ainsi résister à l'envie d'aller rejoindre leur fille qui, après avoir passé son bac, est partie en Provence en compagnie de sa mère et de la nouvelle famille( ajoutez une gouvernante interprétée par Line Renaud)*[12]. (16)

- Nous identifions les niveaux organisationnels d'un texte dépassant celui de la phrase (tels les paragraphes (<par>), les titres et les sous-titres (<titre>), les bouts de code (<bloccode>)[13] et les parties non traitées (<nontraite>)).
- L'antécédent d'une anaphore peut aussi être un syntagme verbal ou une proposition. Dans ces cas, nous déterminons l'antécédent, de façon similaire à ce que nous avons fait pour les syntagmes nominaux, en attribuant à l'anaphore le numéro d'identification du syntagme verbal ou de la proposition :

[10] Notre schématisation a grandement été inspirée par les travaux réalisés dans le cadre des conférences *MUC* (voir [HC97]).
[11] Oui, *XML* est très verbieux. Un texte de trois phrases peut nécessiter presque trois pages d'annotations *XML* !
[12] Tiré de [6].
[13] Dans les *HOWTO Linux*.

*Contre toute attente (enfin...), le jeune Mays Gilliam saura lentement, par son attitude désinvolte, sa franchise et ses opinions populistes,* <sv id="r1"> *gagner le coeur des Américains* </sv>.
<anaph idref="r1" fonction="sujet"> <sn id="r2"> <pro type="demonstratif">*Ce*</pro> </sn> </anaph> *qui n'était pas au programme*[14]... (17)

– Pour respecter les normes *XML*, il est nécessaire d'avoir un préambule conforme aux standards du World Wide Web Consortium (*W3C*). Sans entrer dans les détails, en voici un exemple :

```
<?xml version="1.0" encoding="iso-8859-1"?>
<?xml-stylesheet type="text xsl"
   href="../../Transformations/chaines.xsl"?>
<chaines domaine="Fusion"
   xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
   xsi:noNamespaceSchemaLocation=
     "file:../../Schemas/chainesCoref.xsd">
```

*Le texte*

```
</chaines>
```
(18)

## 4.2 Autres outils *XML*

À ce point-ci, nous pouvons nous demander en quoi *XML* peut nous être utile. S'il n'est qu'une façon de structurer les parties d'un texte et que nous choisissons nous-mêmes l'organisation et les éléments à baliser, comment *XML* peut-il être meilleur qu'un concordancier ?

*XML*, utilisé seul, sert essentiellement à mettre en évidence la structure des constructions linguistiques, nous obligeant à concevoir la schématisation de la structure cadrant le phénomène[15]. Cependant, la grande force de *XML* provient principalement des outils facilitant le traitement des fichiers *XML*.

Nous parlerons brièvement de quelques-uns de ces outils dans cette section.

### 4.2.1 *XHTML*

Le *XHTML* (Extensible HyperText Markup Language) est une reformulation du *HTML* (HyperText Markup Language) en un sous-ensemble de *XML*. Les éléments et les attributs du *XHTML* sont prédéfinis et peuvent être visualisés par un fureteur Web. Voici un exemple simple de fichier *XHTML*. Notons que nous y reconnaissons la syntaxe de *XML*.

[14] Tiré de [12].

[15] Il existe plusieurs outils permettant de formaliser cette schématisation : *DTD* (Document Type Definition), *XML Schemas* et plusieurs autres. L'utilisation de ces outils permet la validation d'un document *XML* selon des schémas prédéfinis, impose la consistance du balisage et prévient les fautes de frappe de l'étiquetage manuel.

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Strict//EN"
   "http://www.w3.org/TR/xhtml1/DTD/xhtml1-strict.dtd">
<html>
   <head>
      <title>Titre</title>
   </head>
   <body>
      <h1>Titre du document</h1>
      <p>Ceci est le texte d'un paragraphe.</p>
   </body>
</html>
```
(19)

Dans cet exemple, le texte est situé entre les balises `<body>`, le titre entre les balises `<h1>` et les paragraphes entre les balises `<p>`.

### 4.2.2 *CSS*

*CSS* (Cascading Style Sheets) est un mécanisme pour ajouter du style à des documents Web. Il est donc intimement lié au *(X)HTML* et permet de régler certains détails présentationnels. Par exemple, pour afficher le titre du document (`<h1>`) en orange et justifier les paragraphes (`<p>`) à 20 pixels de la gauche, nous ajoutons les lignes suivantes au fichier *XHTML* (19)[16]. :

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Strict//EN"
   "http://www.w3.org/TR/xhtml1/DTD/xhtml1-strict.dtd">
<html>
   <head>
      <title>Titre</title>
      <style type="text/css">
         h1 {color: orange}
         p {margin-left: 20px}
      </style>
   </head>
   <body>
      <h1>Titre du document</h1>
      <p>Ceci est le texte d'un paragraphe.</p>
   </body>
</html>
```
(20)

*CSS* peut être utilisé avec les balises existantes du *XHTML* mais peut aussi être utilisé dans des fichiers *XML* pour des balises que nous avons créées. Donc, pour afficher les syntagmes

[16]Pour les cas plus complexes, il est possible d'inclure une feuille de style externe contenue dans un fichier séparé.

nominaux du texte dans un cadre vert, il suffit d'ajouter :

```
sn {border-style:groove; border-color:green}
```
(21)

à l'intérieur de l'élément `<style>`.

Dans notre projet, nous avons utilisé *CSS* pour présenter, colorer et attirer l'oeil sur les champs d'intérêt que nous voulions créer (comme la frontière des syntagmes référentiels ou les types de déterminants). Il est à noter que puisque cette information tient en quelques lignes seulement, il est facile de changer la couleur ou la présentation des éléments en cours d'analyse.

### 4.2.3 *XPath*

*XPath* (*XML* Path Language) est un ensemble de règles syntaxiques servant à déterminer des parties de document *XML*. Nous utilisons *XPath* comme un langage de requêtes pour extraire les fragments de *XML* correspondants à une expression régulière. Par exemple :

1. L'ensemble de tous les syntagmes nominaux contenant un nom propre et contenant un déterminant s'écrit ainsi :

   ```
   //sn[det and npr]
   ```
   (22)

2. La liste des syntagmes anaphoriques ayant un syntagme indéfini comme antécédent est :

   ```
   //anaph[@idref = //sn[det/@type="indéfini"]/@id]
   ```
   (23)

3. La proportion des pronoms personnels par rapport a tous les pronoms s'obtient par :

   ```
   count( //pro[@type="personnel"]) div count( //pro)
   ```
   (24)

La syntaxe peut sembler quelque peu compliquée mais elle permet de produire des expressions assez complexes. Pour aider à la lecture des expressions *XPath*, voici quelques règles :

- À l'image de la navigation de l'arborescence des répertoires d'un ordinateur, les «/» servent à parcourir le fichier *XML* selon un chemin déterminé[17].
- `//` veut dire *n'importe quel*.
- Le `@` indique un attribut.
- Les crochets carrés peuvent être lus en les remplaçant par *tel que* ...

Ainsi, l'expression (23) peut être traduite par : *n'importe quel* `anaph` *tel que l'attribut* `idref` *est égal à l'attribut* `id` *de n'importe quel* `sn` *tel que [ce* `sn`*] contient un* `det` *et que [ce* `det`*] possède un attribut* `indéfini`. Ouf !

*XPath* peut être utilisé par l'intermédiaire d'un interpréteur *XPath* ; certains éditeurs de *XML* offrent d'ailleurs cette fonctionnalité à même leur interface. Il suffit de taper la requête dans le champ *XPATH* et l'interpréteur retourne les noeuds *XML* correspondant à la demande. *XPath* peut aussi être utilisé à l'intérieur d'un autre langage, comme *XSLT*, que nous présenterons à l'instant.

[17] On peut parcourir l'arborescence des répertoires sous *UNIX* à l'aide du séparateur «/» (le fichier `/home/boudreau/mémoire/mémoire.ps`) ou du séparateur «\» sous *DOS*.

### 4.2.4 *XSLT*

*XSLT* (Extensible Stylesheet Language Transformation) est un langage pour transformer des documents *XML*. Il permet de présenter des fichiers *XML* en *XHTML*, d'envelopper les requêtes *XPath* et de réorganiser la structure du document.

Par exemple, pour afficher tous les liens coréférentiels d'un texte (sous forme de liste), de sorte que les liens de dépendance entre les antécédents et les coréférences puissent être visibles, nous appliquons la transformation *XSLT* suivante au document *XML* de notre corpus[18] :

[18] La transformation peut être appliquée par certains fureteurs ou par des programmes tels *Xalan* ou *Saxon*.

```
<?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>
            <h1>Les chaînes coréférentielles</h1>
            <ol>
               <!--Pour toutes les têtes de chaînes
                        coréférentielles.-->
               <xsl:for-each select="//node()[name() != 'anaph'
                  and name() != 'coref']/sn">
                  <xsl:call-template name="teteCoref"/>
               </xsl:for-each>
            </ol>
         </body>
      </html>
   </xsl:template>

   <!--Dessine (récursivement) le noeud vedette et ses
   (sous-)chaînes anaphoriques.-->
   <xsl:template name="teteCoref">
      <xsl:variable name="idSn" select="@id"/>
      <li>
         <!--Dessine le noeud vedette -->
         <xsl:value-of select="."/>
      </li>
      <!--Si le syntagme courant est l'antécédent d'un autre sn.-->
      <xsl:if test="(//anaph | //coref)/@idref = $idSn">
         <ul>
            <!--Appel récursif de la fonction courante.-->
            <xsl:for-each select="(//anaph | //coref)
               [@idref = $idSn]/sn">
               <xsl:call-template name="sn"/>
            </xsl:for-each>
         </ul>
      </xsl:if>
   </xsl:template>

</xsl:stylesheet>
```

(25)

Dans cet exemple, `<xsl:stylesheet>`, `<h1>`, `<ol>`[19] et `<li>` sont des balises de *XHTML* ; ce qui se situe entre les balises `<!--` et `-->` sont des commentaires ; les valeurs des attributs `select` et `test` (en gras) sont des requêtes *XPath* et les `<xsl:template>` sont des fonctions.

Le résultat de cette transformation appliquée au texte balisé (26) (l'exemple du chapitre 1) donne le résultat (27).

> ***Une femme***$_1$ *demande à* ***Donna***$_2$ *de* ***lui***$_{3,anaph=1}$ *créer une robe spécialement pour une séance de spiritisme.* ***Celle-ci***$_{4,anaph=1}$ *voudrait entrer en communication avec* ***son défunt mari***$_{5,anaph=1(possession)}$ *afin qu'****il***$_{6,anaph=3}$ ***l'***$_{7,anaph=4}$ *aide à retrouver son bracelet de diamants.* ***Steve***$_8$ *demande à* ***Carly***$_9$ *de* ***l'***$_{10,anaph=8}$*accompagner au bal de la moisson et* ***Cooper***$_{11}$ *en fait de même avec* ***Val***$_{12}$*. La situation financière de* ***David***$_{13}$ *ne s'arrange pas. On* ***lui***$_{14,anaph=13}$ *refuse même une demande de crédit. Au bal,* ***Brandon***$_{15}$*,* ***Kelly***$_{16}$*,* ***Steve***$_{17,coref=8}$ *et* ***Carly***$_{18,coref=9}$ *sont surpris de voir* ***Val***$_{19,coref=12}$ *et* ***Cooper***$_{20,coref=11}$ *danser ensemble.* ***Donna***$_{21,coref=2}$ *arrive seule au bal et provoque la jalousie de* ***Val***$_{22,coref=19}$ *en flirtant avec* ***Noah***$_{23}$*.* ***Ce dernier***$_{24,anaph=23}$ *s'approche tout de même de* ***Val***$_{25,coref=22}$ *et provoque maintenant la jalousie de* ***Cooper***$_{26,coref=20}$ *qui décide de s'en aller. Mais lorsque* ***Noah***$_{27,coref=23}$ *voit le bracelet que* ***Cooper***$_{28,coref=26}$ *a offert à* ***Val***$_{29,coref=25}$*,* ***il***$_{30,anaph=27}$ *devient furieux.* (26)

```
1  Une femme
      o lui
      o Celle-ci
      o l'
2 Donna
      o Donna
3 son défunt mari
      o il
4 Steve
      o l'
      o Steve
5 Carly
      o Carly
6 Cooper
      o Cooper
            - Cooper
                  # Cooper
7 Val
      o Val
            - Val
                  # Val
                        * Val
```

[19] `<ol>` permet de créer une liste dont les éléments sont délimités par les balises `<li>`.

```
8 David
      o lui
9 Brandon
10 Kelly
11 Noah
      o Ce dernier
      o Noah
            - il
```
(27)

De cette façon, il est plus facile de visualiser le texte partitionné en chaînes coréférentielles. Le texte comporte 11 chaînes (classes d'équivalences) dont la plus longue contient cinq éléments (Val).

## 4.3 Conclusion

*XML* est probablement le métalangage idéal pour baliser un corpus manuellement. Sa syntaxe est relativement simple mais est suffisamment expressive pour pouvoir décrire des phénomènes linguistiques complexes. Une des forces de *XML* est sa simplicité. Il est de plus largement utilisé et il existe beaucoup d'outils et de langages de programmation pour faire le traitement de documents *XML*. Par exemple, un éditeur de *XML*[20] facilite le balisage manuel à l'aide d'une interface conviviale et de menus qui réduisent le nombre de caractères à taper et les fautes de frappe.

L'étude des chaînes coréférentielles se prête particulièrement bien à un balisage *XML*. Les syntagmes nominaux peuvent être encadrés par des balises `<sn>` et contenir les informations à propos des fonctions syntaxiques. Les informations d'ordre lexicographique sont aussi facilement identifiables (les déterminants `<det>` et les têtes, `<npr>`, `<nom>` et `<pro>`). Enfin, les liens coréférentiels (ou anaphoriques) peuvent être clairement établis à l'aide des attributs «pointeurs» `id` et `idref`. Bien que les liens coréférentiels soient difficiles à lire directement dans le document *XML*, ce dernier peut facilement être «transformé» de sorte que les liens soient mieux présentés (comme c'est le cas pour (27) page précédente).

[20] Comme par exemple *Oxygen*, utilisé dans le cadre du présent mémoire.

# Chapitre 5

# Méthodologie

Maintenant que nous avons mieux compris le fonctionnement des rouages des outils *XML* ; nous sommes désormais en mesure de développer une stratégie pour observer et traiter les phénomènes anaphoriques et coréférentiels. Plus précisément, notre étude comporte deux aspects :

**Analyse :** Observer les phénomènes coréférentiels et leur variation selon le type de texte.

**Production :** Déterminer les étapes importantes d'un algorithme permettant d'établir de façon automatique les liens coréférentiels d'un texte.

Pour répondre à nos attentes, nous devons avoir un corpus composé de différents types de textes. De plus, nous devons élaborer une méthodologie pour produire de bonnes règles *knowledge-poor*. La sélection des textes du corpus et l'élaboration de l'algorithme nécessitent un certain nombre d'étapes :

- la sélection et balisage des textes du corpus ;
- l'analyse du corpus d'entraînement et la construction d'un algorithme permettant d'établir automatiquement les chaînes coréférentielles d'un texte ;
- l'évaluation de l'algorithme dans le corpus de test.

## 5.1 Sélection du corpus

Dans l'introduction de ce mémoire (voir section 1.2.2 page 7), nous avons émis l'hypothèse que certains phénomènes de coréférence varient selon le type de texte. Pour vérifier notre supposition, nous avons sélectionné un corpus composé de types de textes choisis de trois domaines différents. De plus, pour nous permettre d'évaluer nos résultats, nous avons divisé notre corpus en corpus d'entraînement et en corpus de test. Nous avons utilisé le corpus d'entraînement pour faire nos observations. C'est le corpus avec lequel nous avons travaillé pour développer les règles *knowledge-poor* qui entrent dans la composition de notre algorithme. Lors

de la sélection de nos sous-corpus, nous avons aussi pris soin de réserver une proportion des textes (20% des textes) constituant notre corpus de test.

Pour sélectionner notre corpus, nous avons suivi les étapes suivantes :

1. le choix des domaines et des textes du corpus ;
2. le prétraitement des textes ;
3. la division des textes en corpus d'entraînement et en corpus de test ;
4. le balisage manuel des textes du corpus d'entraînement.

### 5.1.1 Choix des domaines et des textes

Puisque nous voulons comparer les phénomènes selon le type de texte, nous devons avoir à notre disposition des textes variés. Cependant, puisqu'il est nécessaire d'avoir un nombre assez grand de textes par type de texte et puisqu'il est assez long de traiter chacun des textes, nous avons choisi de nous limiter à trois types de textes seulement. Nous avons orienté le choix des domaines de manière à avoir des cas anaphoriques intéressants et contrastifs. Des contraintes de disponibilité des textes se sont aussi ajoutées à ces critères.

Notre corpus est constitué des trois types de textes suivants :

**Rapports sur les fusions de compagnies :** Les rapports sur les fusions de compagnies[1] sont des textes journalistiques dont le sujet est la fusion de deux (ou de plusieurs) compagnies. Ce type de texte (en version anglaise) a beaucoup été étudié dans le cadre des conférences *MUC*. Nos textes ont été sélectionnés sur `http://www.biblio.eureka.ca` à partir du serveur mandataire (*proxy*) des bases de données des bibliothèques de l'Université de Montréal. La recherche des textes a été faite à l'aide de la requête «*fusion compagnies*».

**Critiques de films :** Les critiques de films sont aussi des textes de type journalistique. Par contre, dans ces textes, l'auteur y est beaucoup plus présent car il présente son opinion à propos d'un film qu'il vient de visionner. Les textes ont été choisis sur le site de *Cyberpresse* à l'adresse suivante : `http://www.cyberpresse.ca`.

***HOWTO Linux* français :** Les *HOWTO Linux* français sont des traductions des *HOWTO Linux* anglais. Comme le nom l'indique, ces textes décrivent les marches à suivre pour installer ou utiliser des logiciels fonctionnant sur le système d'exploitation *Linux*. Les textes ont été puisés à partir du site Internet `http://www.freenix.fr/unix/linux/HOWTO/`. Puisque ce dernier corpus est assez vaste, nous avons limité nos choix en ne sélectionnant que les chapitres comportant le mot *installation* dans le titre.

Le choix des textes s'est effectué en deux occasions (été 2002 et printemps 2003). Lors de ces séances de choix de textes, nous avons choisi une série de textes satisfaisant aux critères posés

[1] Pour ne pas alourdir inutilement le texte, nous utiliserons l'expression *fusions de compagnies* pour désigner les rapports sur les fusions de compagnies.

plus haut. Nous avons choisi 25 textes pour chacun des types de textes déterminés. Pour sélectionner et partitionner le corpus en corpus d'entraînement et corpus de test sans biais, nous avons utilisé un générateur aléatoire mécanique (c'est-à-dire un dé à six côtés). Nous retrouvons un exemple de chacun des types de texte en annexe (voir annexe C page xv). Dans ces exemples, quelques-unes des chaînes que nous jugions importantes ont été mises en évidence (nous avons utilisé des couleurs différentes pour les différencier).

### 5.1.2 Prétraitement

Pour avoir une uniformité des données, nous avons transformé tous les textes recueillis en format texte brut (*flat text*). L'utilisation d'un tel format nous a permis de convertir aisément nos données en *XML*. Les textes de fusions de compagnies étant déjà en format texte, aucune conversion ne fut nécessaire. Par contre, les critiques de films et les *HOWTO Linux* étaient originellement en *HTML*. Un simple `lynx -dump` *fichier_HTML* (dans un environnement *UNIX*) a permis de transformer les textes dans le format désiré[2].

Pour chacun des textes à analyser, nous avons procédé à un petit travail d'identification (semi-)manuelle de certaines méta-informations concernant le texte. Ainsi, nous avons délimité les titres, les sous-titres, les paragraphes et les phrases. Pour ne pas introduire trop de bruit dans l'analyse des textes, nous avons aussi mis de côté des bouts de textes qui étaient moins représentatifs du français (les tableaux, les bouts de code (dans les *HOWTO Linux*), certaines méta-informations (auteur, date, provenance, etc.) Cette identification a en grande partie été effectuée manuellement.

### 5.1.3 Corpus d'entraînement et de test

Voici quelques statistiques concernant la constitution de notre corpus :

TAB. 5.1 – COMPOSITION – CORPUS D'ENTRAÎNEMENT

| | textes | phrases | S.N. réf. | mots |
|---|---|---|---|---|
| Fusions de compagnies | 20 | 356 | 972 | 9481 |
| Critiques de films | 20 | 414 | 1012 | 9730 |
| *HOWTO Linux* | 20 | 1516 | 2468 | 28717 |

### 5.1.4 Balisage

Nous désirions baliser les chaînes coréférentielles d'un texte. Pour réduire le travail, nous avons limité l'étiquetage aux chaînes contenant un nom propre ou un pronom. Pour ce faire, nous avons divisé le travail en quatre étapes :

[2]Malheureusement, lors de cette conversion, nous avons perdu quelques informations présentationnelles précieuses présentes dans le formatage du texte (division en liste, police de caractères, etc.) Nous reparlerons de l'impact de cette perte d'information dans le chapitre *Discussion* (section 7.4 page 89).

TAB. 5.2 – COMPOSITION – CORPUS DE TEST

| | textes | phrases | mots |
|---|---|---|---|
| Fusions de compagnies | 5 | 87 | 2518 |
| Critiques de films | 4 | 95 | 2916 |
| *HOWTO Linux* | 5 | 357 | 7668 |

1. Identifier tous les noms propres du corpus. Ensuite, baliser les éléments grammaticaux (déterminants, pronoms, prépositions, adverbes, etc.) qui ont été rencontrés dans les textes étudiés.
2. Baliser les syntagmes référentiels dont la tête$_r$ est un nom propre ainsi que les syntagmes contenant un pronom. Ensuite, identifier la fonction syntaxique de ces éléments.
3. Partager les syntagmes référentiels de l'étape précédente en chaînes de coréférence.
4. Compléter les chaînes obtenues en ajoutant les autres syntagmes nominaux s'y référant.

Les chaînes coréférentielles comportant un nom propre **ou** un pronom ont été identifiées ainsi dans le corpus d'entraînement. Elles ont été codifiées et validées selon la schématisation élaborée en 4.1.5 page 42.

## 5.2 Élaboration de l'algorithme

Il est probable que toutes les personnes ayant balisé manuellement un corpus de textes dans le but d'en faire une analyse linguistique s'entendent sur un point : le balisage manuel d'un corpus est une tâche longue et ennuyeuse ! En effet, puisque le corpus doit, généralement, être assez gros pour contenir un nombre considérable d'occurrences et contenir une diversité assez grande du phénomène à étudier, il s'ensuit qu'il y a beaucoup de manifestations à baliser.

Pour se donner une idée de la taille de certains corpus, Cobuild, une organisation qui publie des ouvrages de référence pour les apprenants de l'anglais tel des dictionnaires et des grammaires, utilise un corpus d'environ 450 millions de mots[3] constitué de manuels, de nouvelles, d'articles de journaux, etc. Ce corpus n'est certainement pas entièrement étiqueté à la main mais nécessite toutefois une grande équipe de spécialistes pour l'analyse des résultats.

C'est en pensant à des projets comme Cobuild que l'on réalise l'importance de l'automatisation du balisage des corpus linguistiques. Étiqueter à la main un corpus comme le nôtre (environ 50000 mots) est un travail assez long mais un corpus 10000 fois plus grand est impensable. Ce n'est cependant pas en vain que nous avons effectué cet exercice de balisage :

- Nous avons une connaissance intime des corpus étudiés.
- Nous avons constaté que certains phénomènes apparaissent fréquemment. Une personne suffisamment «paresseuse» pensera à comprendre les phénomènes répétitifs dans le but de les automatiser (par exemple en effectuant un «chercher/remplacer»).

[3] Voir `http://www.cobuild.collins.co.uk/` pour plus de renseignements.

- Nous avons remarqué que ce ne sont pas tous les syntagmes nominaux qui méritent d'être identifiés.
- Nous avons maintenant une source d'informations linguistiques plus facilement interrogeable à l'aide d'outils comme *XPath* (présenté à la section 4.2.3 page 45).
- Nous pouvons comparer le corpus d'entraînement avec le balisage manuel et les résultats de l'algorithme.

Toutes ces constatations ont contribué à dégager les étapes essentielles à la construction d'un algorithme permettant de résoudre les chaînes coréférentielles d'un texte. Bien certainement, nous y avons intégré les généralités que nous avons relevées à propos des liens coréférentiels *knowledge-poor* (voir section 3.3 page 33) :

- La compatibilité de certains traits syntaxiques (genre, nombre, personne) et sémantiques (animation) entre les antécédents et les anaphores est très importante.
- Les sujets sont les antécédents les plus fréquents. Viennent ensuite les compléments d'objet direct, les compléments d'objet indirect et les autres compléments.
- La distance entre l'antécédent et l'anaphore n'est jamais très grande ; au plus quelques phrases. Certaines conditions permettent aussi à l'antécédent et au pronom d'être dans la même phrase.
- Les réitérations lexicales sont de bons indicateurs de liens coréférentiels.
- Les référents associés au domaine se comportent de façon différente des autres éléments référentiels du texte. En particulier, le vocabulaire utilisé pour désigner ces entités est beaucoup plus prévisible que pour les autres entités référentielles et est relativement stable pour un type de texte donné. Ces référents semblent aussi plus faciles à résoudre et sont les constituants des chaînes coréférentielles les plus longues du texte.

La nature même de la référence implique une procédure divisée selon les niveaux linguistiques (c'est-à-dire lexical, syntaxique et textuel). Nous avons séparé les niveaux syntaxiques en deux parties en distinguant les phénomènes qui se produisent à l'intérieur du syntagme référentiel (niveau syntaxique I) des phénomènes qui impliquent l'interaction des syntagmes entre eux (niveau syntaxique II)[4].

**Niveau lexical :** Identifier le vocabulaire clé du domaine et les autres mots utiles pour les étapes suivantes.

**Niveau syntaxique I :** Borner certains syntagmes référentiels (les syntagmes les plus susceptibles de se retrouver à l'intérieur des chaînes coréférentielles importantes).

**Niveau syntaxique II :** Déterminer les fonctions syntaxiques des syntagmes identifiés.

**Niveau textuel :** Établir les liens coréférentiels et anaphoriques entre les syntagmes référentiels du texte.

Pour chacune de ces étapes, nous ferons une brève description des stratégies empruntées pour développer des règles *knowledge-poor*. Nous y décrirons quelques indices relevés à même le corpus d'entraînement ou à partir de la théorie présentée dans les chapitres précédents. Ces

[4]Cette nomenclature est tirée de K. Jonasson [Jon94].

indices seront les éléments qui nous aideront à construire les règles présentées au prochain chapitre (voir section 6.2 page 65). Nous espérons, à l'aide d'une combinaison de ces indices, pouvoir construire un algorithme permettant d'établir automatiquement les liens coréférentiels d'un texte. Nous en reparlerons à la section 6.2.4 page 73 du chapitre suivant[5].

### 5.2.1 Niveau lexical

Un des principaux obstacles du traitement automatique de la langue est l'élaboration d'un dictionnaire assez complet pour reconnaître et désambiguïser tous les mots d'un texte. On peut diviser les mots d'une langue en deux catégories : les mots appartenant à des **classes fermées** (les mots grammaticaux) et les mots appartenant à des **classes ouvertes** (les mots lexicaux).

La première catégorie de mots contient relativement très peu de mots ; elle est constituée des déterminants, des prépositions, des pronoms, des conjonctions, des auxiliaires et de certains adverbes. L'identification de ces mots servira à déterminer les bornes possibles des syntagmes référentiels ainsi qu'à déterminer les fonctions syntaxiques (section 6.2.3 page 71). Nous insistons sur le caractère fermé de ces catégories de mots (c'est-à-dire qu'il y en a un nombre limité) car ils sont plus facilement identifiables.

La deuxième catégorie de mots regroupe les noms communs, les verbes, les adjectifs et la plupart des adverbes. C'est cette classe de mots qui pose habituellement problème pour la catégorisation des mots d'un texte. Cependant, notre étude ne nécessite essentiellement que l'identification des noms communs. Nous pensons qu'avec l'aide des déterminants et de l'établissement d'une **liste de noms communs importants (selon le domaine)**, nous pourrons contrecarrer une partie de la difficulté. Nous devrons, par ailleurs, identifier les noms propres, certains noms communs faisant référence à la situation d'énonciation, ainsi que certains verbes impersonnels. Nous reparlerons de la composition de ces catégories de mots à la section 6.2.1 page 65.

### 5.2.2 Niveau syntaxique I

Nous avons élaboré, au chapitre *Notions théoriques*, une typologie des syntagmes nominaux (voir section 2.1 page 14). Nous avons déterminé deux aspects importants concernant la référence : la nature de la tête du syntagme et les déterminants utilisés. Nous avons donc, d'un côté, les syntagmes chapeautés par des noms communs, des noms propres, des pronoms ou des ellipses et, d'un autre côté, les syntagmes indéfinis, définis et démonstratifs.

Pour identifier tous les éléments référentiels d'un texte, il est habituellement nécessaire de passer par une analyse syntaxique du texte. Cependant, certains éléments référentiels peuvent être facilement identifiables grâce notamment aux mots appartenant aux classes fermées relevées à l'étape précédente. Revoyons chacun des types de syntagmes référentiels selon la possibilité de les identifier grâce à ces éléments :

[5]Notons que la division en niveaux linguistiques correspond aux niveaux d'analyse du corpus des textes balisés manuellement ainsi qu'aux étapes du traitement automatique des chaînes coréférentielles que nous élaborerons.

**Pronoms :** Les syntagmes contenant un pronom sont simples à identifier car ils ne contiennent habituellement que le pronom. La classe des pronoms est une classe assez vaste. Elle comprend les pronoms personnels, les pronoms démonstratifs, les pronoms possessifs, les pronoms indéfinis, les pronoms relatifs et les pronoms interrogatifs. Les plus intéressants pour nous sont les pronoms personnels et les pronoms démonstratifs. Nous laisserons de côté les autres pronoms soit parce qu'ils sont trop peu nombreux dans le corpus (pronoms indéfinis et interrogatifs), qu'ils demandent une résolution double (pronoms possessifs) ou bien parce qu'ils sont entièrement déterminés par la syntaxe (pronoms relatifs).

**Ellipses :** Les ellipses sont les éléments qui sont possiblement les plus anaphoriques. Leur entité associée est tellement saillante qu'elle est implicite dans le texte. Il va donc de soi que la résolution complète des anaphores nécessite de les identifier. Par contre, il est très difficile de reconnaître les ellipses car ce qui les caractérise est justement l'absence de mots. Nous choisissons donc de ne pas tenir compte des ellipses dans notre étude car ceci demande une analyse syntaxique complexe.

**Syntagmes contenant un nom propre :** Bien souvent, les syntagmes dont la tête est un nom propre ne contiennent que ce nom propre. Cependant, il y a des cas où le syntagme nom propre contient un autre nom propre (*Marie Curie*) un déterminant (*Le Canada*), des compléments (*Brown Boveri, de Suisse,*) des adjectifs (*La petite Marie*) ou même des noms communs (*le metteur en scène Beno Besson*). Dans ce dernier cas, nous remarquons toutefois que le mot entre le déterminant et le nom propre est souvent un nom commun identifié comme étant un nom associé au domaine.

**Syntagmes démonstratifs :** Les syntagmes démonstratifs sont typiquement anaphoriques. Ils sont aussi très simples à identifier car ils sont habituellement constitués d'un déterminant démonstratif et d'un nom commun seulement. Quelquefois, ils peuvent être suivis d'un complément du nom ou d'une apposition. Parmi les syntagmes démonstratifs, il faut cependant prendre soin de trier les syntagmes contenant un nom commun référant à la situation d'énonciation (*cette année*) et les expressions dont le démonstratif est employé comme un pronom (*ce que* ou *c'est*)[6]. Ces trois dernières expressions ne sont pas référentielles pour nous.

**Autres syntagmes :** Les autres syntagmes sont plus complexes à identifier. Pour ces cas, nous misons sur le fait que ceux qui méritent notre attention contiennent un nom commun associé au domaine.

En résumé, voici les syntagmes que nous jugeons faciles à identifier :

- les syntagmes dont la $\text{tête}_r$ est un pronom ;
- les syntagmes dont la $\text{tête}_r$ est un nom propre ;
- les syntagmes dont la $\text{tête}_r$ est un nom commun associé au domaine ;
- les syntagmes comportant un déterminant démonstratif.

Par contre, nous ignorerons :

[6]On peut reconnaître ces dernières car elles sont suivies d'un verbe auxiliaire ou d'un pronom relatif.

- les ellipses ;
- les syntagmes indéfinis ne comportant pas de nom commun associé au domaine ;
- les syntagmes définis ne comportant pas de nom commun associé au domaine.

### 5.2.3 Niveau syntaxique II

Les fonctions syntaxiques sont très importantes pour résoudre les liens coréférentiels. Par exemple, les appositions sont habituellement coréférentielles à l'élément qui précède. À la section 3.3 page 33, nous avons aussi déterminé que certaines fonctions syntaxiques, tels les sujets, sont des antécédents plus probables. Nous observons plus ou moins quatre types de fonctions syntaxiques dans le texte :

- les sujets ;
- les compléments du verbe ;
- les compléments du nom ;
- les appositions.

Il est assez difficile d'assigner une fonction syntaxique à un syntagme référentiel sans passer par une analyse syntaxique globale de la phrase. On peut toutefois obtenir une bonne approximation de la fonction syntaxique en analysant :

**La position dans la phrase :** Lorsqu'un syntagme référentiel est en début de phrase, il s'agira probablement d'un sujet.

**La présence d'une préposition :** Les prépositions introduisent des compléments du nom ou des compléments indirect.

**La position par rapport aux autres syntagmes de la phrase :** La position des syntagmes référentiels par rapport aux autres syntagmes de la phrase peut permettre d'identifier les compléments du nom ou les appositions.

Un procédé similaire a justement été proposé par A. Siddharthan ([Sid03]) pour les textes anglais (voir section 3.2.6 page 32).

### 5.2.4 Niveau textuel

Dans les chaînes coréférentielles, nous avons trois types de relations possibles (voir section 2.2.1 page 20) :

- une $\text{tête}_c$ ;
- une coréférence ;
- une anaphore.

**$\text{Tête}_c$ et coréférence :** Les têtes de chaînes coréférentielles peuvent être des noms propres, des syntagmes indéfinis et des syntagmes définis autonomes. Ces syntagmes ont aussi la possibilité d'être coréférentiels à un autre élément dans le texte.

**Anaphore :** Parmi les syntagmes référentiels, nous avons déterminé que les pronoms, les ellipses, les syntagmes définis de reprise, les syntagmes définis associatifs, et les syntagmes démonstratifs sont nécessairement anaphoriques. Pour tous ces cas, il est important de prendre en considération les compatibilités syntaxiques et sémantiques entre l'anaphore et l'antécédent. Le genre et le nombre du syntagme pourront dépendre des déterminants, des pronoms et des noms communs qui le constituent. De plus, dans les cas définis (et quelques cas démonstratifs), nous observons des cas de réitération lexicale. Ceci constitue un autre indice important pour trouver le bon antécédent. Il est à noter que les liens anaphoriques lexicaux peuvent couvrir une plus grande distance que les liens anaphoriques pronominaux, ces derniers étant confinés à une distance d'une ou deux phrases, tout au plus.

## 5.3 Mesures d'efficacité de l'algorithme

Pour mesurer l'efficacité de notre algorithme, il est essentiel d'établir une métrique pour la validation et la comparaison de nos résultats. Les mesures de précision et de rappel sont les mesures les plus utilisées dans les travaux de linguistique computationnelle. Elles sont essentiellement des mesures de comparaison entre les résultats idéaux et les résultats obtenus dans la pratique.

### 5.3.1 Précision et rappel

Les résultats d'un algorithme peuvent être classés en quatre cas de figure (tiré de [MS99]) :

| | visé | non visé |
|---|---|---|
| S.N. sélectionné | $vp$ | $fp$ |
| S.N. non sélectionné | $fn$ | $vn$ |

où $vp$ tient pour vrai positif, $fp$ pour faux positif, $fn$ pour faux négatif et $vn$ pour vrai négatif.

Dans le cas idéal, nous avons une correspondance parfaite entre les résultats théoriques et les résultats réels (les résultats de l'algorithme). Ainsi, dans le cas d'un algorithme parfait, il n'y a aucun faux positif et aucun faux négatif.

Cependant, la réalité n'est habituellement pas aussi parfaite ; certains syntagmes référentiels dans la chaîne peuvent avoir été choisis en trop (faux positifs) ou en moins (faux négatifs) :

**Précision :** Nous définissons la précision ($P$) comme la proportion de bons résultats obtenus parmi le nombre total de résultats obtenus. Ainsi :

$$P = \frac{vp}{vp + fp} \tag{5.1}$$

Une bonne précision privilégie l'obtention de bonnes réponses.

**Rappel :** Le rappel ($R$) se définit comme étant le nombre de bons résultats trouvés, divisés par le nombre total de bons résultats. Soit :

$$R = \frac{vp}{vp + fn} \tag{5.2}$$

Un bon rappel s'obtient lorsque presque toutes les bonnes réponses possibles sont identifiées.

Nous utilisons les mesures de précision et de rappel pour évaluer la résolution des anaphores pronominales dans le corpus d'entraînement et le corpus de test (c'est-à-dire les pronoms personnels de troisième personne). Ces mesures nous permettront aussi de comparer nos résultats à d'autres algorithmes sur ces mêmes points. Notons que la précision et le rappel sont des valeurs comprises entre zéro et un. Une bonne précision et un bon rappel ont des valeurs s'approchant le plus possible de un.

### 5.3.2 *B-CUBED*

Bien que les mesures de précision et de rappel soient adéquates pour mesurer certains phénomènes précis (comme la résolution des pronoms personnels), ces mesures permettent difficilement de vérifier les résultats plus globaux. Par exemple, l'efficacité de la résolution des chaînes coréférentielles est plus difficilement mesurable. En effet, les erreurs d'identification des chaînes coréférentielles peuvent être de différentes natures :

- Les chaînes coréférentielles peuvent avoir été bien identifiées par partie (c'est-à-dire que certaines des sous-chaînes peuvent avoir été bien identifiées) mais globalement elles peuvent manquer de connectivité. Par exemple, une chaîne peut avoir été coupée en deux.
- Certains bouts de chaînes peuvent avoir été identifiés en trop dans une chaîne coréférentielle donnée.
- L'importance de la chaîne n'est pas prise en considération. Ainsi, nous voudrions accorder plus d'importance aux chaînes plus longues car elles sont plus pertinentes pour comprendre le contenu d'un texte (et nous semblent donc plus importantes à résoudre).

Mentionnons toutefois qu'il existe des mesures permettant d'évaluer les chaînes coréférentielles en tenant compte de la précision et du rappel d'un élément dans la chaîne coréférentielle ainsi que de la précision et du rappel de la chaîne coréférentielle en entier[7]. Par exemple, l'algorithme *B-CUBED* présenté par B. Baldwin [BMB+95] :

**Précision et rappel d'un élément référentiel :** Pour tous les éléments référentiels identifiés ($i$), nous pouvons calculer la précision ($P_i$) et le rappel ($R_i$) de cet élément de la façon suivante :

$$P_i = \frac{\text{nombre de bons éléments dans la chaîne trouvée contenant } i}{\text{nombre total d'éléments dans la chaîne contenant } i} \tag{5.3}$$

[7]Nous n'avons pas effectués les calculs *B-CUBED* dans notre recherche. Nous en expliquerons les raisons à la section 6.3 page 75.

$$R_i = \frac{\text{nombre de bons éléments dans la chaîne trouvée contenant } i}{\text{nombre total d'éléments dans la chaîne réelle contenant } i} \tag{5.4}$$

**Précision et rappel de la chaîne coréférentielle :** La précision finale ($P_f$) et le rappel final ($R_f$) –c'est-à-dire la précision et le rappel des chaînes coréférentielles pour le texte en entier– s'obtiennent de la façon suivante :

$$P_f = \sum_{i=1}^{N} w_i \times P_i \tag{5.5}$$

$$R_f = \sum_{i=1}^{N} w_i \times R_i \tag{5.6}$$

où $N$ est le nombre de références au total et $w_i$ est le poids accordé à l'élément $i$ (par exemple, soit $l$, la longueur de la chaîne ; nous pouvons choisir $\frac{1}{l}$ comme valeur $w_i$ de tous les éléments de la chaîne).

## 5.4 Conclusion

Nous avons fait une analyse des phénomènes coréférentiels en utilisant un échantillon de trois types de textes français :

- les fusions de compagnies (style journalistique) ;
- les critiques de films (style journalistique) ;
- les *HOWTO Linux* (textes procéduraux).

Dans le but d'alléger le traitement des chaînes coréférentielles, nous avons simplifié et réduit le nombre de cas traités.

Pour faire suite à l'analyse manuelle et statistique de ces textes, nous avons essayé de déterminer quelles sont les ressources minimales nécessaires pour l'identification du maximum de chaînes coréférentielles à l'aide d'un programme informatique simple. Ces ressources sont, pour la plupart, de nature lexicale, syntaxique et textuelle :

**Niveau lexical :**

- Les mots grammaticaux appartenant à des classes fermées ;
- quelques noms communs associés au domaine ;
- quelques noms communs référant à la situation d'énonciation ;
- quelques verbes impersonnels ;
- les noms propres.

**Niveau syntaxique I :**

- Les pronoms ;
- les syntagmes comportant un nom propre ;
- les syntagmes comportant un nom commun associé au domaine ;

- les syntagmes comportant un déterminant démonstratif.

**Niveau syntaxique II :**

- Les sujets ;
- les compléments du verbe ;
- les compléments du nom ;
- les appositions.

**Niveau textuel :**

- Les têtes$_c$ ;
- les coréférences ;
- les anaphores.

Pour nous aider à résoudre les liens coréférentiels et anaphoriques entre les syntagmes référentiels, nous utiliserons aussi les indices suivants :

- la compatibilité syntaxique et sémantique ;
- la distance entre un syntagme référentiel et son antécédent ;
- la réitération lexicale ;
- les fonctions syntaxiques trouvées par approximation.

Dans le prochain chapitre, nous utiliserons ces indices pour élaborer un système de règles *knowledge-poor* établissant les liens coréférentiels et anaphoriques entre les éléments référentiels d'un texte. Par la suite, nous évaluerons les résultats avec les mesures de précision et de rappel.

# Chapitre 6

# Présentation et analyse des résultats

Notre étude vise avant tout à dégager les généralités et les particularités des phénomènes coréférentiels selon le type de texte. Nous croyons que ce qui est général pour tous les textes est relativement simple. Il serait donc possible de résoudre ces liens coréférentiels à l'aide d'un algorithme utilisant des règles *knowledge-poor*. Cependant, les différences des moyens coréférentiels entre les types de textes en termes de fréquence des moyens, ou l'emploi de constructions particulières au domaine offrent de plus grands défis pour le traitement automatique de la langue. Pour pouvoir résoudre ces types de liens coréférentiels, une étude détaillée pour chacun des types de textes est nécessaire.

La présentation et l'analyse des résultats de notre travail s'orientent selon plusieurs aspects. Dans un premier temps, nous présenterons quelques résultats tirés du corpus d'entraînement balisé manuellement. Pour faire suite à l'analyse de ces résultats, nous proposerons un système de règles simple pour résoudre les liens coréférentiels «les plus faciles» d'un texte. Nous tenterons, dans un troisième temps, d'évaluer sommairement la pertinence de ces règles à l'aide des mesures proposées au chapitre précédent (voir section 5.3 page 58). Enfin, nous ferons une étude comparative des types de textes de notre corpus afin d'établir les différences entre chacun.

## 6.1 Corpus balisé manuellement

Au chapitre portant sur l'utilisation des outils *XML* (voir chapitre 4 page 35), nous avons présenté quelques astuces pour «interroger» un texte étiqueté en *XML*. Pour notre étude, nous nous sommes servis de *XSLT* pour nous aider à analyser les textes balisés. Voici quelques éléments qui ont contribué à notre analyse :

- la présentation des textes balisés ;
- la fréquence des noms communs ;
- la comparaison des types de syntagmes référentiels selon les différents facteurs influençant la coréférence.

### 6.1.1 Présentation des textes balisés

Pour étudier la coréférence dans les textes, il peut être utile de présenter les textes de deux façons : la présentation en partitionnement des textes en chaînes coréférentielles et la présentation des chaînes coréférentielles en contexte. Notons que ces présentations ont été utilisées pour analyser les textes balisés manuellement et les textes balisés automatiquement. Nous avons ici exploité un des avantages des outils *XML* : la réutilisation.

#### 6.1.1.1 Textes en chaînes coréférentielles

La présentation d'un texte en partitionnement des éléments référentiels en chaînes coréférentielles offre un point de vue global sur l'ensemble des interactions entre les éléments référentiels du texte. Nous avons, à la section 4.2.4 page 46, présenté une transformation *XSLT* permettant de restructurer un texte balisé sous forme de chaînes coréférentielles (en utilisant des listes imbriquées). Nous avons utilisé une transformation semblable à celle-ci pour présenter tous les textes balisés de notre corpus sous forme de «tableau de dépendances». Dans ce cas-ci, nous avons utilisé un tableau pour distinguer les cas coréférentiels des cas anaphoriques.

Pour nous donner une idée des éléments présents dans les chaînes coréférentielles du corpus d'entraînement, nous retrouvons en annexe un exemple de texte **balisé manuellement** présenté en chaînes de coréférence (voir annexe E page xxiii). Il est à remarquer que les chaînes sont, dans ce cas-ci, ordonnées selon l'ordre d'occurrence de la $\text{tête}_c$ dans le texte. Il aurait aussi été possible d'ordonner les chaînes selon leur longueur.

Voici les conventions que nous avons utilisées pour la présentation des chaînes coréférentielles :

- Les cases les plus larges sont les têtes de chaînes coréférentielles.
- À l'intérieur des tableaux sous-divisés, les cases à gauche sont des liens coréférentiels et les cases à droite sont les liens anaphoriques.
- L'antécédent des anaphores et des coréférences se situe dans la case juste en haut.
- La distance entre un élément coréférentiel et son antécédent (c'est-à-dire le nombre de syntagmes balisés les séparant) est indiquée dans le coin supérieur gauche de la case.
- Les numéros d'identification et la fonction syntaxique sont notés dans le coin inférieur droit de la case.

#### 6.1.1.2 Chaînes coréférentielles en contexte

Bien que la représentation d'un texte en chaîne de coréférence soit pratique pour visualiser les structures coréférentielles d'un texte, il est aussi utile de voir ces chaînes en contexte. Ainsi, il est plus facile de comparer la distance entre les éléments coréférentiels, leur distribution et l'enchâssement des chaînes coréférentielles. Pour avoir une telle vue sur le phénomène, nous avons utilisé une présentation comme celle qui est montrée en annexe (voir annexe F page xxxi). Notons que ces exemples sont des **résultats obtenus à partir de l'algorithme** présenté à la

section 6.2 page suivante. Nous y avons choisi un exemple de texte mettant en évidence un exemple de chaîne importante dans le texte.

### 6.1.2 Fréquence des noms communs

Au chapitre précédent, nous avons décrit notre méthodologie pour le balisage manuel de notre corpus. Revoici les étapes de l'étiquetage manuel (voir 5.1.4 page 52) :

1. Identifier tous les noms propres du corpus. Ensuite, baliser les éléments grammaticaux (déterminants, pronoms, prépositions, adverbes, etc.) que nous supposons pertinents.
2. Baliser les syntagmes référentiels dont la tête$_r$ est un nom propre ainsi que les syntagmes contenant un pronom et identifier la fonction syntaxique de ces éléments.
3. Partager les syntagmes référentiels de l'étape précédente en chaînes de coréférence.
4. Compléter les chaînes obtenues en ajoutant les autres syntagmes nominaux s'y référant.

Il y a relativement peu de noms communs étiquetés de cette façon car nous n'avons relevé que les noms communs appartenant à une chaîne coréférentielle contenant un nom propre ou un pronom[1]. Les noms communs identifiés de cette façon sont assez similaires d'un texte à un autre pour un type de texte donné. Ils sont cependant très différents d'un type de texte à un autre. La liste des noms communs identifiés de cette façon (classés selon la fréquence) est disponible en annexe (voir annexe G page xlviii). Elle sera utile pour l'identification des noms communs associés au domaine[2].

### 6.1.3 Comparaison des types de syntagmes référentiels

Des tableaux présentant quelques statistiques de fréquence (selon les trois types de textes étudiés) se trouvent aussi en annexe (voir annexe H page liv). Ces tableaux contiennent les comptes de certains événements relevés à même le corpus balisé manuellement et ont été obtenus à l'aide d'une transformation *XSLT* appliquée à tout les sous-corpus. Ils mettent en correspondance les têtes des syntagmes référentiels, leur déterminant, leur complexité[3] et leur fonction syntaxique. Pour chacun des domaines, le premier des quatre tableaux contient tous les syntagmes nominaux des chaînes coréférentielles contenant au moins deux éléments coréférentiels. Le second tableau représente les têtes de chaînes coréférentielles (le premier élément) seulement. Enfin, les deux derniers tableaux représentent les syntagmes anaphoriques et les syntagmes coréférentiels (mais non anaphorique) des textes étudiés.

[1] Nous avons peut-être identifié la moitié des noms communs du texte.
[2] Nous en reparlerons à la section 6.4.1 page 78.
[3] Un syntagme nominal est complexe s'il contient une préposition, une conjonction, un pronom relatif ou un signe de ponctuation.

## 6.2 Règles *knowledge-poor*

Au chapitre précédent, nous avons déterminé quatre niveaux importants pour l'identification des liens coréférentiels d'un texte (voir section 5.2 page 53). Cette même division est utilisée ici pour l'élaboration et l'application des règles *knowledge-poor*[4] :

**Traitement lexical :** Identifier le vocabulaire clé du domaine et les autres mots utiles pour les étapes suivantes.

**Traitement syntaxique I :** Borner certains syntagmes référentiels (les syntagmes les plus susceptibles de se retrouver à l'intérieur des chaînes coréférentielles importantes).

**Traitement syntaxique II :** Déterminer les fonctions syntaxiques des syntagmes identifiés.

**Traitement textuel :** Établir les liens coréférentiels et anaphoriques entre les syntagmes référentiels du texte.

### 6.2.1 Traitements au niveau lexical

Déterminer les parties du discours dans un texte est habituellement une tâche difficile sans l'utilisation d'un dictionnaire. Pour simplifier le problème, nous misons sur le fait qu'il est possible de faire une identification partielle des parties du discours. Ainsi, nous n'identifierons que certains mots appartenant à des classes de mots comportant un nombre restreint d'occurrences. Ce à quoi nous ajoutons les noms propres qui peuvent être reconnus grâce à la présence de la majuscule.

L'identification de ces mots fut probablement l'une des parties les plus difficiles de notre recherche. Elle a été faite de façon plus ou moins *ad hoc* en consultant diverses grammaires et les textes du corpus. Cette liste de mots est aussi l'élément central de notre algorithme. L'ensemble des mots du vocabulaire que nous avons jugé nécessaire est contenu dans le fichier que nous nommons `vocabulaire.xml` (voir annexe B page iv). Ce vocabulaire ne contient que quelques centaines de mots[5].

Une partie importante de ce vocabulaire est constitué des noms communs associés au domaine (environ une centaine de mots différents[6]). Pour déterminer les mots importants selon le domaine, nous avons essentiellement analysé la liste de mots présentée à la section 6.1.2 page précédente (voir aussi annexe G page xlviii). Nous reparlerons des différences entre le vocabulaire selon le type de texte à la section 6.4.1 page 78. De plus, nous discuterons de la possibilité d'automatiser l'extraction automatique d'un vocabulaire associé au domaine à la section 7.4.1 page 89.

[4] Notons qu'à partir de maintenant, les exemples cités sont directement puisés à même les résultats de l'analyse du corpus. Pour cette raison, nous ne citerons plus explicitement la provenance des exemples.

[5] Nous estimons qu'environ le quart des syntagmes référentiels identifiés par notre algorithme contiennent un nom associé au domaine.

[6] Nous avons considéré les mots dérivés morphologiquement comme étant des mots différents (par exemple (*fichier* et *fichiers*). Cette distinction n'a pour but que de simplifier le traitement automatique.

Pour l'identification des parties du discours, nous ne faisons essentiellement qu'un «chercher/remplacer» dans tous les textes du corpus. Ainsi, lorsqu'un mot du texte est contenu dans la liste du vocabulaire, il est remplacé par la balise identifiant la partie du discours ainsi que les attributs qu'elle contient. Par exemple, le mot *film* (la première entrée du fichier) est remplacé par <nom traits="domaine" genre="m" nombre="sing">*film*</nom>.

Pour faciliter le traitement, nous avons aussi choisi de décomposer les formes contractées. Ainsi *aux* est remplacé par la forme *à les*. Le cas de *des* est cependant problématique car il peut être la forme contractée de *de les* ou la forme du déterminant indéfini pluriel. Nous avons choisi de décomposer la forme *des* en *de les* car il s'agit du cas le plus courant [7].

Examinons en détail chacune des grandes classes de mots que nous traiterons dans cette étape :

- les mots outils ou appartenant à des classes grammaticales (par exemple : *un, le, par, avec, elle, lui, alors, mais, que, dont, pas*) ;
- les noms communs associés au domaine (par exemple : *fusion, compagnie, metteur en scène, acteur, fichier, disque*) ;
- les noms communs référant à la situation du discours (par exemple : *semaine, chapitre, article*) ;
- les verbes impersonnels (par exemple : *semble, faut, faudra*) ;
- les noms propres (par exemple : *Coline, Marie, Canada*).

#### 6.2.1.1 Mots grammaticaux

Les mots outils sont de bons indices pour l'identification des bornes des syntagmes référentiels ainsi que de leur fonction syntaxique (voir les sections 6.2.2 page 68 et 6.2.3 page 71). Nous ajoutons à cette liste les signes de ponctuation en prenant soin de distinguer les ponctuations simples des signes de ponctuation encadrant (parenthèses, crochets, guillemets et tirets). Les signes de ponctuation encadrant ont un statut particulier dans l'identification des appositions (voir 5.2.3 page 57).

Nous identifions ainsi :

- les déterminants (<det>) ;
- les pronoms (<pro>) ;
- les prépositions (<pre>) ;
- les conjonctions (<conj>) ;
- les pronoms relatifs (<rel>) ;
- certains adverbes (<adv>) ;
- certains verbes auxiliaires (<verbe>) ;
- les signes de ponctuation (<ponc>).

[7] Bien que les simplifications de l'identification des parties du discours des mots grammaticaux soient plus ou moins approximatives, nous ne croyons pas que ceci ait un impact majeur sur les résultats car ces mots ne sont que des indices indirects pour l'identification des syntagmes référentiels.

Cette liste détermine la catégorie grammaticale des mots et contient aussi d'autres informations syntaxiques et sémantiques (le genre, le nombre, la personne, etc.) Notons qu'il y a peu de cas d'ambiguïté dans la liste des mots grammaticaux. Le cas de *en* en est un exemple car il peut être un pronom ou une préposition. Dans ce cas, nous avons choisi de l'identifier en tant que préposition car c'est sa catégorie grammaticale la plus courante. Nous avons mis de côté certains pronoms personnels possiblement ambigus (*se*, *y* et *le*) car ils risquent d'introduire de mauvaises interprétations. Parmi les pronoms personnels, nous ne traiterons pas non plus les pronoms de première et de deuxième personne car ils réfèrent à la situation d'énonciation. De plus, certains pronoms démonstratifs seront exclus (*cela*, *ça*, *ce*) car ils réfèrent souvent à des propositions ou des entités plutôt vagues (*ça va barder*). Pour tous ces cas non traités, nous avons au préalable marqué ces pronoms par le trait `aTraiter="false"` dans le vocabulaire (voir annexe B page x).

#### 6.2.1.2 Noms communs associés au domaine

Dans les textes étudiés, il existe un sous-ensemble de noms communs qui sont assez fréquents dans les chaînes coréférentielles. Bien que les noms communs associés au domaine varient selon le type de texte, il semble toutefois que tous les types de textes possèdent une liste de mots importants pour la référence. Les mots appartenant à cette liste seraient en nombre limité et pourraient être regroupés en classes sémantiques plus ou moins régulières. Donc, pour chacun des domaines, il serait nécessaire d'identifier ces mots.

Voici quelques mots de cette liste :

**Fusion de compagnies :** *mariage, société, compagnie, conseil* ;

**Critiques de films :** *mère, acteur* ;

***HOWTO Linux* :** *fichier, disque, serveur*.

Nous reparlerons du vocabulaire des textes du corpus à la section 6.4.1 page 78. Retenons toutefois que leur nombre est limité ; ceci facilite leur identification dans le texte.

#### 6.2.1.3 Noms communs référant à la situation du discours

En faisant la revue de la littérature à la section 2.1.2 page 17, nous avons énoncé que l'antécédent d'un syntagme démonstratif s'identifie par reprise et qu'il n'est habituellement pas très loin. Cependant, il peut arriver que le référent ne soit pas dans le texte mais se situe dans la situation d'énonciation. Par exemple, nous retrouvons les mots suivant à l'intérieur de cette liste :

> *Les films d'inspiration martiale qui sortent sur nos écrans* ***ces jours-ci*** *sombrent allégrement, il me semble, dans la schizophrénie.* (1)
>
> *Si vous ne voyez pas ce dont il retourne, continuez la lecture et revenez à* ***cette section*** *ultérieurement.* (2)

*Le titre d'ABB a chuté de 85% **cette année**.* (3)

Nous nommerons ces noms *noms communs référant à la situation du discours*. Nous marquons donc ces noms pour bloquer le calcul préférentiel des antécédents d'un texte (dans le corpus, nous avons noté ces noms avec l'attribut `traits="déictique"`). Nous en reparlerons à la section 6.2.2.4 page 70.

#### 6.2.1.4 Verbes impersonnels

Le pronom *il* est souvent utilisé à la forme impersonnelle dans les textes. Bien entendu, nous ne voulons pas repérer un antécédent pour ces pronoms ! Malheureusement, la forme du pronom *il* n'indique rien à propos de son statut impersonnel. Pour aider à détecter les cas impersonnels, nous nous servirons de la nature du verbe suivant le pronom (par exemple : *il **semble**, il **s'agit**, il **faut***).

Ces verbes sont marqués par la présence de l'attribut `trait="impersonnel"` dans le fichier `vocabulaire.xml`. Ce trait servira à bloquer le calcul référentiel des pronoms.

#### 6.2.1.5 Noms propres

Les noms propres sont les éléments référentiels importants de notre étude. Bien qu'ils peuvent se présenter sous plusieurs formes, ils ont une caractéristique qui peut les distinguer : la majuscule.

**RÈGLE 6.1 (IDENTIFIER LES NOMS PROPRES)**
*La majuscule identifie un nom propre définitivement sauf au début de phrase. Si un mot débutant par une majuscule est en début de phrase, il est considéré comme un nom propre s'il existe un autre mot identique à celui-ci dans le texte et que ce dernier a déjà été identifié comme étant un nom propre.*

Cette règle très simple permet d'identifier la plupart des noms propres. Notons qu'il est toutefois possible de manquer les noms propres ne se trouvant qu'en début de phrases. Nous supposons que pour ces cas, si le nom propre n'est pas répété ailleurs dans le texte et dans le corps d'une phrase, il est probablement moins important.

### 6.2.2 Traitements au niveau syntaxique I

En étudiant les tableaux H page liv, nous confirmons quelques suppositions que nous avons faites au chapitre précédent. En particulier, les déterminants définis sont les déterminants les plus nombreux du corpus. De plus, lorsqu'ils sont en position de $\text{tête}_c$, ils sont significativement plus complexes.

À l'aide des parties du discours identifiées à l'étape précédente, nous pouvons reconnaître plusieurs syntagmes référentiels. L'identification d'un syntagme référentiel comporte trois parties importantes :

1. identifier la tête$_r$ ;
2. identifier le début du syntagme ;
3. identifier la bonne fin du syntagme (incluant les compléments).

Notons que l'identification des bornes gauches pose parfois problème pour les cas complexes. En effet, il se peut qu'il y ait des ambiguïtés syntaxiques impliquant l'attachement du complément du nom ou du verbe. Ce qui est important pour nous est l'identification des bonnes têtes référentielles ; l'identification de la frontière gauche sera parfois approximative mais nous ne croyons pas que cela soit très critique. Nous tenterons toutefois d'identifier les cas enchâssés les plus évidents.

Rappelons les syntagmes référentiels que nous avons jugés faciles à identifier (voir section 5.2.2 page 55) :

- les syntagmes dont la tête$_r$ est un pronom ;
- les syntagmes dont la tête$_r$ est un nom propre ;
- les syntagmes dont la tête$_r$ est un nom commun associé au domaine ;
- les syntagmes comportant un déterminant démonstratif.

#### 6.2.2.1 Syntagmes dont la tête$_r$ est un pronom

Un syntagme dont la tête$_r$ est un pronom est facilement identifiable car il ne contient que le pronom. Il faut par contre prendre soin d'éliminer la forme *il* lorsqu'elle précède un verbe impersonnel.

**RÈGLE 6.2 (IDENTIFIER LES SYNTAGMES DONT LA TÊTE$_r$ EST UN PRONOM)**
*Les pronoms de troisième personne sont encadrés d'une balise* `<sn>` *sauf s'ils précèdent immédiatement un élément identifié comme étant un verbe impersonnel (par exemple* `<verbe traits="impersonnel">`*semble*`</verbe>`*).*

***elle***

*(*`<sn><pro>`*elle*`</pro><sn>`*)* (4)

#### 6.2.2.2 Syntagmes dont la tête$_r$ est un nom propre

Nous avons relevé, à la section précédente, que les syntagmes contenant un nom propre ont des formes assez régulières. Nous croyons qu'ils peuvent être déterminés à l'aide de la règle suivante :

**RÈGLE 6.3 (IDENTIFIER LES SYNTAGMES DONT LA TÊTE$_r$ EST UN NOM PROPRE)**
*Un syntagme dont la tête$_r$ est un nom propre est constitué de toute suite de un ou plusieurs noms propres. Cette suite peut être précédée d'un déterminant ou même d'un déterminant suivi d'un mot quelconque.*

Par exemple :

*Caroline Serreau* (5-a)

*le Québec* (5-b)

*la compagnie ABB*

*(compagnie est un mot important du domaine des fusions de compagnies)* (5-c)

*la petite Marie*

*(petite n'est pas un mot du vocabulaire)* (5-d)

Dans le dernier cas, le syntagme est bien identifié même si *petite* n'est pas dans le dictionnaire. Tout ce que la règle demande est que ce qui peut se retrouver entre le déterminant et le nom propre soit un mot.

#### 6.2.2.3 Syntagmes dont la tête$_r$ est un nom commun associé au domaine

Tout comme les syntagmes référentiels dont la tête est un nom propre, les syntagmes dont la tête$_r$ est un nom commun associé au domaine ont des formes assez simples.

**RÈGLE 6.4 (IDENTIFIER LES SYNTAGMES DONT LA TÊTE$_r$ EST UN NOM COMMUN)**
*Un syntagme dont la tête$_r$ est un nom commun associé au domaine est constitué d'un mot identifié comme étant important (selon le domaine). Il peut aussi être précédé d'un (ou plusieurs) déterminant.*

*la* ***fille***

*(fille est un mot important du domaine des critiques de films)* (6-a)

*la* ***compagnie***

*(compagnie est un mot important du domaine des fusions de compagnies)* (6-b)

Notons que l'ordre dans lequel nous appliquons les règles n'est pas sans importance. Par exemple, *la compagnie ABB* sera traitée par la règle 6.3 page précédente car elle contient un nom propre. Par contre, *la compagnie* (tout court) sera traitée par la règle 6.4. Il faut donc ordonner 6.3 page précédente avant 6.4.

#### 6.2.2.4 Syntagmes comportant un déterminant démonstratif

Bien que nous ayons identifié plusieurs cas de syntagmes démonstratifs à l'étape précédente (règle 6.4), il reste encore un certain nombre de syntagmes démonstratifs dans les textes (c'est-à-dire les syntagmes ne comportant pas un nom associé au domaine). Pour identifier les cas restants, nous procédons ainsi :

RÈGLE 6.5 (IDENTIFIER LES SYNTAGMES COMPORTANT UN DÉTERMINANT DÉMONSTRATIF)
*Un syntagme est un syntagme démonstratif si le mot qui suit le déterminant démonstratif n'est pas un mot de la situation d'énonciation (cette année), ni un verbe (c'est) ni un pronom relatif (ce que). Dans les autres cas, le syntagme démonstratif est formé par le déterminant démonstratif suivi du mot qui le suit immédiatement.*

> ***Cette solution** est certes la plus dispendieuse mais elle donne les meilleurs résultats.*
>
> *(solution n'étant pas un nom commun associé au domaine)* (7)

#### 6.2.2.5 Syntagmes référentiels enchâssés

Jusqu'à maintenant, nous n'avons identifié que les syntagmes simples, c'est-à-dire qu'aucune de nos règles ne permet d'identifier les syntagmes référentiels enchâssés. Nous pouvons toutefois traiter les cas les plus simples :

RÈGLE 6.6 (IDENTIFIER LES SYNTAGMES RÉFÉRENTIELS ENCHÂSSÉS)
*Si deux syntagmes référentiels sont séparés par une préposition, le premier syntagme englobe le deuxième.*

> ***la fille** de **Coline Serreau***
>
> *(<sn>la fille <prep>de</prep> <sn>Coline Serreau</sn></sn>)* (8)

### 6.2.3 Traitements au niveau syntaxique II

Nous avons déterminé que certaines fonctions syntaxiques sont plus susceptibles d'être de bons candidats pour l'antécédent de certaines anaphores. Cependant, il est très difficile de déterminer les fonctions syntaxiques des éléments référentiels sans faire une analyse syntaxique au préalable. Nous pouvons toutefois avoir une bonne approximation des fonctions syntaxiques en considérant la position du syntagme dans la phrase et en vérifiant si une préposition précède le syntagme.

L'exactitude de l'attribution syntaxique n'est pas une priorité pour nous. Il aurait peut-être été préférable de pondérer les syntagmes référentiels selon leur position dans la phrase plutôt que de leur attribuer une fonction syntaxique approximative. Ainsi, le premier syntagme référentiel d'une phrase aurait reçu un bon pointage (disons 1 plutôt que `sujet`) alors qu'un syntagme référentiel enchâssé dans un autre aurait reçu un mauvais pointage (4 plutôt que `cNom`). En pratique, les mots (tels *sujet* ou *compléments du nom*) nous semblent plus évocateurs que des nombres.

Nous tenterons de déterminer quatre fonctions syntaxiques des syntagmes nominaux :

- les sujets ;
- les compléments du verbe ;
- les compléments du nom ;
- les appositions.

#### 6.2.3.1 Complément du nom

Un syntagme référentiel précédé par une préposition peut être un complément d'objet indirect (ou circonstanciel) ou un complément du nom. La règle précédente identifie déjà les cas de compléments du nom.

**RÈGLE 6.7 (IDENTIFIER LES COMPLÉMENTS DU NOM)**
*Un syntagme référentiel enchâssé dans un autre syntagme référentiel suivant l'application de la règle 6.6 page précédente est considéré comme étant un complément du nom.*

*la fille de* ***Coline Serreau*** (9)

#### 6.2.3.2 Complément d'objet indirect

Nous retrouvons ici les autres cas de syntagmes référentiels précédés par une préposition. Ces cas seront traités par la règle suivante. Nous considérons ces cas comme des compléments d'objet indirect.

**RÈGLE 6.8 (IDENTIFIER LES COMPLÉMENTS D'OBJET DIRECT)**
*Un syntagme référentiel non traité par 6.7 et précédé d'une préposition est un complément d'objet indirect.*

*appuyez sur* ***la touche F5*** (10)

#### 6.2.3.3 Apposition

Certains signes de ponctuation permettent d'encadrer les appositions (les parenthèses par exemple). Nous traitons ces cas à l'aide de la règle suivante :

**RÈGLE 6.9 (IDENTIFIER LES APPOSITIONS)**
*Un syntagme référentiel borné par des ponctuations encadrantes et précédé immédiatement par un autre syntagme référentiel est une apposition.*

*Cléopâtre (****la superbe Monica Bellucci****)* (11)

#### 6.2.3.4 Sujet

La fonction sujet est la fonction par défaut (et aussi la plus courante dans les textes). Nous attribuerons donc la fonction sujet à tous les autres syntagmes référentiels n'ayant pas déjà de fonction syntaxique.

**RÈGLE 6.10 (IDENTIFIER LES SUJETS)**
*Tout syntagme référentiel dont aucune fonction syntaxique n'a été attribuée par l'application des règles 6.7, 6.8 et 6.9 est un sujet.*

*Gatlif persiste et déçoit* (12)

Il est à noter que de cette façon, les fonctions de complément d'objet direct sont presque tous évaluées comme des sujets. Choisir de donner la fonction sujet par défaut à tous les syntagmes référentiels n'ayant pas reçu de fonction syntaxique n'est pas sans conséquence. Par exemple, à la règle 6.16 page 75, nous privilégions les sujets pour l'identification de l'antécédent. Cependant, en pratique, ceci ne semble pas avoir causé trop de problèmes.

### 6.2.4 Traitements au niveau textuel

À l'intérieur des textes, nous retrouvons quatre types de liens coréférentiels (section 2.2.1 page 20) :

- les coréférences ;
- les anaphores de reprise ;
- les anaphores lexicales ;
- les anaphores pronominales.

Chacun de ces liens coréférentiels nécessite une stratégie de solutionnement différente.

#### 6.2.4.1 Coréférence

Nous traitons deux cas de liens coréférentiels : les liens coréférentiels entre deux noms propres ainsi que les structures appositives.

**RÈGLE 6.11 (IDENTIFIER LES LIENS CORÉFÉRENTIELS ENTRE LES NOMS PROPRES)**
*Un syntagme contenant un nom propre $snpr_2$ est coréférentiel à un autre nom propre $snpr_1$ dans les cas suivants :*

- *si $snpr_1$ et $snpr_2$ sont des noms propres simples et sont identiques ($snpr_1$ = Arcand et $snpr_2$ = Arcand) ;*
- *si $snpr_1$ et $snpr_2$ sont des suites de noms propres et sont identiques ($snpr_1$ = Denys Arcand et $snpr_2$ = Denys Arcand) ;*
- *si $snpr_1$ est une suite de noms propres, $snpr_2$ est un nom propre simple et que cet élément est identique au dernier nom propre de $snpr_1$ ($snpr_1$ = Denys Arcand et $snpr_2$ = Arcand) ;*
- *si $snpr_1$ est un nom propre simple, $snpr_2$ est une suite de noms propres et que le deuxième élément de cette suite est identique à $snpr_1$ ($snpr_1$ = Arcand et $snpr_2$ = Denys Arcand).*

*Les syntagmes $snpr_1$ et $snpr_2$ sont non coréférentiels sinon.*

**RÈGLE 6.12 (IDENTIFIER LES LIENS CORÉFÉRENTIELS DES STRUCTURES APPOSITIVES)**
*Un syntagme dont la fonction est* `apposition` *(voir règle 6.9 page précédente) est coréférentiel au syntagme situé immédiatement avant celui-ci.*

*Cléopâtre$_1$ (**la superbe Monica Bellucci**$_{2,coref=1}$)* (13)

#### 6.2.4.2 Anaphore de reprise

Nous incluons parmi les anaphores de reprise les liens anaphoriques impliquant des réitérations (partielles ou exactes) des noms communs associés au domaine. Notons que pour tous les cas anaphoriques, s'il existe plus d'un antécédent possible, l'antécédent le plus près (mais situé avant) sera sélectionné.

**RÈGLE 6.13 (IDENTIFIER LES LIENS ANAPHORIQUES DE REPRISE)**
*Un syntagme contenant un nom commun associé au domaine $snDom_2$ est anaphorique à un autre syntagme $sn_1$ dans les cas suivants :*

- *si $snDom_2$ est un syntagme défini et que $snDom_2$ et $sn_1$ partagent une identité lexicale (La société $ABB_1$ ... la société$_{2,anaph=1}$) ;*
- *si $snDom_2$ est un syntagme démonstratif, que $snDom_2$ et $sn_1$ partagent une identité lexicale et que $snDom_2$ et $sn_1$ ne sont pas séparés par plus de deux phrases (La société $ABB_1$ ... cette société$_{2,anaph=1}$).*

*Le syntagme $snDom_2$ est non anaphorique au syntagme $sn_1$ sinon.*

#### 6.2.4.3 Anaphore lexicale

Les syntagmes contenant un déterminant démonstratif sont habituellement anaphoriques (voir 2.1.2 page 17). À la règle précédente 6.13, nous avons traité les cas des syntagmes démonstratifs contenant un nom associé au domaine. Ces cas sont traités comme étant des anaphores de reprise. Les autres cas sont pour nous des cas d'anaphores lexicales.

**RÈGLE 6.14 (IDENTIFIER LES LIENS ANAPHORIQUES LEXICAUX)**
*Un syntagme contenant un déterminant démonstratif est considéré possiblement anaphorique dans les cas suivants :*

- *s'il ne contient pas un nom commun marqué* `traits="domaine"` *(ces cas ont été traités par la règle 6.13) ;*
- *s'il ne contient pas un nom commun marqué* `traits="deictique"` *(voir 6.2.1.3 page 67).*

*Ce syntagme (nommé $snDm_2$) est anaphorique à un autre syntagme $sn_1$ si ce dernier contient un nom commun associé au domaine ou bien un nom propre et que $snDm_2$ et $sn_1$ sont compatibles en genre et en nombre*[8].

*Le syntagme est non anaphorique sinon.*

Par exemple :

> *Méfiez-vous d'un éventuel accaparement des évènements de saisie par* ***X11***$_1$ *si* ***ce dernier***$_{2,anaph=1}$ *fonctionne également sur le même écran.* (14)

[8]Nous n'expliquons pas les détails de la compatibilité en genre et en nombre pour les noms propres. Essentiellement, nous utilisons les déterminants présents dans la chaîne coréférentielle déjà existante pour déduire ces traits.

#### 6.2.4.4 Anaphore pronominale

Les syntagmes pronominaux sont traités en deux cas, selon qu'ils sont dans une proposition enchâssée (c'est-à-dire qu'il y a une conjonction, un mot relatif ou une ponctuation quelque part avant) ou au début de la phrase.

RÈGLE 6.15 (IDENTIFIER LES LIENS ANAPHORIQUES PRONOMINAUX ENCHÂSSÉS)
*Les syntagmes pronominaux en position enchâssée sont anaphoriques au premier sujet (qui n'est pas un pronom) de la phrase courante compatible en genre et en nombre. S'il n'existe pas de tel antécédent, il est anaphorique au premier complément d'objet indirect respectant les mêmes conditions.*

> *Alors que* ***cette fusion***$_1$ *n'aura aucun effet sur l'exploitation d'affaires de Téléglobe,* ***elle***$_{2,anaph=1}$ *permettra une plus grande flexibilité opérationnelle, une simplification des communications financières et l'optimisation de la structure financière, du fait que tous les instruments de dette seront regroupés sous Téléglobe, a précisé cette dernière.* (15)

RÈGLE 6.16 (IDENTIFIER LES LIENS ANAPHORIQUES PRONOMINAUX EN DÉBUT DE PHRASE)
*Les syntagmes pronominaux en début de phrase sont anaphoriques au premier syntagme sujet de la phrase précédente compatible en genre et en nombre. S'il n'existe pas de tel antécédent, il est anaphorique au premier complément d'objet indirect respectant les mêmes conditions.*

> *Lors d'une escapade un peu longue de la mère,* ***Julien***$_1$ *se voit obligé de traîner Elsa avec lui dans une excursion à la montagne.* ***Il***$_{2,anaph=1}$ *n'a que quelques jours pour capturer un papillon rare qu'il a promis d'apporter à un être cher.* (16)

## 6.3 Mesures d'efficacité

Pour tester la validité de nos règles, nous avons tout d'abord implanté l'algorithme d'analyse ci-dessus en utilisant le langage *XSLT* et l'avons appliqué aux fichiers du corpus d'entraînement. Nous n'avons certainement pas obtenu les résultats de la section précédente du premier coup. Nous avons dû faire certains ajustements avant d'obtenir l'ensemble final des règles. Au moment où nous avons jugé ces règles satisfaisantes, nous avons appliqué l'algorithme au corpus.de test.

Nous avons mesuré la validité de nos résultats selon les critères établis au chapitre précédent (voir section 5.3 page 58). Rappelons que la précision est calculée par $P = \frac{vp}{vp+fp}$ et que le rappel est mesuré par $R = \frac{vp}{vp+fn}$. Nous avons utilisé ces mesures pour évaluer la résolution des pronoms personnels de troisième personne de notre corpus. Les tableaux 6.1 page suivante et 6.2 page suivante présentent quelques chiffres.

Notons que ce que nous avons considéré comme étant un vrai positif est un pronom personnel bien résolu. Par contre, un faux positif est un pronom personnel à qui nous avons attribué

un mauvais antécédent. De plus, les vrais négatifs sont les pronoms pour lesquels il n'existe aucun antécédent identifié et qui ont été notés ainsi par l'algorithme (par exemple, le *il* impersonnel). Enfin, les faux négatifs sont les pronoms pour lesquels il existe un antécédent dans le texte mais pour lequel aucun antécédent n'a été identifié par l'algorithme.

TAB. 6.1 – MESURES D'EFFICACITÉ – RÉSOLUTION DES PRONOMS (CORPUS D'ENTRAÎNEMENT)

| | Fusions | Films | *HOWTO Linux* | Global |
|---|---|---|---|---|
| Précision | 0.50 | 0.58 | 0.56 | 0.55 |
| Rappel | 0.86 | 0.78 | 0.75 | 0.78 |

TAB. 6.2 – MESURES D'EFFICACITÉ – RÉSOLUTION DES PRONOMS (CORPUS DE TEST)

| | Fusions | Films | *HOWTO Linux* | Global |
|---|---|---|---|---|
| Précision | 0.42 | 0.9 | 0.55 | 0.59 |
| Rappel | 0.40 | 0.9 | 1 | 0.73 |

Sans être décevants, nous obtenons des résultats un peu inférieurs à ceux obtenus par d'autres algorithmes pour des textes anglais. Dans les travaux étudiés, les mesures de précision de certains algorithmes peuvent atteindre 0.6[9]. En fait, ces résultats sont peut-être un peu mieux que ce que nous espérions. Nous nous attendions à avoir un très faible rappel, car nous avons simplifié en enlevant beaucoup de cas problématiques. Fort heureusement, les cas problématiques ainsi éliminés n'étaient pas nombreux et n'ont pas influencé beaucoup les résultats.

Nous expliquons la faiblesse de notre précision par le fait que nous avons, contrairement aux autres algorithmes, très peu d'informations linguistiques (nous n'avons pas utilisé de dictionnaire, ni d'analyseur syntaxique). Il faut mentionner que notre algorithme n'est encore qu'à un stade d'ébauche. Il pourrait certainement être amélioré à l'aide d'une analyse plus poussée. Dans notre recherche, nous ne cherchions qu'à faire une étude de faisabilité des approches *knowledge-poor* communes à tous les types de textes. C'est pourquoi nous avons utilisé le même algorithme sur tous les types de textes. Il est probable qu'en utilisant les différences selon le type de texte, nous pourrions améliorer sensiblement les résultats. Nous reparlerons des différences selon le type de texte dans le reste de ce chapitre (voir 6.4 page suivante).

Nous avons très peu de textes et aussi très peu d'occurrences de pronoms dans chacun des textes rencontrés car les textes étudiés sont relativement courts (voir annexe H page liv). L'interprétation des chiffres de précision et de rappel sont donc à prendre avec un grain de sel car ils ne sont pas statistiquement significatifs. Par exemple, il est très étrange d'obtenir (dans certains cas) une meilleure précision ou un meilleur rappel dans le corpus de test que dans le corpus d'entraînement. Nous avons tout de même choisi d'inclure ces résultats dans ce mémoire pour donner une idée des performances obtenues. Toutefois, nous avons choisi de ne pas mesurer la performance de notre algorithme avec les mesures *B-CUBED*[10]. Avant d'accomplir

[9] Voir [BM01].

[10] Nous avons discuté des mesures *B-CUBED* à la section 5.3.2 page 59.

une telle tâche, nous croyons qu'il est nécessaire d'avoir un corpus plus important. Aussi, le choix des types de textes représente probablement une «histoire d'horreur» en ce qui concerne les chaînes coréférentielles d'un texte. Par exemple, dans les fusions de compagnies, certaines chaînes peuvent se combiner (la compagnie $A$ et la compagnie $B$ se fusionnent pour donner la compagnie $C$). Dans les critiques de films, certaines chaînes peuvent avoir une double identité (le nom de l'acteur et le nom du personnage). Ces cas un peu bizarres compliquent le calcul de précision et de rappel des chaînes de coréférence en entier car la notion de chaîne coréférentielle y est moins bien définie.

Il est aussi à noter que les mesures de précision et de rappel ne permettent pas de saisir l'ensemble des performances d'un algorithme. Nous avons mis quelques exemples de mauvais résultats de notre algorithme en annexe (voir F page xxxi) :

- Le premier texte (intitulé *invasions*[11]) démontre qu'il y a une confusion possible lorsque deux personnes portent le même nom de famille (ici *Denys Arcand* et *Gabriel Arcand*).
- Le deuxième texte (*Oracle-HOWTO-2*) illustre certaines difficultés pour l'identification de certains syntagmes. Par exemple, les numéros de sous-sections ont été inclus dans le syntagme *1.Préparation du Serveur*. Notons aussi que *un Utilisateur Oracle*, *Oracle*, *le noyau Oracle*, *l'Installateur Oracle sur le CD SCO* et *les fichiers Oracle de* ont tous étés considérés comme étant des coréférents, ceci est assez discutable (nous reparlerons de quelques-uns de ces cas à la section 7.4 page 89). Par contre, *l'utilisateur ORACLE*, à cause de sa forme en majuscule, n'a pas été identifié dans la même chaîne coréférentielle que le syntagme *un Utilisateur Oracle*.
- Le troisième texte nous montre qu'il est possible d'identifier une chaîne coréférentielle qui s'étend sur toute la longueur d'un texte. Cependant, nous remarquons que la chaîne n'est pas complète et qu'elle omet quelques pronoms qui devraient appartenir à cette chaîne.

## 6.4 Variations selon le type de texte

Bien que les textes français, en général, partagent beaucoup de ressemblances concernant le solutionnement des liens coréférentiels, nous ne pouvons cependant nier le fait qu'ils ont des particularités qui leur sont propres selon le type de texte. Essentiellement, nous observons les types de variations suivants :

- le vocabulaire utilisé ;
- les différences entre les éléments référentiels (types de syntagmes nominaux) ;
- les constructions particulières ;
- l'organisation des chaînes coréférentielles à l'intérieur d'un texte.

Il est à noter que même si nous étudions ici les textes dans le but de dégager les différences entre les types de textes, notre analyse pourrait permettre un travail orienté dans le sens inverse. Ainsi, certaines des constatations relevées dans la présente section pourraient servir à

[11] Ces textes sont les sorties *HTML* générées par notre algorithme. Le premier champ encadré correspond au nom du fichier traité et ne correspond donc pas nécessairement au titre de l'article.

déterminer à quel type de texte appartient un texte. Ceci pourrait être fait en étudiant le vocabulaire, les types d'éléments référentiels, les constructions particulières et les chaînes coréférentielles observées dans un texte donné.

### 6.4.1 Vocabulaire

Nous avons écrit, dans la première section de ce chapitre, que le vocabulaire varie selon les types de textes étudiés. La variation est particulièrement importante pour les noms communs. Puisque dans certains textes il est question surtout de personnes, dans d'autres d'objets ou de procédures, le choix du vocabulaire en sera certainement affecté.

Pour faire suite à l'analyse de fréquence sur les textes balisés (voir annexe G page xlviii), nous observons que nous ne retrouvons pas le même type de nom commun et de nom propre dans les trois textes. Nous constatons qu'il y a une variété de noms communs assez restreinte et qu'ils peuvent être regroupés en classes de noms communs. Par exemple :

**Fusions de compagnies :** Les noms communs présents à haute fréquence dans les fusions de compagnies sont les synonymes de *fusion* (*mariage, acquisition*) de *proposition* (*offre, accord*), de *compagnie* (*société, groupe*) et des mots associés au domaine des affaires (*actionnaire, conseil, direction*).

**Critiques de films :** Les noms communs présents dans les critiques de films sont les métiers (*acteur, superhéros*), les relations familiales (*père, mère*) et le vocabulaire du cinéma (*scénario, caméra*).

***HOWTO Linux* :** Les noms communs présents dans les *HOWTO Linux* sont tous les termes associés à l'informatique (*fichier, disque, serveur*).

Nous observons aussi une variabilité dans les référents associés aux noms propres dans les types de textes étudiés. Dans les fusions de compagnies, ce sont les entités non humaines qui dominent, suivi par les noms des lieux et les noms des entités humaines. Par contre, les entités humaines sont les éléments les plus présents des critiques de film, suivi des entités non humaines. C'est d'ailleurs le même genre de classification qui caractérise une bonne partie des noms communs associés au domaine, soit les relations familiales et les métiers que pratiquent les personnages. À cela, nous ajoutons les mots appartenant au domaine du film (caméra, spectateur, histoire, etc.) Enfin, les humains et les lieux sont presque absents des *HOWTO Linux*, laissant une grande place aux entités non humaines. Les noms communs associés au domaine jouent un rôle beaucoup plus important dans le cas des *HOWTO Linux*.

Remarquons qu'il est assez difficile de quantifier l'importance d'identifier les noms communs associés au domaine. Leur identification est assez fastidieuse mais nous croyons toutefois qu'ils sont très importants. Dans les textes étudiés, nous estimons qu'il y a environ une chaîne coréférentielle parmi les trois plus longues qui contiennent un (ou plusieurs) nom commun associé au domaine. Il arrive aussi que les syntagmes référentiels dont la $\text{tête}_r$ est un nom commun associé au domaine soient l'antécédent d'un pronom personnel. Dans ces cas, le pronom semble être bien résolu par l'algorithme.

### 6.4.2 Éléments référentiels

Les syntagmes démonstratifs, les syntagmes possessifs, et les pronoms se trouvent dans les trois sous-corpus étudiés. Cependant, selon le type de texte, nous remarquons que la façon d'employer ces syntagmes varie beaucoup notamment au niveau de la personne, du nombre et de l'appartenance aux chaînes de coréférence importantes. Par exemple, nous remarquons que la deuxième personne est très fréquente dans les *HOWTO Linux* car l'auteur s'adresse de façon plus directe à ses destinataires. Pour les critiques de films et les fusions de compagnies, c'est plutôt la troisième personne qui est privilégiée mettant à l'avant-plan l'aspect plus descriptif de ces textes. Notons toutefois la présence de l'utilisation de la première personne de ces textes journalistiques qui permet à l'auteur de se rapprocher de ses lecteurs dans les critiques de films ou bien qui rapporte les paroles de quelqu'un dans les fusions de compagnies. De plus, les critiques de films comportent beaucoup de déterminants possessifs. Enfin, dans les *HOWTO Linux*, les noms propres définis de la façon donnée ci-dessus sont très peu fréquents. Remarquons toutefois que certains éléments comme les noms de programmes, les commandes, les bouts de codes informatiques, etc. se comportent comme des noms propres (par exemple, *le fichier* ***zut.doc***). Cette différence est particulièrement marquée au niveau de leurs comportements vis-à-vis des déterminants et des compléments (voir aussi section 7.4.5 page 91). Enfin, les *HOWTO Linux* utilisent fréquemment des références associées à la situation d'énonciation et de verbes impersonnels.

Ces différences entre les types de syntagmes utilisés pourraient influencer l'importance des règles que nous avons développées. Nous avons dit, dans la section précédente, que l'ordre d'application des règles est important. De plus, dans certains cas, la distance entre un élément et son antécédent peut être plus grande dans certains types de texte. Par exemple, une étude menée par [Kit82] semble démontrer que dans les contes pour enfants, cette distance peut être un peu plus grande.

### 6.4.3 Constructions particulières

Puisque chaque type de texte a des objectifs différents, il se peut que certains utilisent des «raccourcis» pour faire l'économie de constructions pouvant devenir assez lourdes sinon. Nous avons parlé des «vides» que comportent les recettes de cuisine (voir 1.2.2 page 7). Lors de l'analyse des textes de notre corpus, nous avons constaté qu'il y a aussi des formes raccourcies. Ces formes sont assez diversifiées mais semblent plutôt **stables dans un type de texte donné**. Il est à remarquer qu'il est peu probable de retrouver une phrase comme (19) page suivante dans un *HOWTO Linux*. Pourtant, ce genre de construction abonde dans les critiques de films.

Examinons quelques-unes de ces constructions que nous avons relevées :

**Fusions de compagnies :**

- Dans les fusions de compagnies, il arrive fréquemment que les chaînes coréférentielles, à l'image des compagnies qu'elles représentent, soient fusionnées dans le texte. Une

construction assez productive pour exprimer cette agglomération dans le texte est regrouper les deux compagnies par un tirait (*Pétro-Canada-Ultramar*). Par contre, dans d'autres textes, si une compagnie est sensiblement plus grosse, il peut arriver qu'elle absorbe la deuxième compagnie. Dans ces cas, il est plutôt difficile de distinguer la différence de la compagnie avant et après la fusion. Dans ce cas, la compagnie conserve habituellement le même nom.

- L'utilisation de métonymies est très fréquente dans ces textes.

  *La compagnie Adherex a annoncé qu'elle fusionnait.* (17)

- Le discours direct est aussi très courant.

  *«Nous venons de conclure une première fusion, mais ce ne sera pas la dernière», lance M. Grey.* (18)

**Critiques de films :** Dans ces textes, nous retrouvons souvent des structures de la forme $X$ $(Y)$ en $Z$ où $X, Y$ et $Z$ réfèrent à la même entité mais sous divers aspects.

***En pilote de brousse et pompier volant, Marc (Jean-Nicolas Verreault)** incarne avec brio cette nouvelle vision du mâle québécois, à la fois viril et sensible.* (19)

Bien entendu, cette expression peut être décomposée. Par exemple, nous retrouvons très souvent la structure simplifiée $X$ $(Y)$ (comme dans *Cléopâtre (la superbe Monica Bellucci)*).

*HOWTO Linux* **:** Les *HOWTO Linux* ressemblent un peu aux recettes de cuisine car ils sont aussi des textes procéduraux. Nous y retrouvons donc aussi les «vides» que l'on retrouve dans ces textes. De plus, la mise en forme est très importante et contient une certaine information référentielle. Par exemple, les bouts de code informatique peuvent parfois être considérés en un tout et être repris anaphoriquement. Notons que les bouts de code informatique sont habituellement introduits par « :». Nous avons soulevé à la section 1.1 page 3 que ce signe introduit bien souvent une cataphore.

### 6.4.4 Chaînes coréférentielles

Nous introduisons ici une nouvelle notion : celle de **chaîne importante d'un texte**. Il peut sembler un peu circulaire de décrire ce qu'est une chaîne importante seulement après les avoir identifiées dans l'étape de balisage manuel du corpus. Nous n'avions, au moment du balisage, qu'une idée intuitive de ce qui nous paraissait être une chaîne coréférentielle importante. Après le balisage, nous avons maintenant les ressources pour donner une définition productive d'une chaîne importante, c'est-à-dire de définir un mécanisme permettant de les identifier. Dans le chapitre *Travaux précédents* (section 3.2.5 page 32), nous avons invoqué quelques propriétés de certaines chaînes coréférentielles importantes d'un texte :

- ces chaînes sont habituellement plus faciles à résoudre ;
- elles contiennent les informations importantes d'un texte ;
- leur vocabulaire est plus restreint.

Le concept de chaîne importante est plus ou moins flou, il correspond grosso modo à ce que S. Bergler a présenté dans [Ber97].

DÉFINITION 6.1 (CHAÎNE CORÉFÉRENTIELLE IMPORTANTE D'UN TEXTE)
*Les chaînes coréférentielles importantes d'un texte sont des chaînes coréférentielles qui sont proportionnellement longues par rapport aux autres chaînes du texte, qui ont une distribution qui s'étend sur la longueur du texte et dont le premier élément (la tête de la chaîne coréférentielle) est habituellement dans le titre ou les deux premières phrases du texte.*

Les chaînes importantes réfèrent souvent aux entités qu'on s'attend à rencontrer dans un texte (sachant le type de texte). Par exemple, dans les textes de fusion de compagnies, on s'attend au moins à avoir les deux compagnies (parfois plus) avant la fusion ainsi que la compagnie résultante de la fusion. Ces chaînes contiennent souvent beaucoup de cas anaphoriques, dont beaucoup d'anaphores lexicales.

On peut aussi considérer, de façon un peu moins importante, les référents associés aux référents principaux du texte. Dans le cas des fusions de compagnies, il peut s'agir du président de la compagnie, des employés ou des actionnaires. Nous pouvons considérer ces derniers cas comme référents des **chaînes coréférentielles associées aux chaînes importantes**. Leur distribution s'étend bien souvent sur un seul paragraphe mais il peut arriver que ces chaînes soient plus étendues.

Enfin, les autres chaînes coréférentielles du texte sont beaucoup plus courtes ; ce sont typiquement des éléments référentiels isolés (des chaînes de longueur 1) ou pouvant contenir un élément anaphorique situé à l'intérieur de la même phrase (anaphore syntaxique ou quelques cas d'anaphores pronominales).

Empiriquement, dans les textes de notre corpus, nous retenons environ trois ou quatre chaînes que nous jugeons importantes ou associées à une chaîne importante par texte. Pour sélectionner ces chaînes, il suffit, dans bien des cas, d'ordonner ces chaînes selon leur longueur et de sélectionner les trois ou quatre chaînes les plus longues. Nous croyons que l'identification de ces chaînes est très importante pour la recherche d'informations. Ces chaînes peuvent aider à déterminer les entités importantes d'un texte et à préciser quelques éléments importants à propos de ces entités.

Dans bien des types de texte, il arrive souvent que les chaînes importantes soient chapeautées par des noms propres car ils sont des référents importants. Par contre, pour les types de textes où les entités importantes ne sont pas nécessairement des personnes, il faut identifier les entités à l'aide des mots du domaine ou de classes de mots.

Les chaînes coréférentielles sont le squelette du texte. Nous pouvons tenter de caractériser les chaînes appartenant à un type de texte donné. Par exemple, les chaînes présentes dans les fusions de compagnies sont habituellement longues et prévisibles. Par contre, dans les *HOWTO-Linux*, les chaînes sont relativement courtes et ont une étendue se limitant à quelques paragraphes. Pour aider à caractériser les chaînes coréférentielles d'un texte, voici quelques

questions auxquelles il faut répondre[12] :

**Les chaînes sont-elles en parallèle ou en série dans le texte ?** Lorsqu'il y a plusieurs chaînes importantes dans le texte, elles peuvent parfois être en parallèle dans le texte. C'est le cas normalement des chaînes représentant le film ou le metteur en scène dans les critiques de films. Ces chaînes sont habituellement introduites dans les premières phrases du texte et sont aussi rappelées vers la fin du texte. C'est un peu la même chose qui se produit dans les fusions de compagnies. Dans ces textes, les chaînes représentant la compagnie importante sont présentes tout au long de la lecture du texte. Les autres chaînes associées à ces textes peuvent apparaître à l'occasion selon le thème abordé. Par contre, les *HOWTO-Linux* ont plutôt une structure en série car les chaînes coréférentielles sont les unes à la suite des autres.

**Quelle est la distribution des chaînes ?** Retrouvons-nous des éléments de ces chaînes dans tous les paragraphes du texte ? Par exemple, il est fréquent que certaines chaînes présentes dans les *HOWTO-Linux* soient concentrées dans un seul paragraphe. Dans d'autres textes, les premier et dernier paragraphes peuvent avoir un statut particulier. Par exemple, les critiques de films sont particulièrement sensibles à la division du texte en paragraphes. Les paragraphes se trouvant au début et à la fin du texte contiennent une concentration assez importante d'éléments référentiels.

**De quelle longueur sont les chaînes ?** Un critère empirique jugeant l'importance d'une chaîne coréférentielle peut très certainement être basé sur la longueur des chaînes coréférentielles. Dans les fusions de compagnies, nous remarquons qu'il existe souvent une chaîne proportionnellement plus longue que les autres.

## 6.5 Conclusion

Certains phénomènes coréférentiels sont assez simples à résoudre. Ils semblent en même temps plus ou moins communs à tous les types de textes rencontrés. Nous avons fait un algorithme qui utilise un vocabulaire simple (plus ou moins prévisible). Les résultats sont plus ou moins satisfaisants mais suggèrent qu'il est essentiel de prendre en considération les différences entre les types de textes pour améliorer les performances.

Ces différences sont orientées selon les axes suivants :

- le vocabulaire utilisé ;
- les différences entre les éléments référentiels (types de syntagmes nominaux) ;
- les constructions particulières (assez diversifiées mais plutôt stables dans un type de texte donné) ;
- l'organisation des chaînes coréférentielles à l'intérieur d'un texte.

Ces différences pourraient aussi être traitées dans le sens inverse et servir à déterminer à quel type de texte appartient un texte. Voici une brève comparaison sommaire des types de textes que nous avons étudiés :

[12] Les exemples présentés à l'annexe C page xv illustrent bien les points que nous développons ici.

**Fusions de compagnies :**

- Utilisation du trait d'union pour fusionner deux chaînes ;
- utilisation du discours direct ;
- utilisation de métonymies ;
- les chaînes coréférentielles sont longues et prévisibles ;
- certaines chaînes peuvent être en parallèle dans le texte.

**Critiques de films :**

- Beaucoup de cas possessifs ;
- les noms communs sont très utilisés dans les chaînes importantes du texte ;
- utilisation des constructions $X$ `(`$Y$`)` `en` $Z$ ;
- certaines chaînes peuvent avoir une double identité (le nom de l'acteur et le nom du personnage) ;
- le texte peut comporter plusieurs chaînes en parallèle ;
- les premier et dernier paragraphes sont plus importants (concentration assez importante d'éléments référentiels).

*HOWTO Linux* **:**

- Utilisation du « :» pour introduire des bouts de code informatique ;
- utilisation de la deuxième personne ainsi que du pronom *le* en position C.O.D. ;
- certains mots se comportent comme des noms propres (par exemple les noms de programmes) mais n'ont pas nécessairement de majuscule ;
- très peu d'entités humaines ;
- les chaînes sont en série ;
- beaucoup de chaînes sont concentrées sur un seul paragraphe ;
- les chaînes sont relativement courtes et beaucoup de ces chaînes sont de longueur 1.

# Chapitre 7

# Discussion

Dans cette étude, nous avons déterminé les facteurs théoriques influençant les liens coréférentiels d'un texte (section 3.3 page 33). Nous avons ensuite étudié ces facteurs en contexte en observant les chaînes importantes d'un corpus comportant trois types de textes. Enfin, nous avons fait l'ébauche d'un algorithme simple permettant d'identifier quelques-uns des éléments référentiels des chaînes importantes d'un texte (section 6.2 page 65).

Il y a certainement encore beaucoup de réajustements à apporter à notre algorithme avant d'obtenir des résultats qui soient utilisables dans un contexte réel (d'entreprise, par exemple). Cependant, avant d'aller plus loin dans la résolution des liens coréférentiels selon des approches *knowledge-poor*, il serait bien de prendre un peu de recul vis-à-vis de notre méthode et de nos résultats pour examiner le phénomène dans ses grandes lignes afin d'avoir une meilleure vue d'ensemble sur le problème.

Dans ce chapitre, nous analyserons sommairement les problèmes rencontrés dans notre étude et nous discuterons de certains choix que nous avons faits. Nous analyserons aussi les points forts et les points faibles de notre approche. Nous proposerons ensuite quelques solutions pour améliorer les résultats. Enfin, nous donnerons quelques pistes pour aller plus loin.

## 7.1 Problèmes rencontrés

Avant d'entreprendre notre recherche, nous nous étions fixé quelques buts. Notre hypothèse générale était qu'il est possible d'établir les chaînes de coréférence automatiquement avec une stratégie générale qui prévoit certaines étapes d'adaptation au vocabulaire et aux structures syntaxiques spéciales du sous-langage. Compte tenu de la restriction des ressources linguistiques à notre disposition (dictionnaire, analyseur syntaxique, corpus), nous avons dû faire certains compromis. Voici quelques-unes des solutions que nous avons envisagées pour réduire la complexité du problème :

# Chapitre 7

# Discussion

Dans cette étude, nous avons déterminé les facteurs théoriques influençant les liens coréférentiels d'un texte (section 3.3 page 33). Nous avons ensuite étudié ces facteurs en contexte en observant les chaînes importantes d'un corpus comportant trois types de textes. Enfin, nous avons fait l'ébauche d'un algorithme simple permettant d'identifier quelques-uns des éléments référentiels des chaînes importantes d'un texte (section 6.2 page 65).

Il y a certainement encore beaucoup de réajustements à apporter à notre algorithme avant d'obtenir des résultats qui soient utilisables dans un contexte réel (d'entreprise, par exemple). Cependant, avant d'aller plus loin dans la résolution des liens coréférentiels selon des approches *knowledge-poor*, il serait bien de prendre un peu de recul vis-à-vis de notre méthode et de nos résultats pour examiner le phénomène dans ses grandes lignes afin d'avoir une meilleure vue d'ensemble sur le problème.

Dans ce chapitre, nous analyserons sommairement les problèmes rencontrés dans notre étude et nous discuterons de certains choix que nous avons faits. Nous analyserons aussi les points forts et les points faibles de notre approche. Nous proposerons ensuite quelques solutions pour améliorer les résultats. Enfin, nous donnerons quelques pistes pour aller plus loin.

## 7.1 Problèmes rencontrés

Avant d'entreprendre notre recherche, nous nous étions fixé quelques buts. Notre hypothèse générale était qu'il est possible d'établir les chaînes de coréférence automatiquement avec une stratégie générale qui prévoit certaines étapes d'adaptation au vocabulaire et aux structures syntaxiques spéciales du sous-langage. Compte tenu de la restriction des ressources linguistiques à notre disposition (dictionnaire, analyseur syntaxique, corpus), nous avons dû faire certains compromis. Voici quelques-unes des solutions que nous avons envisagées pour réduire la complexité du problème :

**Réduction du prétraitement syntaxique :** Pour résoudre parfaitement tous les liens coréférentiels d'un texte, il semble qu'une analyse syntaxique complète (et parfaite) du texte soit nécessaire. Malheureusement, nous ne pouvons pas obtenir une telle analyse de façon automatique avec les outils dont nous disposons. C'est pourquoi nous avons préféré ne pas utiliser d'analyseur syntaxique dans le prétraitement de notre corpus. Ainsi, le seul prétraitement nécessaire à notre algorithme est l'identification des frontières des paragraphes et des phrases. Ces étapes d'identifications étant plus simples, elles peuvent être automatisées plus facilement.

**Réduction du nombre de syntagmes référentiels traités :** De façon générale, nous avons simplifié le problème d'identification des items référentiels en ne traitant que les syntagmes référentiels les plus courants. C'est pourquoi nous avons porté une attention particulière aux noms propres et aux pronoms personnels de troisième personne. Les ellipses ont été éliminées d'emblée car elles sont très difficiles à identifier. Pour compléter les chaînes coréférentielles contenant un nom propre ou un pronom, nous avons misé sur le fait que plusieurs des syntagmes référentiels s'y référant (possiblement les plus «importants») contiennent un nom associé au domaine. Cette supposition est appuyée par l'analyse d'un corpus que nous avons balisé manuellement.

**Simplification de l'analyse syntaxique :** Beaucoup de liens anaphoriques sont dépendants de la syntaxe. Notamment, les pronoms réfléchis (*lui-même*) et les pronoms relatifs. Cependant, nous nous sommes plutôt intéressés aux liens anaphoriques (et coréférentiels) dépassant les frontières de la phrase. Le traitement exclusif de ces liens a permis de réduire l'analyse syntaxique nécessaire à l'identification des syntagmes référentiels seulement. Nous avons donc choisi d'inclure, dans notre algorithme, une étape permettant d'identifier sommairement les syntagmes référentiels que nous jugions importants ainsi qu'une identification approximative de leurs fonctions syntaxiques. Pour nous aider à accomplir une telle tâche, nous avons utilisé les éléments syntaxiques déjà présents dans le texte. Ainsi, nous avons utilisé les mots grammaticaux (déterminants, prépositions, conjonctions, etc.), les signes de ponctuation ainsi que les positions de ces éléments dans la phrase.

**Simplification de l'identification des liens coréférentiels :** Pour simplifier la résolution des liens coréférentiels d'un texte, nous avons choisi de ne traiter que les liens coréférentiels impliquant l'identité des référents. Dans bien des cas, le lien peut être déduit grâce à la réitération de certains mots (comme les noms propres ou les noms communs). Dans d'autres cas, le lien peut s'établir en examinant les autres syntagmes référentiels dans un contexte rapproché. En plus de la proximité, des contraintes de compatibilité des traits syntaxiques et sémantiques doivent être satisfaites.

Suivant ces simplifications, nous avons élaboré, dans le chapitre précédent, un système de règles formelles permettant d'établir certaines chaînes coréférentielles se trouvant à l'intérieur d'un texte (voir section 6.2 page 65). Par la suite, nous avons fait une implantation de ces règles en utilisant le langage *XSLT* et l'avons appliqué à l'ensemble des textes du corpus (d'entraîne-

ment et de test). Cet algorithme est très simple[1] mais permet néanmoins de traiter un grand nombre de cas anaphoriques et coréférentiels.

Nous avons discuté, à la section *Mesures d'efficacité* (voir 6.3 page 75), des résultats obtenus à partir de l'algorithme. Nous convenons qu'ils ne sont peut-être pas exceptionnels. Cependant, compte tenu des ressources très limitées que nous avons utilisées, nous sommes toutefois étonnés de ce que nous avons réussi à accomplir. L'algorithme n'est pas, en tant que tel, une fin en soi. Il est plutôt un moyen simple de traiter les cas coréférentiels les plus simples et communs à tous les types de textes. Pour un traitement plus large des liens coréférentiels d'un texte, nous croyons qu'il est essentiel de prendre en considération les différences selon le type de texte (voir section 6.4 page 77).

Dans ce mémoire, nous n'avons qu'effleuré les variations des moyens coréférentiels entre les types de textes. Nous n'avons étudié que trois types de textes différents sans proposer explicitement de règles pour résoudre les différences coréférentielles. Cependant, nous avons relevé quelques problèmes pour chacun des types de textes traités. Les problèmes les plus flagrants sont pour nous l'identification des syntagmes référentiels. En voici quelques exemples :

**Fusions de compagnies :** Certains syntagmes référentiels sont complexes, ils peuvent être composés de sigles, de noms de titres et des compléments du nom complexes.

*le ministère de l'Agriculture des Pêcheries et de l' Alimentation (MAPAQ)* (1-a)

*le Conseil des productions végétales du Québec (CPVQ)* (1-b)

*les Bourses américaines* (1-c)

*le Bureau canadien de la concurrence* (1-d)

*la Commission commerciale fédérale* (1-e)

*la Chambre de commerce régionale des entrepreneurs de Québec* (1-f)

*le Tribunal de la concurrence* (1-g)

**Critiques de films :** Dans les critiques de films, c'est essentiellement l'identification des titres de films qui cause problème. Ces titres sont composés de syntagmes complexes, ils peuvent contenir un verbe conjugué ou un signe de ponctuation et ils sont parfois écrits dans une autre langue que le français.

*Trois Hommes et un couffin* (2-a)

*The Fast and the Furious* (2-b)

*Astérix et Obélix : Mission Cléopâtre* (2-c)

*L'Auberge espagnole* (2-d)

*Chacun cherche son chat* (2-e)

*The Hunt for Red October* (2-f)

*le Meurtre de Roger Ackroyd* (2-g)

$8\frac{1}{2}$ (2-h)

*E le nave va* (2-i)
*Il ne faut pas parier sa tête avec le diable* (2-j)
*The Book of Eve* (2-k)
*Crimes and Misdemeanors* (2-l)
*les Dix Petits Nègres* (2-m)
*les Invasions barbares* (2-n)
*Love and Human Remains* (2-o)
*It Runs in the Family* (2-p)
*Raiders of the Lost Ark* (2-q)
*La Turbulence des fluide* (2-r)

***HOWTO Linux*** : Le traitement des *HOWTO Linux* fut très problématique pour nous. Un des problèmes étant l'identification des noms de programmes, de commandes ou des messages d'erreurs. Ces entités se comportent comme des noms propres mais n'ont pas nécessairement de signes distinctifs (comme la majuscule). Remarquons toutefois que ces expressions comportent parfois des nombres ou des signes de ponctuation qui pourraient aider à les identifier.

*lancez* ***gcc -v*** (3-a)
*un* ***make*** *devrait compiler* ***gcc*** *correctement*
*(ici,* ***un*** *pourrait être optionnel)* (3-b)
*Dans* ***../html/****, vous devez trouver un fichier* ***index.htm*** (3-c)
*Un cat /proc/cpuinfo /proc/meminfo* (3-d)
*Esc-x doctor* (3-e)
*Press A Key To Begin Test* (3-f)
*configure -target=i486-linux -host=XXX* (3-g)
*aha152x=port,irq,scsi_id,1* (3-h)
*/etc/init.d/proftpd start ** (3-i)
*rpcinfo : can't contact portmapper : RPC : Remote system error -* (3-j)
*Connection refused , RPC_PROG_NOT_REGISTERED* (3-k)
*un nouveau sous-répertoire Linux-(votre version)-votre cpu* (3-l)

Ces textes contiennent aussi des renvois vers d'autres documents. Par ailleurs, lors de la transformation de ces textes en format texte brut, les hyperliens ont été remplacés par des crochets.

*Le [8] Linux Kernel HOWTO* (4-a)
*Le [7] Guide d'installation de Linux (Linux Installation HOWTO)* (4-b)
*le site web [9] www.3dfx.co* (4-c)
*l' adresse [13] http ://lemures.shinma.symix.com/ zureal/cdu31a.html* (4-d)

Bien entendu, certaines de ces difficultés pourraient être contournées en tenant compte du balisage préexistant (pour les documents *HTML*) ou de la typographie (par exemple l'utilisation de caractères gras ou italiques).

[1] L'algorithme tient en moins de 1500 lignes de code *XSLT*.

## 7.2 Points forts de la démarche

Au début de notre recherche, nous avions quelques attentes vis-à-vis des règles que nous voulions inclure dans notre algorithme (voir sections 1.2.5 page 9 et 3.1 page 26). Voici quelques caractéristiques de ces règles :

- être simples ;
- ne pas nécessiter de dictionnaire ou de système de connaissances complexe ;
- ne pas nécessiter de prétraitement complexe ;
- s'appliquer à un grand nombre de cas ;
- donner le moins possible de mauvais résultats.

Nous croyons avoir en grande partie répondu à ces attentes. La simplicité de notre algorithme constitue selon nous un point distinctif important de notre recherche. En effet, à l'aide d'un dictionnaire ne contenant que quelques centaines de mots (en grande partie, un vocabulaire constitué de mots appartenant à des classes de mots fermées) et d'un algorithme constitué de moins de 1500 lignes de code, nous avons réussi à résoudre un nombre important de liens coréférentiels. Le seul prétraitement nécessaire pour l'application de cet algorithme est essentiellement l'identification des frontières des phrases. Pour réussir cet exploit, nous avons exploité les informations déjà présentes dans le texte. Par exemple, nous avons identifié les noms propres à l'aide de la majuscule et nous avons déterminé les bornes des syntagmes référentiels grâce aux déterminants et aux prépositions. De plus, le genre et le nombre des syntagmes référentiels ont pu être déduits en utilisant les déterminants et les chaînes coréférentielles déjà existantes.

En principe, les règles que nous avons mises au point peuvent s'appliquer à tout type de textes. Bien que nous n'ayons pas formellement vérifié cette hypothèse, nous pouvons au moins espérer appliquer ces règles à à peu près n'importe quel texte spécialisé. Nous expliquons la grande couverture de nos règles grâce à leur caractère général. En effet, la réitération d'une $\text{tête}_r$, la distance entre un antécédent et une anaphore ainsi que la compatibilité en genre et en nombre sont des critères assez stables pour les textes de la langue en général. Ce sont aussi les critères que nous avons privilégiés pour l'établissement des liens coréférentiels de notre algorithme.

Indirectement, en étudiant les moyens coréférentiels dans des textes du français, nous avons été confrontés au problème de la représentation des chaînes coréférentielles à l'intérieur d'un texte. Nous avons aussi dû trouver des moyens pour manipuler ces chaînes à l'aide d'outils informatiques. Le chapitre *Outils XML* (chapitre 4 page 35) présente le cheminement que nous avons emprunté pour schématiser les chaînes coréférentielles à l'intérieur d'un corpus. Pour ce faire, nous avons utilisé *XML* et les autres langages qui lui sont liés.

## 7.3 Points faibles de la démarche

Notre besoin de simplicité n'est malheureusement pas sans conséquence sur les résultats obtenus. Nous retenons trois types de problèmes possibles engendrés par notre approche :

**Simplifications excessives :** Les simplifications entraînent un problème d'exhaustivité des résultats. Nous n'avons traité que certains syntagmes référentiels, laissé de côté les références introduites par les verbes et les propositions et n'avons traité que quelques types d'anaphores. Par exemple, nous n'avons identifié que quelques pronoms et quelques déterminants. De plus, nous n'avons pas identifié les pronoms possessifs, les ellipses et les pronoms interrogatifs. Nous croyons cependant qu'il est possible d'augmenter le nombre de cas couverts à l'aide de procédés similaires.

**Travail préalable laborieux :** Il y a assez de travail à faire pour traiter un nouveau type de texte. La nature de ce travail est essentiellement l'identification du vocabulaire et l'identification des structures anaphoriques variant selon le type de texte. Pour faire ce travail, il est nécessaire de consulter un très grand nombre de textes. Cependant, avec l'émergence d'ontologies formalisées (donc, lisibles par un ordinateur) portant sur plusieurs domaines spécialisés, cette faiblesse risque de s'estomper grandement[2].

**Performances :** Dans ce mémoire, nous n'avons pas pris en compte les contraintes de performances informatiques :

- L'algorithme, tel que présenté, nécessite le parcours du texte plusieurs fois.
- *XSLT* n'est pas (encore) un langage très efficace. Le traitement des chaînes coréférentielles avec *XSLT* n'est donc pas très rapide et est difficilement utilisable à grande échelle (quelques minutes pour traiter 60 textes)[3].

Ces considérations sont importantes pour l'utilisation d'un algorithme dans un but concret.

## 7.4 Améliorations possibles

Nous croyons qu'il est certainement possible d'améliorer les résultats obtenus. Voici quelques points qu'une recherche ultérieure portant sur la résolution des liens coréférentiels peut emprunter :

### 7.4.1 Construction du vocabulaire

Au chapitre précédent (section 6.2.1 page 65), nous avons exposé la façon dont nous avions élaboré notre liste de mots appartenant à notre dictionnaire. Cette liste de mots a été obtenue de façon plus ou moins *ad hoc* en consultant les textes du corpus et les grammaires du français. Pourtant, le vocabulaire ainsi trouvé est une ressource-clé pour le reste de l'élaboration de

[2] Voir aussi la section suivante concernant les améliorations possibles.

[3] Cependant, *XSLT* peut être compilé (voir `http://www.developer.com/xml/article.php/721241`). Donc, les performances pourraient, en milieu réel, être améliorées très facilement.

notre algorithme. Il est donc probable qu'une construction plus minutieuse pourrait contribuer à améliorer les résultats obtenus.

Bien que l'identification des mots grammaticaux (classes fermées) fut une étape laborieuse de notre recherche, la liste de mots ainsi obtenue peut être réutilisée pour tout autre type de texte. Par contre, la liste des noms communs associés au domaine est appelée à changer selon le domaine[4]. Pour relever ces noms communs, nous avons analysé leur fréquence dans les textes étudiés. Cette façon de faire n'est pas la seule. Une ontologie ou un dictionnaire terminologique peuvent aussi être des outils utiles pour cette tâche. Une autre méthode peut être l'utilisation des méthodes stochastique pour l'identification du vocabulaire «anormalement fréquent» dans différents types de textes (par rapport à la langue en général)[5]. Cependant, il se peut que la difficulté de l'identification des mots importants selon le domaine puisse dépendre de la fermeture du domaine.

### 7.4.2 Longueur des chaînes existantes

Statistiquement, si une chaîne coréférentielle est longue, il est plus probable qu'elle contienne un élément qui est l'antécédent d'un autre élément. Il serait donc avantageux d'ajouter ceci dans notre système de règles. D'ailleurs, nous ne croyons pas que ceci apporte beaucoup de complexifications.

Dans ce mémoire, nous avons travaillé avec des textes relativement courts. Cependant, pour des textes plus longs, il se peut que la distance et l'étendue des chaînes puissent influencer les règles de coréférences. Nous croyons qu'une étude plus approfondie des structures des chaînes coréférentielles puisse aussi être utile pour améliorer les résultats.

### 7.4.3 Mise en forme du texte

Bien que les textes soient lus de façon linéaire, ils comportent tout de même des éléments qui peuvent donner de la dimension à un texte. Par exemple, les listes, les sections et les paragraphes sont des éléments qui peuvent servir à structurer un texte. Nous n'avons pas exploité ces éléments mais ils influencent certainement la référence dans les textes.

De plus, lors de la mise en forme d'un texte, l'auteur a choisi de faire ressortir plusieurs éléments du texte soit par une typographie différente, la présence de titres, de sous-titres et même les urls référant à d'autres documents. Nous croyons que ces éléments renferment beaucoup d'informations qui puissent nous aider à résoudre certaines anaphores. Nous avons dû laisser de côté ces détails typographiques à cause d'une trop grande variabilité selon les auteurs mais tenir compte de la mise en forme du texte (ainsi que des indices typographiques d'un texte) aurait certainement aidé à dénouer certains liens coréférentiels.

[4]Nous avons parlé de la variation du vocabulaire des textes du corpus à la section 6.4.1 page 78.

[5]Voir [MS99] pour plus de détails concernant l'utilisation des méthodes statistiques appliquées au traitement automatique de la langue.

### 7.4.4 Nombre de textes et de domaines traités

Nous avons fait une très petite analyse. Les chiffres avec lesquels nous avons travaillé sont très petits ; trois types de textes est réellement trop peu pour être représentatif de tous les types de textes français. Il est certain qu'un plus grand nombre de textes et de domaines traités sont nécessaires pour une étude plus poussée.

### 7.4.5 Constructions `det nom_commun npr`

Les constructions du type `det nom_commun npr` se retrouvent occasionnellement dans les textes que nous avons étudiés. Ces constructions sont surtout présentes dans les *HOWTO Linux*. Nous croyons qu'une partie des mauvais résultats obtenus dans les *HOWTO Linux* est due à un mauvais traitement de ces syntagmes.

Bien que les constructions `det nom_commun npr` aient la même forme de surface, elles semblent se partager en trois catégories[6] : nominalisation, attribution[7] et hyponymie.

**Nominalisation :** Ces productions semblent être possibles pour les noms communs en général. Elles permettent de nommer (baptiser) un objet pour la première fois. Ces constructions devraient se retrouver, en principe, exclusivement en tête de chaîne coréférentielle.

*la fusion Air Alliance-Air Nova* (5-a)

*l' affiliation Petro-Canada-Ultramar* (5-b)

*le tandem Pétro-Canada-Ultramar* (5-c)

*le nom peu intuitif «/usr »* (5-d)

*le nom YP* (5-e)

**Attribution :** Ces expressions contiennent dans leur sens (à la manière de *npr*) ou (à propos de *npr*). Ces expressions semblent interdire les relations coréférentielles avec d'autres syntagmes. Par exemple, *la formule Excel* ne devrait pas être coréférentielle avec *Excel*.

*la folie Spider Man* (6-a)

*l'écurie Marvel* (6-b)

*l'effet Invasion* (6-c)

*l'aventure EnCana* (6-d)

*la formule Excel* (6-e)

*les formats RedHat et Debian* (6-f)

[6] Nous ne sommes pas les premiers à nous intéresser aux constructions `det nom_commun npr`. Voir G. Kleiber [Kle81], K. Jonasson [Jon94] et Gary-Prieur [GP94].

[7] Terme emprunté à [Mel93].

**Hyperonymie :** Ces constructions sont les plus nombreuses dans les textes. Dans ces cas, le nom commun et le nom propre sont en relation d'hyperonymie et sont interchangeables (pas seulement en contexte). Aussi, il est habituellement possible d'insérer le verbe *être* entre les deux éléments. Dans les textes de notre corpus, ces expressions contiennent habituellement des mots du domaine. Elles peuvent donc être reprises anaphoriquement (*la société ABB ... la société*) ou être coréférentielles à un autre syntagme (*la société ABB ... ABB*).

*le metteur en scène Beno Besson* (7-a)

*le chien Idéfix* (7-b)

*l'architecte Numérobis* (7-c)

*son père Kirk* (7-d)

*la société ABB* (7-e)

*le directeur financier Peter Rubenovitch* (7-f)

*la variable d'environnement LD_LIBRARY_PATH* (7-g)

*le programme bin/detect* (7-h)

## 7.5 Pour aller plus loin...

Il est rare qu'une étude soit entièrement terminée. Dans cette optique, nous pouvons considérer que notre algorithme permettant d'établir automatiquement les liens coréférentiels d'un texte est encore à un stade primitif. Nous n'avons que tracé les grandes lignes directrices permettant d'utiliser les règles *knowledge-poor* permettant de résoudre les types de liens anaphoriques et coréférentiels communs à tous les types de textes. Cependant, nous croyons que notre approche peut inspirer une vaste possibilité pour des travaux futurs. En voici quelques exemples :

### 7.5.1 Traiter d'autres cas simples

Dans notre recherche, nous avons essayé de trouver des règles les plus simples possible permettant d'établir certaines chaînes coréférentielles d'un texte. Cependant, nous ne croyons pas avoir traité tous les cas faciles. Nous pensons qu'il est encore possible d'ajouter quelques autres cas coréférentiels ou anaphoriques sans trop complexifier le traitement.

Par exemple, il est possible de reprendre quelques cas que nous avons laissé tomber. Nous pouvons mentionner le pronom *le* (en position complément d'objet direct) qui se positionne devant certains verbes d'action. Le pronom *le* est relativement rare dans les textes de type journalistique mais il est très fréquent dans les textes procéduraux (comme dans les *HOWTO Linux*). Une étude approfondie de ces types de textes pourrait certainement apporter quelques indications sur la stratégie à adopter pour traiter ces pronoms.

### 7.5.2 Traiter d'autres types de cas coréférentiels

Au chapitre *Travaux précédents* (voir section 3.1 page 26) nous avons mentionné qu'il existe d'autres relations que les relations coréférentielles impliquant l'identité des références. Nous rappelons ici que nous pouvons retrouver des liens de possession, d'association et de partie-tout :

***Martin***$_1$ *aimerait bien retrouver* ***son chien***$_{2,anaph=1(possession)}$. (8-a)

*Il y a* ***une jolie fenêtre***$_1$ *dans le salon mais* ***les rideaux***$_{2,anaph=1(association)}$ *sont affreux.* (8-b)

*Il y a* ***une lampe***$_1$ *dans le salon mais* ***l'ampoule***$_{2,anaph=1(partie-tout)}$ *est brûlée.* (8-c)

Particulièrement, les cas possessifs nous semblent assez réguliers. Puisque *son* $x$ peut à peu près toujours être paraphrasé par l'expression *le* $x$ *de* $y$, nous croyons que ces cas pourraient être résolus facilement. La résolution anaphorique des syntagmes possessifs consiste à déterminer qui est $y$.

### 7.5.3 Utiliser les différences des moyens coréférentiels selon le type de texte

Un des points importants de notre analyse selon le type de texte est qu'il existe beaucoup de facteurs qui varient selon le type de texte. Nous rappelons ici ces facteurs (section 6.4 page 77) :

- le vocabulaire utilisé ;
- les différences entre les éléments référentiels (types de syntagmes nominaux) ;
- les constructions particulières ;
- l'organisation des chaînes coréférentielles à l'intérieur d'un texte.

Dans ce mémoire, nous n'avons cependant pas exploité ces différences dans le calcul des chaînes coréférentielles. Il nous semble que ces facteurs soient importants et qu'ils nécessitent une étude approfondie.

## 7.6 Conclusion

Malgré le petit nombre de types de textes étudiés, nous avons tout de même réussi à dégager quelques principes pouvant être utiles pour le l'établissement automatique des liens coréférentiels. Voici les étapes que nous jugeons essentielles pour le traitement des chaînes coréférentielles d'un nouveau type de texte (voir section 6.4 page 77) :

- Identifier le vocabulaire utilisé.
- Déterminer les différences entre les éléments référentiels.
- Identifier les constructions anaphoriques particulières.
- Étudier l'organisation des chaînes coréférentielles à l'intérieur d'un texte.

Ces étapes peuvent demander assez de travail de la part du chercheur ; les ressources déjà existantes peuvent néanmoins réduire la tâche. Par exemple, le chercheur peut utiliser des dictionnaires terminologiques ou des ontologies du domaine étudié. Il peut aussi arriver qu'un type de texte partage des similarités avec d'autres types de textes. De ce point de vue, une étude plus poussée des différents types de textes nous semble nécessaire.

Il est plutôt difficile de prédire les résultats de notre démarche appliquée à un nouveau type de texte. Nous avons vu, dans le cas des *HOWTO Linux*, que notre approche n'est pas tout à fait au point. Certains ajustements pourraient toutefois améliorer les résultats de façon globale. Notamment, considérer la longueur des chaînes déjà existantes dans la pondération des antécédents, utiliser certaines informations contenues dans la mise en forme du document et utiliser une classification des noms propres.

# Chapitre 8

# Conclusion

La résolution des chaînes coréférentielles d'un texte est encore de nos jours une tâche très ambitieuse pour le traitement automatique de la langue. Au cours de nos recherches, nous estimons avoir réussi à mieux comprendre la référence en explorant le sujet sous toutes ses formes. Notre parcours a emprunté plusieurs chemins, examinant de façon assez large les aspects théoriques et pratiques du phénomène. Lors de notre exploration, nous avons aussi réussi à dégager quelques repères pour aider à résoudre les chaînes anaphoriques et coréférentielles.

Dans l'introduction de ce mémoire, nous avons fait une liste des objectifs que nous nous étions fixés. Revoici cette liste (section 1.2 page 6) :

- Caractériser les facteurs linguistiques influençant la coréférence.
- Dégager les comportements coréférentiels variant selon le type de texte.
- Concevoir une façon de baliser la coréférence dans le but de faciliter le traitement informatique.
- Adapter au français certains travaux permettant la résolution des liens coréférentiels qui ont été faits pour des textes anglais.
- Déterminer les ressources linguistiques minimales nécessaires au traitement automatique des liens coréférentiels.

## 8.1 Facteurs influençant la coréférence

Nous considérons que les expressions référentielles sont les syntagmes nominaux dont la $\text{tête}_r$ est un nom commun ou un nom propre. Les syntagmes contenant un pronom de troisième personne ou une ellipse font aussi partie de cette catégorie. Voici donc les éléments linguistiques impliqués dans la référence :

**Nom commun :** Le nom commun $x$ (utilisé seul) représente *l'ensemble des individus étiquetés* $x$ : ainsi, *chien* évoque l'ensemble de tous les chiens. Cet ensemble est la classe référentielle représentée par $x$.

**Nom propre :** Le nom propre a une signification semblable à celle du nom commun : (celui qui est appelé $x$). La différence entre un nom commun et un nom propre se situe au niveau de la classification (la catégorisation) des concepts qu'ils représentent. Il n'y a pas véritablement de classe référentielle associée à un nom propre sinon une classe ne contenant qu'un seul élément : l'individu appelé $x$.

**Pronom et ellipse :** Les syntagmes chapeautés par les pronoms et les ellipses sont totalement dépendants de leur antécédent car l'antécédent contient une grande partie de l'interprétation lexicale du syntagme. Cette dépendance garantit par le fait même l'unicité référentielle des syntagmes pronominaux et des ellipses.

**Déterminant :** Puisque le nom commun, seul, ne fait qu'identifier la classe référentielle sur laquelle porte la référence ; le syntagme nominal, en entier, contient le reste de l'information nécessaire pour déterminer un individu particulier (ou un groupe d'individus particulier). Ce sont les déterminants qui remplissent le mieux la fonction individualisante du syntagme nominal. Chacun des types de déterminants utilise toutefois des moyens différents pour le faire.

La référence est un phénomène qui se déroule tout au long de la lecture d'un texte. Ainsi, les référents peuvent apparaître pour une première fois, obligeant le lecteur à réserver une place pour cette nouvelle référence virtuelle dans sa mémoire. Par la suite, cette référence peut se présenter à nouveau, se dissiper (si elle n'est plus évoquée dans le texte), être mise en relief ou être présentée sous un autre jour.

L'apparition et l'évolution des éléments référentiels peuvent être représentées à l'aide des chaînes anaphoriques et coréférentielles. Rappelons qu'une identité éventuelle entre deux éléments référentiels dont les interprétations sont indépendantes est une coréférence et que l'interprétation d'un élément qui tire nécessairement son interprétation d'un autre élément est une anaphore.

Pour faire suite à une analyse des facteurs influençant les liens coréférentiels à l'intérieur d'un texte, nous avons retenu une série de points importants :

- La compatibilité de certains traits sémantiques (genre, nombre, personne, animation) entre les antécédents et les anaphores est très importante.
- Les sujets sont de meilleurs antécédents. Viennent ensuite les compléments d'objet direct, les compléments d'objet indirect et les autres compléments.
- La distance entre l'antécédent et l'anaphore n'est jamais très grande ; au plus quelques phrases. Certaines conditions permettent aussi à l'antécédent et au pronom d'être dans la même phrase.
- Les réitérations lexicales sont de bons indicateurs de liens coréférentiels.
- Les référents associés au domaine se comportent de façon différente des autres éléments référentiels du texte. Ces référents sont aussi plus faciles à résoudre.

## 8.2 Comportements coréférentiels selon le type de texte

Les règles pour résoudre les types de liens coréférentiels communs à tous les types de textes sont relativement simples. Nous avons d'ailleurs fait une ébauche d'algorithme pour résoudre ces cas. Par contre, les particularités référentielles variant selon le type de texte demandent un peu plus de travail. Lors de notre analyse, nous croyons néanmoins avoir dégagé quelques pistes pour guider l'application de la démarche à un nouveau type de texte.

### Vocabulaire selon le domaine

Le vocabulaire est une ressource-clé pour notre recherche. Pour limiter les ressources lexicographiques, nous utilisons un vocabulaire composé de noms communs empiriquement fréquents et prévisibles selon le domaine. Cette liste devrait contenir quelques centaines de mots (tout au plus).

### Type d'éléments référentiels

La fréquence des différents types d'éléments référentiels varie selon le type de texte. Par exemple, pour indiquer la reprise d'une entité dans le texte, certains types de textes utilisent plus souvent certains types de pronoms (par exemple le pronom *le* en position C.O.D dans les *HOWTO Linux*) tandis que d'autres types de textes utilisent des syntagmes possessifs (comme dans les critiques de films). Il est important de caractériser la fréquence des éléments coréférentiels et anaphoriques pour faciliter leur résolution. Ainsi, il est possible que dans certains types de textes, l'importance de certaines règles (la priorité d'une règle *knowledge-poor*) soit influencée par ces choix.

### Constructions particulières

Certaines constructions sont propres à certains types de texte. Ces constructions sont parfois anaphoriques et nécessitent un traitement particulier. Par exemple, les compléments elliptiques sont très présents dans les recettes de cuisine et représentent bien souvent des éléments en transformation dont le référent est plus ou moins défini.

### Charpente coréférentielle

L'évolution des éléments référentiels varie selon les textes. Dans certains, une entité importante se retrouve dans presque tous les paragraphes. Il peut aussi arriver que l'élément se transforme (comme dans les recettes de cuisine) ou s'agglutine à un autre élément (fusions de compagnie). Par contre, dans d'autres textes, il n'existe pas de telles entités importantes. L'ensemble de toutes les chaînes coréférentielles des textes représente plus ou moins le squelette du texte et la morphologie du squelette est plus ou moins typique selon le type de texte.

## 8.3 Schématisation de la coréférence

*XML* est un outil pratique en linguistique et beaucoup plus souple que le concordancier. Il est cependant plus complexe car il nécessite l'utilisation de plusieurs outils qui lui sont reliés.

Nous avons utilisé *XML* pour étiqueter les éléments lexicaux, les syntagmes référentiels et les liens anaphoriques et coréférentiels entre les syntagmes. Nous avons utilisé *CSS*, à l'intérieur de *XHTML*, pour ajouter des éléments de style et pour mettre en relief les éléments importants de notre étude. Nous avons fait des requêtes dans les textes balisés à l'aide de *XPath* et nous avons transformé les textes pour présenter les résultats sous un autre jour à l'aide de *XSLT*.

## 8.4 Adaptation au français

Pour l'établissement des chaînes coréférentielles, les algorithmes adaptés au français ne sont pas fondamentalement différents des algorithmes pour l'anglais. Notons toutefois quelques différences :

- le genre des mots est plus important ;
- l'utilisation des prépositions facilite la détection des syntagmes référentiels et l'attribution des fonctions syntaxiques ;
- le système des déterminants est un peu différent de celui de l'anglais.

## 8.5 Minimalisation des ressources

Nous croyons avoir démontré qu'il est possible de résoudre beaucoup de liens coréférentiels avec relativement peu de ressources linguistiques. Selon notre analyse, il existe une sous-classe de chaînes coréférentielles qui est beaucoup plus facile à résoudre. Aussi, ces chaînes sont habituellement les chaînes les plus importantes pour comprendre le contenu d'un texte.

Pour établir les chaînes coréférentielles importantes d'un texte, nous avons utilisé un vocabulaire constitué de listes de mots grammaticaux (listes fermées). Ainsi, nous avons utilisé les déterminants, les pronoms, les prépositions, etc. ainsi que quelques listes de mots particuliers tels les noms communs importants selon le domaine, les noms communs référant à la situation d'énonciation ainsi que les verbes impersonnels. Ces mots sont listés en annexe (voir annexe B page iv).

Par la suite, nous avons identifié les syntagmes référentiels et les liens anaphoriques et coréférentiels grâce à un système de règles assez simple. Pour ces règles, nous avons utilisé les mots identifiés précédemment, les noms propres (mots débutant par une majuscule) ainsi que la position de ces éléments dans la phrase ou par rapport aux autres éléments de la phrase.

# Annexes

# Annexe A

# Glossaire

*CSS* **(Cascading Style Sheets) :** *CSS* est un mécanisme pour ajouter du style (police de caractères, couleur, espacement, etc.) à des documents Web. 44

*HTML* **(HyperText Markup Language) :** *HTML* est un langage de balisage utilisé pour structurer des documents sur Internet. Ces documents peuvent être visualisés à l'aide d'un fureteur. 43

*XHTML* **(Extensible HyperText Markup Language) :** *XHTML* est une reformulation du *HTML* (HyperText Markup Language) en un sous-ensemble de *XML*. 43

*XML* **(Extensible Markup Language) :** *XML* est un langage de balisage permettant, dans un sens très large, de structurer des documents et des données. 36

*XPath* **(*XML* Path Language) :** *XPath* est un ensemble de règles syntaxiques servant à déterminer des parties de document *XML*. 45

*XSLT* **(Extensible Stylesheet Language Transformation) :** *XSLT* est un langage pour transformer des documents *XML*. 46

**anaphore :** L'anaphore est un processus syntaxique consistant à reprendre par un segment, un pronom en particulier, un autre segment du discours, un syntagme nominal antérieur, par exemple. 21

**asymétrie d'une relation :** Soit $A$, un ensemble et $R$ une relation dans $A$, $R$ est asymétrique $\Leftrightarrow [\forall \langle a, b \rangle \in R, \langle b, a \rangle \notin R]$. 21

**chaîne coréférentielle :** Une chaîne coréférentielle est le regroupement de tous les éléments lexicaux référant au même objet extra-linguistique. 19

**chaîne coréférentielle importante d'un texte :** Les chaînes coréférentielles importantes d'un texte sont des chaînes coréférentielles qui sont proportionnellement longues par rapport aux autres chaînes du texte, qui ont une distribution qui s'étend sur la longueur du texte et dont le premier élément (la tête de la chaîne coréférentielle) est habituellement dans le titre ou les deux premières phrases du texte. 81

**classe référentielle :** La classe référentielle est l'ensemble conceptuel de tous les particuliers réunis sous l'étiquette de l'item lexical. 15

**coréférence :** La coréférence est le lien entre deux éléments linguistiques partageant la même référence du monde extra-linguistique. 20

**désignateur rigide :** Le désignateur rigide est une association directe et durable entre un élément linguistique et son référent (dans le monde extra-linguistique). 16

**irréflexivité d'une relation :** Soit $A$, un ensemble et $R$ une relation dans $A$, $R$ est irréflexive $\Leftrightarrow [\forall a \in A, \langle a, a\rangle \notin R]$. 21

**précision :** la précision est la proportion de bons résultats obtenus parmi le nombre total de résultats obtenus. 58

**référence :** La référence est la fonction par laquelle un signe linguistique renvoie à un objet du monde extra-linguistique, réel ou imaginaire. 13

**réflexivité d'une relation :** Soit $A$, un ensemble et $R$ une relation dans $A$, $R$ est réflexive $\Leftrightarrow$ $[\forall a \in A, \langle a, a\rangle \in R]$. La relation est non réflexive sinon. 20

**résolution des anaphores :** La résolution des anaphores est l'opération consistant à repérer l'antécédent des éléments anaphoriques. 4

**rappel :** le rappel d'un résultat est le nombre de bons résultats trouvés, divisé par le nombre total de bons résultats. 59

**relation d'équivalence :** Une relation d'équivalence est une relation qui est réflexive, symétrique et transitive. 21

**relation d'ordre strict :** Une relation d'ordre strict est une relation qui est transitive, irréflexive et asymétrique. 22

**symétrie d'une relation :** Soit $A$, un ensemble et $R$ une relation dans $A$, $R$ est symétrique $\Leftrightarrow$ $[\forall \langle a, b\rangle \in R \Rightarrow \langle b, a\rangle \in R]$. La relation est non symétrique sinon. 20

**tête coréférentielle :** Une tête de chaîne coréférentielle (notée $tête_c$) est le premier élément d'une chaîne coréférentielle (selon l'ordre d'occurrence dans le texte). 20

**transitivité d'une relation :** Soit $A$, un ensemble et $R$ une relation dans $A$, $R$ est transitive $\Leftrightarrow$ $[\forall \langle a, b\rangle$ et $\langle b, c\rangle \in R \Rightarrow \langle a, c\rangle \in R]$. La relation est non transitive sinon. 20

# Annexe B

# Vocabulaire

```
<?xml version="1.0" encoding="iso-8859-1"?>
<vocabulaire>
    <Fusion>
        <noms>
            <nom traits="domaine" genre="f" nombre="sing">fusion</nom>
            <nom traits="domaine" genre="f" nombre="plur">fusions</nom>
            <nom traits="domaine" genre="f" nombre="sing">offre</nom>
            <nom traits="domaine" genre="f" nombre="plur">offres</nom>
            <nom traits="domaine" genre="f" nombre="sing">compagnie</nom>
            <nom traits="domaine" genre="f" nombre="plur">compagnies</nom>
            <nom traits="domaine" genre="f" nombre="sing">proposition</nom>
            <nom traits="domaine" genre="f" nombre="plur">propositions</nom>
            <nom traits="domaine" genre="f" nombre="sing">acquisition</nom>
            <nom traits="domaine" genre="f" nombre="plur">acquisitions</nom>
            <nom traits="domaine" genre="m" nombre="sing">organisme</nom>
            <nom traits="domaine" genre="m" nombre="plur">organismes</nom>
            <nom traits="domaine" nombre="sing">actionnaire</nom>
            <nom traits="domaine" nombre="plur">actionnaires</nom>
            <nom traits="domaine" genre="m" nombre="sing">accord</nom>
            <nom traits="domaine" genre="m" nombre="plur">accords</nom>
            <nom traits="domaine" genre="f" nombre="sing">entente</nom>
            <nom traits="domaine" genre="f" nombre="plur">ententes</nom>
            <nom traits="domaine" genre="m" nombre="sing">groupe</nom>
            <nom traits="domaine" genre="m" nombre="plur">groupes</nom>
            <nom traits="domaine" genre="m" nombre="sing">groupement</nom>
            <nom traits="domaine" genre="f" nombre="sing">société</nom>
            <nom traits="domaine" genre="f" nombre="plur">sociétés</nom>
            <nom traits="domaine" genre="f" nombre="sing">autorité</nom>
            <nom traits="domaine" genre="f" nombre="plur">autorités</nom>
            <nom traits="domaine" genre="m" nombre="sing">chef</nom>
            <nom traits="domaine" genre="m" nombre="sing">conseil</nom>
            <nom traits="domaine" genre="m" nombre="plur">conseils</nom>
            <nom traits="domaine" genre="m" nombre="sing">directeur</nom>
            <nom traits="domaine" genre="m" nombre="plur">directeurs</nom>
            <nom traits="domaine" genre="f" nombre="sing">directrice</nom>
            <nom traits="domaine" genre="f" nombre="plur">directrices</nom>
```

```
            <nom traits="domaine" genre="f" nombre="sing">direction</nom>
            <nom traits="domaine" genre="f" nombre="plur">directions</nom>
            <nom traits="domaine" genre="m" nombre="sing">fabricant</nom>
            <nom traits="domaine" genre="m" nombre="plur">fabricants</nom>
            <nom traits="domaine" genre="m" nombre="sing">projet</nom>
            <nom traits="domaine" genre="m" nombre="plur">projets</nom>
            <nom traits="domaine" genre="m" nombre="sing">actif</nom>
            <nom traits="domaine" genre="m" nombre="plur">actifs</nom>
            <nom traits="domaine" nombre="sing">action</nom>
            <nom traits="domaine" nombre="plur">actions</nom>
            <nom traits="domaine" genre="f" nombre="sing">transaction</nom>
            <nom traits="domaine" genre="f" nombre="plur">transactions</nom>
            <nom traits="domaine" genre="m" nombre="sing">administrateur</nom>
            <nom traits="domaine" genre="m" nombre="plur">administrateurs</nom>
            <nom traits="domaine" genre="f" nombre="sing">administratrice</nom>
            <nom traits="domaine" genre="f" nombre="plur">administratrices</nom>
            <nom traits="domaine" genre="f" nombre="sing">administration</nom>
            <nom traits="domaine" genre="f" nombre="plur">administrations</nom>
            <nom traits="domaine" genre="f" nombre="sing">filiale</nom>
            <nom traits="domaine" genre="f" nombre="plur">filiales</nom>
            <nom traits="domaine" genre="f" nombre="sing">firme</nom>
            <nom traits="domaine" genre="f" nombre="plur">firmes</nom>
            <nom traits="domaine" genre="m" nombre="sing">mariage</nom>
            <!--Des faux npr -->
            <nom>US</nom>
            <nom traits="domaine" genre="m" nombre="sing">PDG</nom>
            <nom traits="domaine" genre="m" nombre="sing">M</nom>
            <nom traits="domaine" genre="m" nombre="sing">M.</nom>
        </noms>
    </Fusion>
    <Film>
        <noms>
            <nom traits="domaine" genre="m" nombre="sing">film</nom>
            <nom traits="domaine" genre="m" nombre="plur">films</nom>
            <nom traits="domaine" genre="m" nombre="sing">personnage</nom>
            <nom traits="domaine" genre="m" nombre="plur">personnages</nom>
            <nom traits="domaine" genre="m" nombre="sing">homme</nom>
            <nom traits="domaine" genre="m" nombre="plur">hommes</nom>
            <nom traits="domaine" genre="m">fils</nom>
            <nom traits="domaine" genre="m" nombre="sing">petit-fils</nom>
            <nom traits="domaine" genre="m" nombre="plur">petits-fils</nom>
            <nom traits="domaine" genre="f" nombre="sing">histoire</nom>
            <nom traits="domaine" genre="m" nombre="plur">scénario</nom>
            <nom traits="domaine" genre="f" nombre="sing">femme</nom>
            <nom traits="domaine" genre="f" nombre="plur">femmes</nom>
            <nom traits="domaine" genre="m" nombre="sing">père</nom>
            <nom traits="domaine" genre="m" nombre="sing">grand-père</nom>
            <nom traits="domaine" genre="m" nombre="sing">réalisateur</nom>
            <nom traits="domaine" genre="f" nombre="sing">réalisatrice</nom>
            <nom traits="domaine" genre="m" nombre="sing">récit</nom>
            <nom traits="domaine" genre="f" nombre="sing">famille</nom>
            <nom traits="domaine" genre="m">héros</nom>
            <nom traits="domaine" genre="m">superhéros</nom>
```

```
            <nom traits="domaine" genre="f" nombre="sing">scène</nom>
            <nom traits="domaine" genre="f" nombre="plur">scènes</nom>
            <nom traits="domaine" genre="m" nombre="sing">metteur en scènes</nom>
            <nom traits="domaine" genre="m" nombre="sing">dialogue</nom>
            <nom traits="domaine" genre="m" nombre="plur">dialogues</nom>
            <nom traits="domaine" genre="f" nombre="sing">intrigue</nom>
            <nom traits="domaine" genre="f" nombre="sing">mère</nom>
            <nom traits="domaine" genre="f" nombre="sing">grand-mère</nom>
            <nom traits="domaine" genre="f" nombre="sing">caméra</nom>
            <nom traits="domaine" genre="m" nombre="sing">couple</nom>
            <nom traits="domaine" genre="f" nombre="sing">image</nom>
            <nom traits="domaine" genre="f" nombre="plur">images</nom>
            <nom traits="domaine" genre="m" nombre="sing">mari</nom>
            <nom traits="domaine" genre="m" nombre="sing">acteur</nom>
            <nom traits="domaine" genre="m" nombre="plur">acteurs</nom>
            <nom traits="domaine" genre="f" nombre="sing">actrice</nom>
            <nom traits="domaine" genre="f" nombre="plur">actrices</nom>
            <nom traits="domaine" genre="m" nombre="sing">amant</nom>
            <nom traits="domaine" genre="m" nombre="sing">cinéaste</nom>
            <nom traits="domaine" genre="f" nombre="sing">dame</nom>
            <nom traits="domaine" nombre="sing">enfant</nom>
            <nom traits="domaine" nombre="plur">enfants</nom>
            <nom traits="domaine" genre="f" nombre="sing">fille</nom>
            <nom traits="domaine" genre="f" nombre="plur">filles</nom>
            <nom traits="domaine" genre="f" nombre="sing">petite-fille</nom>
            <nom traits="domaine" genre="f" nombre="plur">petites-filles</nom>
            <nom traits="domaine" genre="f" nombre="sing">finale</nom>
            <nom traits="domaine" genre="f" nombre="sing">oeuvre</nom>
            <nom traits="domaine" genre="f" nombre="plur">oeuvres</nom>
            <nom traits="domaine" genre="m" nombre="sing">chef-d'oeuvre</nom>
            <nom traits="domaine" genre="m" nombre="plur">chefs-d'oeuvres</nom>
            <nom traits="domaine" genre="f" nombre="sing">partition</nom>
            <nom traits="domaine" genre="m" nombre="sing">projet</nom>
            <nom traits="domaine" genre="f" nombre="sing">quête</nom>
            <nom traits="domaine" genre="m" nombre="sing">rôle</nom>
            <nom traits="domaine" genre="m" nombre="plur">rôles</nom>
            <nom traits="domaine" genre="m" nombre="sing">spectateur</nom>
            <nom traits="domaine" genre="m" nombre="plur">spectateurs</nom>
            <nom traits="domaine" genre="m" nombre="sing">ami</nom>
            <nom traits="domaine" genre="m" nombre="plur">amis</nom>
            <nom traits="domaine" genre="f" nombre="sing">amie</nom>
            <nom traits="domaine" genre="f" nombre="plur">amies</nom>
            <nom traits="domaine" genre="m" nombre="sing">décor</nom>
            <nom traits="domaine" genre="m" nombre="plur">décors</nom>
        </noms>
    </Film>
    <HOWTO>
        <noms>
            <nom traits="domaine" genre="m" nombre="sing">fichier</nom>
            <nom traits="domaine" genre="m" nombre="plur">fichiers</nom>
            <nom traits="domaine" genre="m" nombre="sing">système</nom>
            <nom traits="domaine" genre="m" nombre="plur">systèmes</nom>
            <nom traits="domaine" genre="m" nombre="sing">pilote</nom>
```

```
<nom traits="domaine" genre="m" nombre="plur">pilotes</nom>
<nom traits="domaine" genre="m" nombre="sing">répertoire</nom>
<nom traits="domaine" genre="m" nombre="plur">répertoires</nom>
<nom traits="domaine" genre="m" nombre="sing">sous-répertoire</nom>
<nom traits="domaine" genre="m" nombre="plur">sous-répertoires</nom>
<nom traits="domaine" genre="m" nombre="sing">disque</nom>
<nom traits="domaine" genre="m" nombre="plur">disques</nom>
<nom traits="domaine" genre="f" nombre="sing">disquette</nom>
<nom traits="domaine" genre="f" nombre="plur">disquettes</nom>
<nom traits="domaine" genre="f" nombre="sing">commande</nom>
<nom traits="domaine" genre="f" nombre="plur">commandes</nom>
<nom traits="domaine" genre="f" nombre="sing">partition</nom>
<nom traits="domaine" genre="f" nombre="plur">partitions</nom>
<nom traits="domaine" genre="m" nombre="sing">programme</nom>
<nom traits="domaine" genre="m" nombre="plur">programmes</nom>
<nom traits="domaine" genre="m" nombre="sing">serveur</nom>
<nom traits="domaine" genre="m" nombre="plur">serveurs</nom>
<nom traits="domaine" genre="m" nombre="sing">utilisateur</nom>
<nom traits="domaine" genre="m" nombre="plur">utilisateurs</nom>
<nom traits="domaine" genre="m" nombre="sing">super-utilisateur</nom>
<nom traits="domaine" genre="m" nombre="plur">super-utilisateurs</nom>
<nom traits="domaine" genre="f" nombre="sing">machine</nom>
<nom traits="domaine" genre="f" nombre="plur">machines</nom>
<nom traits="domaine" genre="m" nombre="sing">client</nom>
<nom traits="domaine" genre="m" nombre="plur">clients</nom>
<nom traits="domaine" genre="f" nombre="sing">carte</nom>
<nom traits="domaine" genre="f" nombre="plur">cartes</nom>
<nom traits="domaine" genre="f" nombre="sing">distribution</nom>
<nom traits="domaine" genre="f" nombre="plur">distributions</nom>
<nom traits="domaine" genre="f" nombre="sing">page</nom>
<nom traits="domaine" genre="f" nombre="plur">pages</nom>
<nom traits="domaine" genre="m" nombre="sing">script</nom>
<nom traits="domaine" genre="m" nombre="plur">scripts</nom>
<nom traits="domaine" genre="f" nombre="sing">version</nom>
<nom traits="domaine" genre="f" nombre="plur">versions</nom>
<nom traits="domaine" genre="m" nombre="sing">matériel</nom>
<nom traits="domaine" genre="m" nombre="plur">matériels</nom>
<nom traits="domaine" genre="f" nombre="sing">option</nom>
<nom traits="domaine" genre="f" nombre="plur">options</nom>
<nom traits="domaine" genre="m" nombre="sing">paquetage</nom>
<nom traits="domaine" genre="m" nombre="plur">paquetages</nom>
<nom traits="domaine" genre="f" nombre="sing">archive</nom>
<nom traits="domaine" genre="f" nombre="plur">archives</nom>
<nom traits="domaine" genre="m" nombre="sing">contrôleur</nom>
<nom traits="domaine" genre="m" nombre="plur">contrôleurs</nom>
<nom traits="domaine" genre="m" nombre="sing">utilitaire</nom>
<nom traits="domaine" genre="m" nombre="plur">utilitaires</nom>
<nom traits="domaine" genre="m" nombre="sing">langage</nom>
<nom traits="domaine" genre="m" nombre="plur">langages</nom>
<nom traits="domaine" genre="f" nombre="sing">variable</nom>
<nom traits="domaine" genre="f" nombre="plur">variables</nom>
<nom traits="domaine" genre="m" nombre="sing">démon</nom>
<nom traits="domaine" genre="m" nombre="plur">démons</nom>
```

```
        <nom traits="domaine" genre="m" nombre="sing">périphérique</nom>
        <nom traits="domaine" genre="m" nombre="plur">périphériques</nom>
        <nom traits="domaine" genre="m" nombre="sing">site</nom>
        <nom traits="domaine" genre="m" nombre="plur">sites</nom>
        <nom traits="domaine" genre="f" nombre="sing">source</nom>
        <nom traits="domaine" genre="f" nombre="plur">sources</nom>
        <nom traits="domaine" genre="f" nombre="sing">bibliothèque</nom>
        <nom traits="domaine" genre="f" nombre="plur">bibliothèques</nom>
        <nom traits="domaine" genre="f" nombre="sing">documentation</nom>
        <nom traits="domaine" genre="f" nombre="plur">documentations</nom>
        <nom traits="domaine" genre="m" nombre="sing">format</nom>
        <nom traits="domaine" genre="m" nombre="plur">formats</nom>
        <nom traits="domaine" genre="m" nombre="sing">logiciel</nom>
        <nom traits="domaine" genre="m" nombre="plur">logiciels</nom>
        <nom traits="domaine" genre="m" nombre="sing">module</nom>
        <nom traits="domaine" genre="m" nombre="plur">modules</nom>
        <nom traits="domaine" genre="m" nombre="sing">noyau</nom>
        <nom traits="domaine" genre="m" nombre="plur">noyaux</nom>
        <nom traits="domaine" genre="m" nombre="sing">outil</nom>
        <nom traits="domaine" genre="m" nombre="plur">outils</nom>
        <nom traits="domaine" genre="m" nombre="sing">paramètre</nom>
        <nom traits="domaine" genre="m" nombre="plur">paramètres</nom>
        <nom traits="domaine" genre="f" nombre="sing">station</nom>
        <nom traits="domaine" genre="f" nombre="plur">stations</nom>
        <nom traits="domaine" genre="f" nombre="sing">touche</nom>
        <nom traits="domaine" genre="f" nombre="plur">touches</nom>
        <nom traits="domaine" genre="m">bus</nom>
        <nom traits="domaine" genre="m" nombre="sing">pseudo-code</nom>
        <nom traits="domaine" genre="m" nombre="sing">code</nom>
        <nom traits="domaine" genre="m" nombre="sing">forum</nom>
        <nom traits="domaine" genre="m" nombre="plur">forums</nom>
        <nom traits="domaine" genre="m" nombre="sing">gestionnaire</nom>
        <nom traits="domaine" genre="m" nombre="plur">gestionnaires</nom>
        <nom traits="domaine" genre="f" nombre="sing">image</nom>
        <nom traits="domaine" genre="f" nombre="plur">images</nom>
        <nom traits="domaine" genre="m" nombre="sing">ordinateur</nom>
        <nom traits="domaine" genre="m" nombre="plur">ordinateur</nom>
        <nom traits="domaine" genre="m" nombre="sing">port</nom>
        <nom traits="domaine" genre="f" nombre="sing">session</nom>
        <nom traits="domaine" genre="f">souris</nom>
        <nom traits="domaine" genre="m" nombre="sing">écran</nom>
        <nom traits="domaine" genre="m" nombre="plur">écrans</nom>
        <nom traits="domaine" genre="m" nombre="sing">plein-écran</nom>
        <nom traits="domaine" genre="f" nombre="sing">application</nom>
        <nom traits="domaine" genre="f" nombre="plur">applications</nom>
        <nom traits="domaine" genre="m" nombre="sing">binaire</nom>
        <nom traits="domaine" genre="m" nombre="plur">binaires</nom>
        <nom traits="domaine" genre="m" nombre="sing">compte</nom>
        <nom traits="domaine" genre="m" nombre="plur">comptes</nom>
        <nom traits="domaine" genre="m" nombre="sing">protocole</nom>
        <nom traits="domaine" genre="m" nombre="plur">protocoles</nom>
    </noms>
</HOWTO>
```

```
<tous>
    <noms>
        <nom traits="déictique" genre="m" nombre="sing">jour</nom>
        <nom traits="déictique" genre="m" nombre="plur">jours</nom>
        <nom traits="déictique" genre="f" nombre="sing">semaine</nom>
        <nom traits="déictique" genre="f" nombre="plur">semaines</nom>
        <nom traits="déictique" genre="f" nombre="sing">année</nom>
        <nom traits="déictique" genre="f" nombre="plur">années</nom>
        <nom traits="déictique" genre="m" nombre="sing">chapitre</nom>
        <nom traits="déictique" genre="m" nombre="plur">chapitres</nom>
        <nom traits="déictique" genre="f" nombre="sing">section</nom>
        <nom traits="déictique" genre="f" nombre="plur">sections</nom>
        <nom traits="déictique" genre="m" nombre="sing">article</nom>
        <nom traits="déictique" genre="m" nombre="plur">articles</nom>
    </noms>
    <dets>
        <det type="indéfini">zéro</det>
        <det type="indéfini" genre="m">un</det>
        <det type="indéfini" genre="f">une</det>
        <det type="indéfini" nombre="plur">des</det>
        <det type="indéfini" nombre="plur">deux</det>
        <det type="indéfini" nombre="plur">trois</det>
        <det type="indéfini" nombre="plur">quatre</det>
        <det type="indéfini" nombre="plur">cinq</det>
        <det type="indéfini" nombre="plur">six</det>
        <det type="indéfini" nombre="plur">sept</det>
        <det type="indéfini" nombre="plur">huit</det>
        <det type="indéfini" nombre="plur">neuf</det>
        <det type="indéfini" nombre="plur">dix</det>
        <det type="indéfini" nombre="plur">onze</det>
        <det type="indéfini" nombre="plur">douze</det>
        <det type="indéfini" nombre="plur">treise</det>
        <det type="indéfini" nombre="plur">quatorze</det>
        <det type="indéfini" nombre="plur">quinze</det>
        <det type="indéfini" nombre="plur">seize</det>
        <det type="indéfini" nombre="plur">vingt</det>
        <det type="indéfini" nombre="plur">trente</det>
        <det type="indéfini" nombre="plur">quarante</det>
        <det type="indéfini" nombre="plur">cinquante</det>
        <det type="indéfini" nombre="plur">soxante</det>
        <det type="indéfini" nombre="plur">cent</det>
        <det type="indéfini" nombre="plur">cents</det>
        <det type="indéfini" nombre="plur">mille</det>
        <det type="indéfini" nombre="plur">milles</det>
        <det type="indéfini" genre="m">certain</det>
        <det type="indéfini" genre="f">certaine</det>
        <det type="indéfini" genre="m" nombre="plur">certains</det>
        <det type="indéfini" genre="f" nombre="plur">certaines</det>
        <det type="indéfini">chaque</det>
        <det type="indéfini" nombre="plur">plusieurs</det>
        <det type="indéfini" nombre="plur">quelque</det>
        <det type="indéfini" nombre="plur">quelques</det>
        <det type="défini" genre="m">le</det>
```

```
        <det type="défini">l'</det>
        <det type="défini" genre="f">la</det>
        <det type="défini" nombre="plur">les</det>
        <det type="démonstratif" genre="m">ce</det>
        <det type="démonstratif">c'</det>
        <det type="démonstratif" genre="m">cet</det>
        <det type="démonstratif" genre="f">cette</det>
        <det type="démonstratif" nombre="plur">ces</det>
    </dets>
    <pros>
        <pro type="démonstratif" genre="m">celui</pro>
        <pro type="démonstratif" genre="f">celle</pro>
        <pro type="démonstratif" genre="m" nbPer="plur">ceux</pro>
        <pro type="démonstratif" genre="f" nbPer="plur">celles</pro>
        <pro type="possessif" genre="m" personne="1">mien</pro>
        <pro type="possessif" genre="m" personne="2">tien</pro>
        <pro type="possessif" genre="m" personne="3">sien</pro>
        <pro type="possessif" genre="f" personne="1">mienne</pro>
        <pro type="possessif" genre="f" personne="2">tienne</pro>
        <pro type="possessif" genre="f" personne="3">sienne</pro>
        <pro type="possessif" genre="m" personne="1" nombre="plur">miens</pro>
        <pro type="possessif" genre="m" personne="2" nombre="plur">tiens</pro>
        <pro type="possessif" genre="m" personne="3" nombre="plur">siens</pro>
        <pro type="possessif" genre="f" personne="1" nombre="plur">miennes</pro>
        <pro type="possessif" genre="f" personne="2" nombre="plur">tiennes</pro>
        <pro type="possessif" genre="f" personne="3" nombre="plur">siennes</pro>
        <pro type="possessif" personne="1" nbPer="plur">nôtre</pro>
        <pro type="possessif" personne="2" nbPer="plur">vôtre</pro>
        <pro type="personnel" personne="1">je</pro>
        <pro type="personnel" personne="1">j'</pro>
        <pro type="personnel" personne="2">tu</pro>
        <pro type="personnel" genPer="m" personne="3">il</pro>
        <pro type="personnel" genPer="m" personne="3">-il</pro>
        <pro type="personnel" genPer="m" personne="3">-t-il</pro>
        <pro type="personnel" genPer="f" personne="3">elle</pro>
        <pro type="personnel" genPer="f" personne="3">-elle</pro>
        <pro type="personnel" genPer="f" personne="3">-t-elle</pro>
        <pro type="personnel" personne="3" aTraiter="false">on</pro>
        <pro type="personnel" personne="3" aTraiter="false">-on</pro>
        <pro type="personnel" personne="3" aTraiter="false">-t-on</pro>
        <pro type="personnel" personne="1" nbPer="plur">nous</pro>
        <pro type="personnel" personne="1" nbPer="plur">-nous</pro>
        <pro type="personnel" personne="2" nbPer="plur">vous</pro>
        <pro type="personnel" personne="2" nbPer="plur">-vous</pro>
        <pro type="personnel" genPer="m" personne="3" nbPer="plur">ils</pro>
        <pro type="personnel" genPer="m" personne="3" nbPer="plur">-ils</pro>
        <pro type="personnel" genPer="m" personne="3" nbPer="plur">-t-ils</pro>
        <pro type="personnel" genPer="f" personne="3" nbPer="plur">elles</pro>
        <pro type="personnel" genPer="f" personne="3" nbPer="plur">-elles</pro>
        <pro type="personnel" genPer="f" personne="3" nbPer="plur">-t-elles</pro>
        <pro type="personnel" personne="1">me</pro>
        <pro type="personnel" personne="1">m'</pro>
        <pro type="personnel" personne="2">te</pro>
```

```
        <pro type="personnel" personne="2">t'</pro>
        <pro type="personnel" personne="3" aTraiter="false">se</pro>
        <pro type="personnel" personne="3" aTraiter="false">s'</pro>
        <pro type="personnel" personne="3">lui</pro>
        <pro type="personnel" personne="3" aTraiter="false">-en</pro>
        <pro type="personnel" personne="3" aTraiter="false">y</pro>
        <pro type="personnel" personne="3" aTraiter="false">-y</pro>
        <pro type="personnel" personne="1">moi</pro>
        <pro type="personnel" personne="2">toi</pro>
        <pro type="personnel" personne="3" aTraiter="false">ça</pro>
        <pro type="personnel" personne="3" aTraiter="false">cela</pro>
        <!--«ce» et «c'» ont étés traités avec les dét  -->
        <pro type="personnel" genPer="m" personne="3" nbPer="plur">eux</pro>
        <pro type="personnel" personne="1">soi</pro>
    </pros>
    <preps>
        <prep>à</prep>
        <prep>de</prep>
        <prep>d'</prep>
        <prep>dans</prep>
        <prep>en</prep>
        <prep>pour</prep>
        <prep>sur</prep>
        <prep>avec</prep>
        <prep>chez</prep>
        <prep>contre</prep>
        <prep>sous</prep>
        <prep>vers</prep>
        <prep>par</prep>
        <prep>avant</prep>
        <prep>après</prep>
        <prep>concernant</prep>
        <prep>depuis</prep>
        <prep>derrière</prep>
        <prep>dès</prep>
        <prep>devant</prep>
        <prep>durant</prep>
        <prep>en</prep>
        <prep>entre</prep>
        <prep>envers</prep>
        <prep>excepté</prep>
        <prep>hormis</prep>
        <prep>jusque</prep>
        <prep>jusqu'</prep>
        <prep>jusques</prep>
        <prep>malgré</prep>
        <prep>moyennant</prep>
        <prep>outre</prep>
        <prep>parmi</prep>
        <prep>passé</prep>
        <prep>pendant</prep>
        <prep>plein</prep>
        <prep>près</prep>
```

```
        <prep>proche</prep>
        <prep>sans</prep>
        <prep>sauf</prep>
        <prep>selon</prep>
        <prep>sous</prep>
        <prep>suivant</prep>
        <prep>supposé</prep>
        <prep>touchant</prep>
    </preps>
    <conjs>
        <conj>alors</conj>
        <conj>mais</conj>
        <conj>ou</conj>
        <conj>et</conj>
        <conj>donc</conj>
        <conj>car</conj>
        <conj>ni</conj>
        <conj>or</conj>
        <conj>ainsi</conj>
        <conj>aussi</conj>
        <conj>cependant</conj>
        <conj>combien</conj>
        <conj>comme</conj>
        <conj>encore</conj>
        <conj>enfin</conj>
        <conj>ensuite</conj>
        <conj>lorsque</conj>
        <conj>lorsqu'</conj>
        <conj>néanmoins</conj>
        <conj>partant</conj>
        <conj>pourquoi</conj>
        <conj>pourtant</conj>
        <conj>puis</conj>
        <conj>puisque</conj>
        <conj>puisqu'</conj>
        <conj>quoique</conj>
        <conj>quoiqu'</conj>
        <conj>si</conj>
        <conj>s'</conj>
        <conj>sinon</conj>
        <conj>soit</conj>
        <conj>tantôt</conj>
        <conj>toutefois</conj>
    </conjs>
    <rels>
        <rel>que</rel>
        <rel>qu'</rel>
        <rel>qui</rel>
        <rel>dont</rel>
        <rel>où</rel>
        <rel>lequel</rel>
        <rel>duquel</rel>
        <rel>auquel</rel>
```

```
        <rel>laquelle</rel>
        <rel>lesquels</rel>
        <rel>desquels</rel>
        <rel>auxquels</rel>
        <rel>lesquelles</rel>
        <rel>desquelles</rel>
        <rel>auxquelles</rel>
        <rel>quiconque</rel>
    </rels>
    <advs>
        <adv>plus</adv>
        <adv>pas</adv>
        <adv>n'</adv>
        <adv>ne</adv>
    </advs>
    <verbes>
        <verbe traits="attributif">était</verbe>
        <verbe traits="attributif">été</verbe>
        <verbe traits="attributif">est</verbe>
        <verbe traits="attributif">sont</verbe>
        <verbe traits="attributif">être</verbe>
        <verbe traits="attributif">sera</verbe>
        <verbe traits="impersonnel">semble</verbe>
        <verbe traits="impersonnel">faut</verbe>
        <verbe traits="impersonnel">faudra</verbe>
        <verbe traits="impersonnel">fait</verbe>
        <verbe traits="impersonnel">s'agit</verbe>
        <verbe traits="impersonnel">reste</verbe>
        <verbe traits="impersonnel">vaut</verbe>
        <verbe traits="impersonnel">existe</verbe>
        <verbe traits="impersonnel">suffit</verbe>
        <verbe traits="impersonnel">convient</verbe>
        <verbe>a</verbe>
        <verbe>ont</verbe>
        <verbe>avait</verbe>
        <verbe>va</verbe>
    </verbes>
    <poncs>
        <ponc>!</ponc>
        <ponc>.</ponc>
        <ponc>?</ponc>
        <ponc>:</ponc>
        <ponc>;</ponc>
        <ponc>%</ponc>
        <ponc>$</ponc>
        <ponc type="encadrant">,</ponc>
        <ponc type="encadrant">(</ponc>
        <ponc type="encadrant">)</ponc>
        <ponc type="encadrant">[</ponc>
        <ponc type="encadrant">]</ponc>
        <ponc type="encadrant">«</ponc>
        <ponc type="encadrant">»</ponc>
        <ponc type="encadrant">"</ponc>
```

```
            <ponc type="encadrant">-</ponc>
        </poncs>
    </tous>
</vocabulaire>
```

# Annexe C

# Exemples de textes

## C.1 Fusion de comapgnies

### L'expansion a desservi ABB[1]

**Bloomberg**

Zurich - Dans les années 1990, l'expansion de la société ABB a permis au PDG d'alors, Percy Barnevik, de devenir l'un des chefs d'entreprise les plus respectés au monde. Mais cette quête de mondialisation est maintenant à la source des malheurs de la compagnie de génie, soulignent des investisseurs.

Créée en 1988 à la suite de la fusion de Asea AB, de Suède, et de Brown Boveri, de Suisse, ABB a cherché à concurrencer des géants tels General Electric en réalisant plus de 200 acquisitions en une décennie et en mettant sur pied une division de services financiers.

Aujourd'hui, au moment où la code de crédit de la compagnie a été abaissée (ses titres de dette sont considérés à risque), le PDG Juergen Dormann tente de faire échec aux pertes en vendant des actifs et en réduisant l'effectif.

L'une des plus importantes acquisitions fut le rachat, en 1990, de Combustion Engineering, au coût de 1,6 milliard US. Cette compagnie américaine fabriquait des chaudières isolées à l'amiante, produit que l'on associa plus tard au cancer. Bien que la plupart des actifs de cette division eussent été vendus en l'an 2000, ABB doit faire face aujourd'hui à une facture d'au moins 1,1 milliard US pour régler 111 000 poursuites liées à l'amiante.

Pour limiter les paiements, ABB va demander la protection de la loi contre les créanciers pour sa division américaine. Mais les poursuites liées à l'amiante coûtent déjà environ 865 millions US à ABB, fabricant de robots d'usine et de câbles électriques. À ce chapitre, les réclamations contre les compagnies aux États-Unis ont atteint plus de 54 milliards US à la fin de l'an 2000 et elles pourraient coûter 210 milliards US supplémentaires, selon le Rand Institute.

[1] Tiré de [11].

**La grande leçon**

"Les gens étaient si emballés par ABB et par ses perspectives ; personne ne parlait d'amiante en 1990", soutient Goeran Espelund, PDG de Lannebo Fonder, qui gère des actifs d'environ 170 millions US. ABB, société établie à Zurich, en Suisse, a connu sa première perte en 2001. Le mois dernier, elle a fait savoir que sa perte au troisième trimestre était plus importante que prévu et a renoncé à une cible de profit pour 2002. Récemment, Moody's Investors Service a réduit sa notation touchant les titres de dette d'ABB, l'abaissant au niveau de "à risque", en raison des craintes suscitées par les poursuites liées à l'amiante.

Le titre d'ABB a chuté de 85 % cette année, conférant à la compagnie une valeur boursière de 2,7 milliards de francs suisses (soit 1,9 milliard US). Elle valait 57 milliards de francs en 1999.

Lorsque ABB, dont les racines remontent au secteur de l'éclairage au 19e siècle, en Suède, fit l'acquisition de Combustion Engineering en 1990, M. Barnevik déclara à la télévision suédoise qu'il était "difficile d'envisager des conséquences négatives".

La compagnie est devenue un modèle dans les années subséquentes. Qualifié de réponse européenne à Jack Welch, PDG de GE, M. Barnevik fit augmenter les ventes d'ABB de plus de 30 milliards US et gonfla l'effectif à plus de 200 000 avant de devenir président de la compagnie en 1996. Le Financial Times le salua comme PDG et président de la compagnie européenne la plus respectée chaque année de 1994 à 1997.

Normes de comptabilité

L'an dernier, les ennuis ont commencé pour ABB. Sous la pression de l'investisseur suisse Martin Ebner, qui avait rassemblé une participation de 11 % et avait joint les rangs du conseil d'administration de la compagnie en 1999, ABB racheta environ 2 % de ses actions pour faire remonter le titre, privant la compagnie de liquidités au moment où la demande commençait à fléchir. M. Ebner a quitté le conseil d'administration le mois dernier tandis que la valeur de sa participation, qui avait déjà atteint plus de 5 milliards de francs, avait chuté à moins de 180 millions de francs.

La compagnie a aussi adopté les principes comptables généralement reconnus aux États-Unis avant d'inscrire son titre aux Bourses américaines, une décision impliquant que la société devait réévaluer ses gains des trois années précédentes. Ce changement révéla que le bénéfice net combiné d'ABB pour 1998, 1999 et 2000 était inférieur d'environ 1,36 milliard US à ce qui avait été antérieurement déclaré.

Puis, en novembre de l'an dernier, M. Barnevik, 61 ans, démissionna de son poste de président au milieu d'une querelle portant sur des paiements de retraite de 140 millions US à lui-même et à un autre ancien PDG. Selon ABB, ces sommes n'auraient pas été approuvées dans les règles. M. Barnevik a refusé d'être interviewé dans le cadre de cet article.

Aujourd'hui, en plus de vendre des actifs, le PDG actuel, Juergen Dormann, projette de réduire les coûts de 800 millions US au cours des 18 prochains mois. Il a refusé de fournir un nouvel objectif de gains pour 2002 en raison de l'incertitude sur le marché. ABB présente la pire performance au sein de l'indice boursier suisse cette année.

La dernière fois que M. Dormann, 62 ans, a évoqué un objectif de profit pour ABB, c'était le 11 septembre dernier, une semaine après son accession au poste de PDG. Il indiquait alors que le bénéfice considéré dans un rapport avec les ventes allait s'établir à entre 4 et 5 %, comparativement à 1,2 % l'an dernier.

Le refus de préciser un nouvel objectif pour 2002 a fait baisser la confiance en l'ancien PDG d'Aventis, compagnie pharmaceutique, qui avait mené à bien la fusion de Rhône-Poulenc, de France, et de Hoechst, d'Allemagne, pour former Aventis. C'était en 1999 et c'était une affaire de 29 milliards US.

"La confiance a disparu", dit M. Folkmar.

## C.2 Critique de film

### La Turbulence des fluides[2]

**Un nouveau souffle**

Pascale Bussières et Julie Gayet dans une scène de La Turbulence des fluides.

À Baie-Comeau, un phénomène bizarre s'est produit : la marée s'est arrêtée, mettant à nu une plage de sable en partie constituée d'une île normalement engloutie par les hautes eaux. C'est sur ce site qu'une sismologue, Alice (Pascale Bussières), est envoyée avec pour mission de trouver la cause du bouleversement. Signe annonciateur d'un tremblement de terre ? Les experts de Tokyo où travaille Alice (dans l'attente du grand séisme) ne peuvent rejeter cette hypothèse.

Bien sûr, la donnée de base de La Turbulence des fluides repose sur une impossibilité physique. À preuve du contraire, la Lune exerce toujours son attraction, même dans le film de Manon Briand. Mais il est des licences poétiques qui peuvent se révéler fécondes. Le scénario de cette Turbulence qui repose justement sur une correspondance entre les lois de la nature et celles de l'âme, exige du spectateur ce consentement préalable à une entorse possible aux lois de Newton.

J'aime personnellement ce télescopage d'intrigues. Le film démarre sur une catastrophe appréhendée et finit par s'arrimer aux effets d'une autre catastrophe : la chute, un an plus tôt, d'un petit avion dont l'un des passagers, une femme, n'a jamais été retrouvé. Dans l'esprit du film, ces deux catastrophes s'emboîtent l'une dans l'autre. Les prétendues victimes de l'absence de marée sont en fait des individus ayant, pour une raison ou pour une autre, souffert de cet accident d'avion. Bel exemple de pensée magique à l'oeuvre dans le cinéma québécois.

Cela étant dit, le fragile échafaudage bâti par Manon Briand s'étoffe peu à peu d'intéressants numéros d'acteurs. Pascale Bussières campe une Alice en apparence insensible, repliée sur elle-même à la suite d'une rupture, et qui va peu à peu s'épanouir sous l'effet d'une rencontre. En pilote de brousse et pompier volant, Marc (Jean-Nicolas Verreault) incarne avec brio cette nouvelle vision du mâle québécois, à la fois viril et sensible. Même si son personnage paraît quelque peu superflu, le rôle de Julie Gayet en Catherine, une journaliste lesbienne, est défendu sur un ton léger et plaisant. Il y a aussi Geneviève Bujold en patronne de café, Colette, religieuse défroquée qui n'a que peu de répliques mais dont les silences paraissent tout aussi éloquents. Sans parler de la petite chinoise (Ji-Yan Séguin), attendrissante en orpheline mue par des réflexes mystérieux.

Manon Briand avait surpris par la spontanéité de son premier long métrage, Deux Secondes. La maturité aidant, La Turbulence des fluides me paraît encore plus maîtrisée et d'une audace plus risquée. Ce n'est pas tout le monde qui voudra endosser son pari difficile. Mais pour qui arrive à la suivre dans ses virages très marqués, son film qui se déroule dans un décor naturel rarement exploité au cinéma représente un souffle nouveau qu'on aurait tort de bouder.

[2] Tiré de [7].

## C.3 *HOWTO Linux*

### Terminal Texte pour Linux[3]

**3. Installation rapide**

Voici une procédure rapide pour installer un terminal sans passer par une procédure de [7]mise en place à la fois pour le terminal et l'ordinateur hôte. Cela ne fonctionnera probablement pas bien s'il se trouve que le terminal a été configuré de manière incompatible avec l'ordinateur. Si vous ne comprenez pas tout ceci vous devrez consulter d'autres parties de ce document pour plus d'informations.

Pour installer un terminal, regardez d'abord dans /etc/termcap ou terminfo.src pour y trouver une entrée le concernant (voir [8]terminfo et termcap (détaillé)). Déterminez sur quel port série vous le brancherez et quelle est le nom tty pour ce port (par exemple, ttyS1, voyez [9]noms de périphériques). En tant qu'utilisateur root, éditez /etc/inittab et ajoutez une commande getty à côté des autres commandes getty. Le format de la commande getty dépend du programme getty que vous utilisez. agetty (simplement appelé getty dans la distribution Debian) est le plus simple (pas de fichier de configuration). Voyez le fichier "info" ou la page de manuel de getty. Pour les paramètres de getty, utilisez le nom terminfo (ou termcap) de votre terminal, comme vt100. Entrez une vitesse de transmission supportée par le terminal. Si vous mettez la vitesse trop haut vous aurez peut-être besoin d'utiliser le [10]contrôle de flux.

Connectez alors physiquement le port série principal du terminal au port série choisi de l'ordinateur avec un câble null-modem et allumez le terminal. N'espérez pas que la plupart des câbles tout prêts soient câblés correctement pour gérer le contrôle de flux matériel. Assurez-vous que la vitesse de transmission du terminal est la même que celle que vous avez donnée à getty et que son paramètre "bits de données" est 8. Alors, sur la console de l'ordinateur tapez "init q" pour faire prendre en compte les changements que vous avez faits au fichier inittab. Vous devriez maintenant voir une invite de login sur le terminal. Sinon, appuyez sur la touche retour chariot du terminal. Si cela ne fonctionne pas, continuez de lire ce document et/ou voyez [11]régler les problèmes.

[3] Tiré de [13].

# Annexe D

# Exemple d'un texte balisé

```
<?xml version="1.0" encoding="iso-8859-1"?>
<?xml-stylesheet type="text/xsl" href="../../Transformations/chaines.xsl"?>
<chaines domaine="Fusion"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:noNamespaceSchemaLocation=
    "file:/home/boudreau/www/private/Schemas/chainesCoref.xsd">
  <meta/>
  <corps>
    <par>
      <titre>mattel</titre>
    </par>
    <nontraite> Le Devoir Économie Mardi 15 décembre 1998 B2 </nontraite>
    <nontraite> En bref ... </nontraite>
    <par>
      <soustitre>
        <sn fonction="sujet" id="d0e40">
          <npr genre="f" traits="humain">Mattel</npr>
        </sn> achète <sn fonction="cod" id="d0e74">
          <det genre="m" type="indéfini">un</det>
          <nom traits="domaine">fabricant</nom>
          <prep>de</prep> logiciels <prep>de</prep> loisirs</sn>
        <nontraite>AFP Los Angeles</nontraite>
      </soustitre>
    </par>
    <nontraite> TYPE : Nouvelle brève LONGUEUR : Court </nontraite>
    <par>
      <ph>
        <coref idref="d0e40">
          <sn fonction="sujet" id="d0e75">
            <npr genre="f" traits="entité">Mattel</npr>
            <ponc>,</ponc>
            <sn fonction="apposition" id="d0e91">
              <nom traits="domaine">fabricant</nom> américain
              <prep>de</prep> jouets
              <conj>et</conj> notamment <prep>de</prep>
              <sn fonction="cNom" id="d0e92">
```

```
            <det genre="f" type="défini">la</det>
            <nom>poupée</nom>
            <npr genre="f" hyponyme="true" tete="false" traits="entité">Barbie
          </npr>
          </sn>
        </sn>
        <ponc>,</ponc>
      </sn>
    </coref>
    <verbe>a</verbe> annoncé hier <sn fonction="cod" id="d0e116">
      <det type="défini">l'</det>
      <nom traits="domaine">acquisition</nom>
      <prep>de</prep>
      <coref idref="d0e74">
        <sn fonction="cNom" id="d0e117">
          <det genre="f" type="défini">la</det>
          <nom traits="domaine">société</nom>
          <npr genre="f" hyponyme="true" tete="false" traits="entité">Learning
          </npr>
          <ponc>,</ponc> spécialisée <prep>dans</prep>
          <det nombre="plur" type="défini">les</det> logiciels éducatifs
          <conj>et</conj>
          <prep>de</prep> loisirs</sn>
      </coref>
    </sn>
    <ponc>.</ponc>
  </ph>
  <ph>
    <coref idref="d0e75">
      <sn fonction="sujet" id="d0e150">
        <npr genre="f" traits="entité">Mattel</npr>
      </sn>
    </coref>
    <verbe>va</verbe> racheter <coref idref="d0e117">
      <sn fonction="cod" id="d0e159">
        <npr genre="f" traits="entité">Learning</npr>
      </sn>
    </coref>
    <ponc>,</ponc>
    <prep>par</prep>
    <anaph idref="d0e116" type="association">
      <sn fonction="cod" id="d0e20">
        <nom traits="domaine">fusion</nom>
      </sn>
    </anaph>
    <ponc>,</ponc>
    <prep>pour</prep>
    <det genre="m" type="indéfini">un</det> montant évalué <prep>à</prep>
    <det nombre="plur" type="indéfini">3,8</det> milliards
    <prep>de</prep> dollars <ponc>.</ponc>
  </ph>
  <ph>
    <prep>Pour</prep>
```

```
        <anaph idref="d0e150" type="identite">
          <sn fonction="coi" id="d0e21">
            <det genre="m" type="défini">le</det>
            <nom traits="domaine">fabricant</nom>
            <prep>de</prep> jouets</sn>
        </anaph>
        <ponc>,</ponc>
        <det type="défini">l'</det> heure <verbe>est</verbe> venue
        <prep>de</prep> diversifier
          <anaph idref="d0e21" type="possession">
          <sn fonction="cod" id="d0e216">
            <det personne="3" type="possessif">son</det>
            <nom traits="domaine">offre</nom>
            <prep>de</prep> produits</sn>
        </anaph> notamment <prep>vers</prep>
        <det nombre="plur" type="défini">les</det> jeux interactifs
        <conj>et</conj> éducatifs <ponc>,</ponc>
        <conj>mais</conj>
        <conj>aussi</conj>
        <prep>sur</prep>
        <det genre="m" nombre="plur" type="indéfini">tous</det>
        <det nombre="plur" type="défini">les</det> supports offerts
        <prep>par</prep>
        <det nombre="plur" type="défini">les</det> nouvelles technologies
        <ponc>,</ponc>
        <conj>comme</conj>
        <det nombre="plur" type="défini">les</det> micro-ordinateurs
        <ponc>.</ponc>
      </ph>
    </par>
    <nontraite> DOC.#:981215LE043 </nontraite>
  </corps>
</chaines>
```

# Annexe E

# Tableaux des coréférences

Cette annexe contient des exemples chaînes coréférentielles présentes dans les textes suivants[1] :

## Fusion de compagnies :

**Superior Propane restera séparée, pour le moment, d'IGG PC Calgary**[2]

## Critique de film :

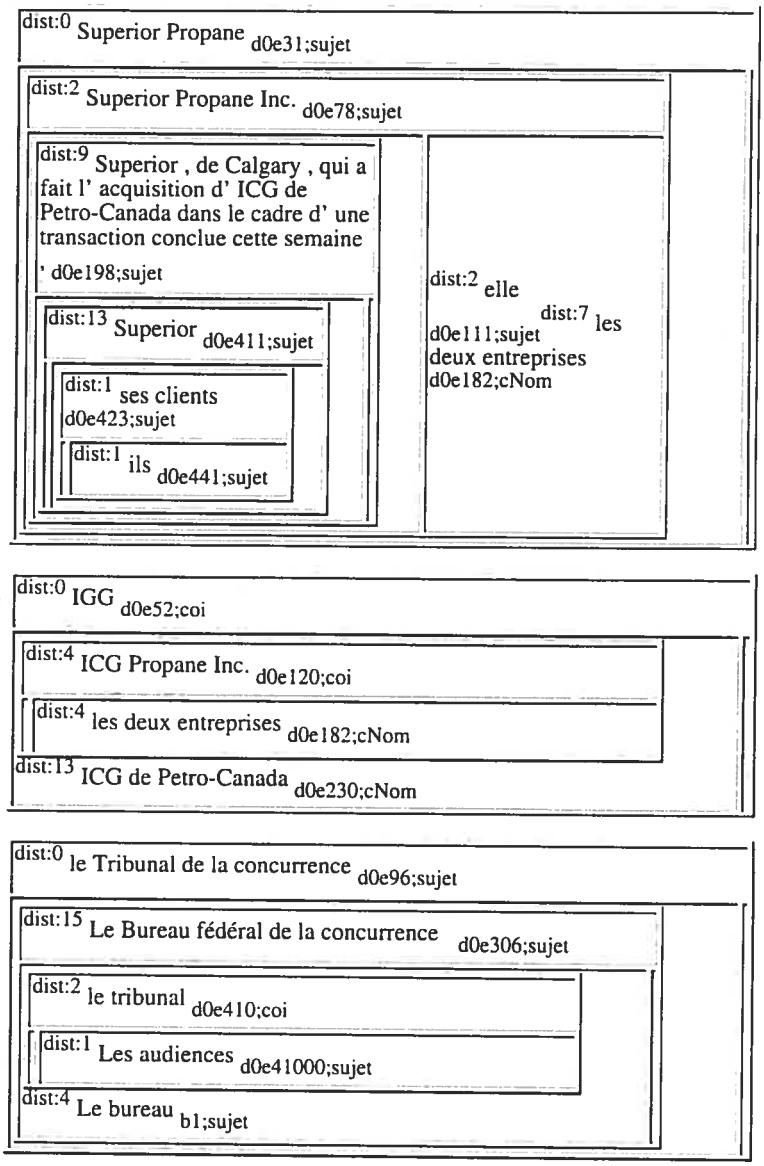
**Papillon**[3]

## *HOWTO Linux* :

**RedHat Linux KickStart HOWTO**[4]

[1] Ces textes ont été balisés manuellement.
[2] Tiré de [14].
[3] Tiré de [10].
[4] Tiré de [15].

# Chaines de «superior_propane»



dist:0 ce qu' ait été examiné le projet de fusion de 175 millions $ de les deux entreprises d0e135;cc

dist:0 le projet de fusion de 175 millions $ de les deux entreprises d0e18;cod

dist:3 cette fusion d0e183;sujet

dist:12 le projet x1;sujet

dist:4 l' entente x2;sujet

dist:0 fusion d0e181;cNom

dist:0 Calgary d0e209;cNom

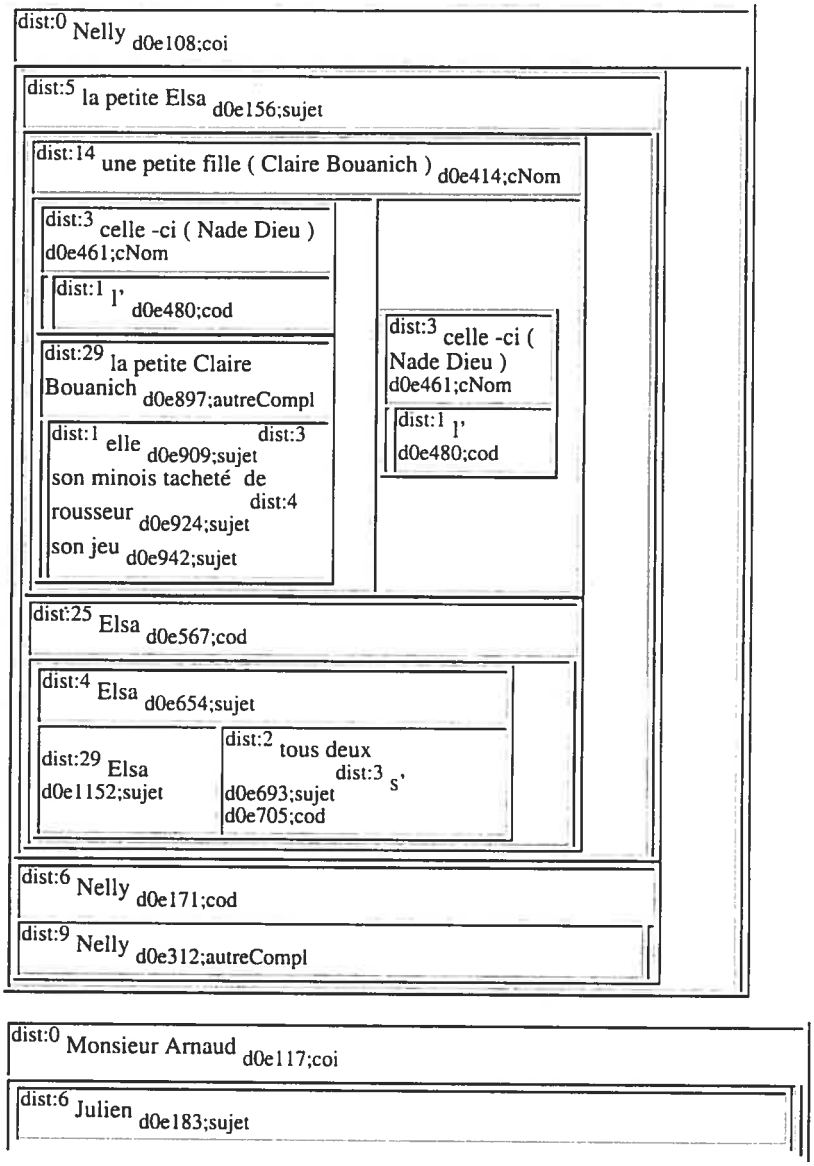
dist:0 l' acquisition d' ICG de Petro-Canada d0e229;cod

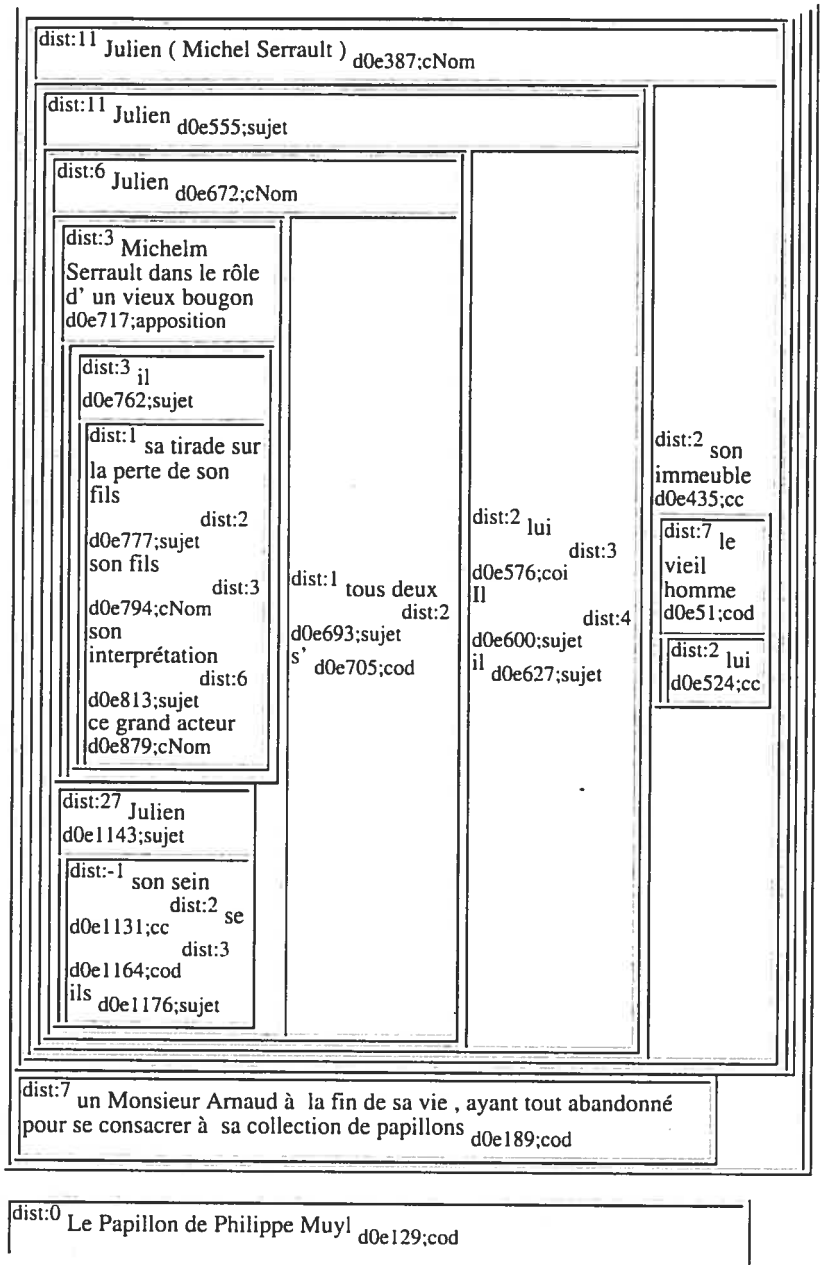
dist:0 Petro-Canada d0e238;cNom

dist:0 une transaction conclue cette semaine d0e256;cNom

dist:0 cette semaine d0e257;cc

# Chaines de «papillon»




dist:11 Julien ( Michel Serrault ) d0e387;cNom
dist:11 Julien d0e555;sujet
dist:6 Julien d0e672;cNom
dist:3 Michelm Serrault dans le rôle d' un vieux bougon d0e717;apposition
dist:3 il d0e762;sujet
dist:1 sa tirade sur la perte de son fils dist:2 d0e777;sujet
son fils dist:3 d0e794;cNom
son interprétation dist:6 d0e813;sujet
ce grand acteur d0e879;cNom
dist:27 Julien d0e1143;sujet
dist:-1 son sein dist:2 se d0e1131;cc dist:3 d0e1164;cod
ils d0e1176;sujet
dist:1 tous deux dist:2 d0e693;sujet
s' d0e705;cod
dist:2 lui dist:3 d0e576;coi
Il dist:4 d0e600;sujet
il d0e627;sujet
dist:2 son immeuble d0e435;cc
dist:7 le vieil homme d0e51;cod
dist:2 lui d0e524;cc
dist:7 un Monsieur Arnaud à la fin de sa vie , ayant tout abandonné pour se consacrer à sa collection de papillons d0e189;cod
dist:0 Le Papillon de Philippe Muyl d0e129;cod

dist:11 sa belle histoire d0e291;sujet dist:13 un film sympathique , qui tente manifestement de nous réchauffer le coeur d0e350;cod

dist:-6 sa lenteur
dist:-5 ses non-dits
d0e69;cc
d0e81;cc

dist:42 le film
d0e854;cNom

dist:11 le film
d0e1025;sujet

dist:2 les personnages
d0e1076;cNom

dist:0 Philippe Muyl d0e137;cNom

dist:8 Philippe Muyl d0e258;sujet

dist:41 Philippe Muyl
d0e963;sujet

dist:2 sa belle histoire
d0e291;sujet

dist:0 tout d0e218;cod

dist:0 un Claude Sautet d0e279;cNom

dist:0 La mère de celle -ci ( Nade Dieu ) d0e450;sujet

dist:5 l' d0e515;cod dist:7 la mère d0e554;cNom

dist:0 l' oublie fréquemment à la sortie de l' école d0e48;

dist:NaN ce qui force le vieil homme à l' accueillir chez lui
d0e501;apposition

dist:0 Michelm Serrault dans le rôled' un vieux bougon d0e71;

dist:NaN ce d0e738;sujet

dist:0 le rôled' un vieux bougon d0e737;cNom

dist:0 beaux enfants d0e923;cNom

dist:0 Là où Philippe Muyl excelle d0e96;

dist:NaN c' d0e972;sujet

dist:0 le Vercors où le film a été tourné d0e996;cNom

dist:2 ce décor naturel d0e1026;cc

dist:2 cette nature attaquée par les braconniers d0e1077;cNom

dist:0 le Game Boy d0e1107;sujet

# Chaines de «KickStart-HOWTO-8»

dist:0 la machine à installer d0e38;cod

dist:0 la disquette d' amorçage RedHat d0e39;cc

dist:14 le processus d' installation RedHat d0e268;cNom

dist:10 la disquette d' amorçage d0e201;cod

dist:0 ENTRÉE d0e75;cod

dist:0 le programme SYSLINUX d0e90;cNom

dist:0 linux ks d0e105;cod

dist:0 tout ce que vous aurez à faire d0e147;attribut

dist:0 ce qui suit d0e226;coi

dist:0 KickStart d0e301;sujet

dist:2 il d0e324;sujet dist:3 il d0e352;sujet dist:5 il d0e376;sujet dist:7 lui d0e394;coi

dist:0 ce qu' il doit faire d0e316;cod

dist:0 votre carte réseau d0e364;cod

dist:4 son RQ d0e400;cod dist:5 son adresse mémoire d' entrée/sortie d0e412;cod

# Annexe F

# Coréférences en contexte :

Cette annexe contient les textes suivants[1] :

## Critique de film :

### Invasions barbares[2]

## *HOWTO Linux* :

### Oracle Database HOWTO[3]

## Critique de film :

### The Favourite Game[4]

[1] Ces textes ont été balisés automatiquement.
[2] Tiré de [16].
[3] Tiré de [17].
[4] Tiré de [8].

# Contexte

| | |
|---|---|
| | **invasions** |
| | Comme la plupart de ses personnages sujet , l' intelligence de le Déclin col est à le rendez-vous dans les Invasions col barbares . |
| | Mais la causticité de le premier volet est ici remplacée par la gravité , le sérieux et la densité . |
| L' Arcand | **L' Arcand sujet nouveau** |
| | Rarement le public d' ici aura -t-il sujet eu l' impression de connaître un film sans sujet l' avoir encore vu tant la rumeur médiatique entourant le lancement de les Invasions col barbares s' est faite insistante . |
| | L' effet Invasions sujet passera peut-être à l' histoire dans col notre petit monde de le spectacle comme l' exemple d' une saturation par un objet absent . |
| | Tellement d' encre a déjà coulé que le critique a l' impression que sa modeste |

| | |
|---|---|
| | contribution devient superflue . |
| | Il est vrai que , pour une fois , le film en sujet cause n' est pas un film sujet ordinaire . |
| Denys Arcand | Il s'agit , et je pèse mes mots , de le chef-d' œuvre de Denys Arcand cNom sujet . |
| Denys Arcand \|\| Denys Arcand Autant \|\| Gabriel Arcand \|\| -t-il | - Aussi - [ 10 ] Denys Arcand sujet : la mort joyeuse [ 11 ] Du Déclin à les Invasions cNom apposition [ 12 ] Les émotions de Denys Arcand Autant coi Le Déclin de ce l'empire américain dont on retrouve presque tous les personnages dans Les Invasions cNom sujet barbares - il ne manque que Geneviève Rioux sujet et Gabriel Arcand sujet - se voulait drôle, caustique et frivole , autant ce nouveau film sujet cherche -t-il sujet la gravité , le sérieux et la densité . |
| | Mais dans les deux cas , on retrouve la même intelligence à le service d' un scénario coi |

| | |
|---|---|
| | brillamment dialogué , l' un de les plus intelligents que notre cinéma ait produits . |
| Arcand | En fait , la première partie est trompeuse à cause de le ton sarcastique qu' Arcand [sujet] utilise pour décrire le système de santé québécois . |
| | Son travelling d' ouverture à travers un couloir d' hôpitaljonché de malades résume tous les discours alarmistes sur la question . |
| | Mais à travers les efforts de Sébastien [coi] ( Stéphane Rousseau [apposition] ) pour trouver une chambre convenable à son père [sujet] ( Rémy Girard [apposition] ) , se glissent plusieurs scènes [sujet] irrésistibles qui rappellent les meilleurs moments de le Déclin [coi] . |
| | La tirade bureaucratique prononcée par Lise Roy [coi] constitue en elle-même une pièce d' anthologie . |
| | Mais le film [sujet] ne se limite pas à les efforts d' un fils pour [coi] aider son |

| | |
|---|---|
| | père_à [sujet] s` éteindre dignement . |
| Arcand | Àtravers une cascade de séquences , Arcand [sujet] déroule son récit_d' [sujet] une manière remarquable , passant allégrement de le thème de la santé à le syndicalisme , puis de la question de la drogue à celle de l' amitié , pour finalement aborder , tout en demi-teintes , le thème final , la mort . |
| | Même les démonstrations d' amitié dans Les Invasions [coi] barbares n' ont pas la même portée que dans Le Déclin [coi] . |
| | Là , on était naturellement amis_à [sujet] cause d' une appartenance commune , la profession d' historien . |
| | Ici , ce lien d' amitié paraît porter le poids d' une longue fréquentation . |
| | Claude ( Yves Jacques [apposition] ) est venu de Rome pour [coi] se retrouver à le chevet de Rémy [coi] . |
| | Louise ( Dorothée Berryman [apposition] ) met de côté ses griefs à l' égard de Rémy pour [coi] |

| | |
|---|---|
| | ne retrouver que l' homme qu' sujet elle cc a aimé et le père de sujet ses enfants sujet . |
| | D' admirables personnages sujet enrichissent aussi le récit sujet , celui en particulier de Johanne Marie Tremblay coi ( soeur Constance sujet ) qui apporte la note spiritualiste de rigueur . |
| Arcand | Mais , bien sûr , la grande innovation de les Invasions coi , c' est la place qu' Arcand sujet réserve à la nouvelle génération . |
| | Ce thème , même déjà abordé dans un film coi moins personnel comme Love sujet and Human Remains sujet , prend ici toute sa signification . |
| | Les personnages sujet défendus par Stéphane Rousseau coi , Marie-Josée Croze sujet et Isabelle Blais sujet témoignent d' un esprit nouveau . |

| | |
|---|---|
| Arcand | C' est sans doute dans l' espoir qu' ils sujet incarnent que se situe le mieux le changement chez Arcand cor . |
| un Arcand | Mais à la vérité , c' est avant tout d' un Arcand cor nouveau que témoigne ce film sujet . |
| | Le cynique d' hier s' est converti . |
| | Mais sa foi ne sera jamais celle de le charbonnier . |
| | On dirait qu' une barrière retenant l' émotion chez lui cor s' est ouverte . |
| | La tendresse qu' on sentait surtout dans le personnage d' Yves Jacques dans Le Déclin cNom cNom cor colore ici l' ensemble de le film cor . |
| Arcand | C' est d' un regard serein et mûri qu' Arcand sujet observe sa société et porte sur elle cor un jugement nuancé . |
| | Dans La Maudite Galette cor , Réjeanne Padovani sujet et même Gina sujet , tout le monde ou presque était pourri . |

| | |
|---|---|
| | Dans Le Déclin coi encore , c' était l' époque de le carpe diem , chacun pour soi , le plaisir avant tout . |
| Arcand | Depuis Jésus de Montréal cNom coi Arcand sujet est sensible à la souffrance de les autres . |
| | Mais ici , dans Les Invasions apposition barbares , sa compassion déteint sur l' ensemble de les personnages coi . |
| | C' est pourquoi , traitant de la mort , atteint -il sujet de les accents inouïs et touche -t-il sujet à de les sommets inoubliables . |

# Contexte

| | |
|---|---|
| | **Oracle-HOWTO-2** |
| | **2 . Installation de le logiciel Oracle** cNom sujet |
| | **2 . 1 Préparation de le Serveur** cNom sujet |
| un Utilisateur Oracle | **Création d' un Utilisateur Oracle** cNom sujet |
| Oracle | Nous avons évidemment besoin d' un utilisateur pour coi maintenir la base de données Oracle sujet . |
| le noyau Oracle Oracle | Comme nous n' avons l' intention de relier le noyau Oracle sujet ( plus sur ceci plus tard ) , nous devons accepter les noms d' utilisateur coi et de groupe par défaut d' Oracle coi . |
| Il | Il sujet inclut l' utilisateur ORACLE sujet et le groupe DBA sujet . |
| | 1 . Se connecter comme root |
| | 2 . Créer l' utilisateur sujet oracle et le groupe dba . |
| | `$ groupadd dba$ useradd oracle` |
| | 3 . S' assurer que le répertoire sujet personnel est créé pour l' utilisateur coi oracle . |
| | `$ mkdir /home/oracle$ mkdir /home/oracle/7.3.3.0.0 (Version of Oracle)$ chown -R oracle.dba /home/oracle` |

| | |
|---|---|
| | 2 . 2 Installation depuis le CDROM cNom sujet |
| l' Installateur Oracle sur le CD SCO le CD SCO | Malheureusement , l' Installateur Oracle sur le CD SCO cNom sujet ne marchera pas . |
| | Beaucoup de problèmes peut être rencontrés , des core dumps à les blocages . |
| | On doit donc copier les fichiers de le CDROM cNom sujet manuellement et les décompresser : |
| | ( S' assurer que le CDROM sujet est monté sur le système coi ) . |
| utilisateur Oracle | 1 . Se connecter comme utilisateur Oracle sujet |
| | 2 . Changer de répertoire pour /home/oracle/7 cNom coi . 3 . 3 . 0 . 0 . |
| | 3 . Copier les fichiers d' sujet installation de le CDROM coi |
| | `$ cp -a /mnt/cdrom/* .` |
| les fichiers Oracle de | **4 . Décompresser sujet les fichiers Oracle de cc le CDROM cc .** |
| | `$ find .- name *_ -exec ~/7.3.3.0.0/orainst/oiuncomp { } \ ;` |
| | **2 . 3 Tâches de Post Installation cNom sujet** |
| | **Tâches pour Root cNom sujet** |

| | |
|---|---|
| Oracle | Ajouter les lignes suivantes dans /etc/profile coi ou dans . profile pour chaque utilisateur d' Oracle cNom coi . |
| | `# Oracle Specific`<br>`ORACLE_HOME=/home/oracle/7.3.3.0.0`<br>`ORACLE_SID=orcl`<br>`ORACLE_TERM=vt100export ORACLE_HOME`<br>`ORACLE_SID ORACLE_TERM # Changer le`<br>`chemin pour Oracle`<br>`PATH="$PATH:$ORACLE_HOME/bin "` |
| l' utilitaire Oracle d' | Nous devons aussi changer le propriétaire et les permissions de l' utilitaire Oracle d' coi augmentation de ulimit . |
| | `$ chown root.root`<br>`$ORACLE_HOME/bin/osh$ chmod u+s`<br>`$ORACLE_HOME/bin/osh` |
| | **Tâches pour Oracle cNom sujet** |
| | Changer les permissions pour les fichiers Oracle pour coi s' assurer de leur bonne exécution . |
| | `$ chmod +x $ORACLE_HOME/bin/*$ chmod`<br>`u+s $ORACLE_HOME/bin/oracle` |
| | Les outils Oracle sujet demandent que les messages soient dans le répertoire coi $ ORACLE_HOME/tool_name/mesg sujet . |

## Contexte

| | favoriteGame |
|---|---|
| | **Portrait** sujet **du poète en play-boy** |
| | Il n' y a pas de poète de fin de semaine ni de poète à temps partiel . |
| | Le poète vit hors du temps et hume le monde comme un homme sujet hume une femme pour sujet capter son essence . |
| | C' est son jeu favori , si l' on accepte le terme « jeu » , puisque ce jeu est sa vie même . |
| | Dans le film The Favourite Game coref , inspiré du roman de jeunesse de Leonard Cohen coref , le réalisateur Bernar Hébert sujet voulait quant à lui coref saisir l'essence du poète sans pour autant se substituer à lui coref -sans coref faire de poésie à sa place- et il y parvient très bien . |
| Léo | Hébert sujet offre une plongée intime vers les sources d'inspiration de |

| | |
|---|---|
| | Léo cor ( JR Bourne apposition ) , jeune homme en quête de cNom sujet ses souvenirs et de sa place dans le monde . |
| Léo | En fait , Léo sujet cherche plus que ses souvenirs ; il est plutôtà l'écoute de leur résonance dans sa mémoire et travaille en véritable archéologue de sa psyché . |
| il | S' il sujet butine de femme en femme avec cNom cor succès -le sujet poète est oisif et à ce je-ne-sais-quoi qui les fait tomber- c' est parce qu' elles sujet l'aident à remonter le fil jusqu' à cette image cor marquante de l'enfance , décisive , délicieusement traumatique : le corps nu d' une fillette dévoilé lors d' un jeu de docteur , vision qui agit sur lui cor comme une révélation de la beauté pure . |
| | Tout le reste n' est , au fond , qu' une tentative de retrouver cette émotion première . |
| | Mais , paradoxalement , c' est justement cette distance dans le temps , la lente |

| | |
|---|---|
| | décantation dans l'esprit de ce choc initial ayant contaminé pour toujours son regard , qui le fait poète . |
| Léo | Avec les femmes cor . Léo sujet est cruel sans le vouloir . |
| Il il il | Il sujet les amène au bord de l'abandon , mais il sujet a rapidement besoin d'air pour retourner à son jeu favori et pour vraiment comprendre ce qu' il sujet ressent . |
| | Ainsi , Tamara apposition ( Sabine Karsenti apposition ) , goûte à sa médecine , puis Shell sujet ( Michèle Barbara Pelletier apposition ) , avec qui il sujet vit la relation la plus signifiante , ce qu' il sujet comprend trop tard . |
| | Ses fréquents allers-retours , entre New York cor et Montréal sujet , entre absence et présence , ont eu raison de cet amour qu' il sujet aurait étiré plus que les autres . |
| | Il sujet finit par retrouver la fillette à l'origine de ses fantasmes , devenue |

| | |
|---|---|
| | femme . |
| | Il sujet a beau la séduire , ce n' est évidemment plus la même chose . |
| | Il lui sujet faudra aller plus loin , jusqu' avant la création du traumatisme « esthétique » qu' elle sujet a provoqué en lui coi , jusqu' à l'innocence d' avant le jeu du docteur : il se souvient d' un autre jeu qui consiste à tomber dans la neige pour ensuite observer la trace que notre corps a laissée , cependant que notre esprit y voit la forme d' un ange . |
| Léo | JR Bourne sujet prête à Léo coi sa beauté froide et son regard métallique où il est difficile de distinguer la naïveté de l'indifférence . |
| Léo | C' est qu' il sujet joue sur les deux registres : Léo sujet est naïvement indifférent à la réalité des autres , trop occupé à analyser la sienne . |
| Il | Il sujet abandonne tout aussi aisément un job qu' une femme sujet , sans vraiment comprendre les reproches qu' on lui sujet |

| | |
|---|---|
| | fait . |
| il il | Il est extraterrestre , c'est-à-dire qu' il sujet ne touche pas le terre-à-terre , ce qui explique probablement sa forte sensualité et son penchant charnel , qui sont sa forme de contact avec ces êtres dont il sujet ne comprend pas les aspirations . |
| | Bernar Hébert sujet a créé ici un « ovni » cinématographique , c'est-à-dire un film sujet comme on en voit peu et dont on pourrait jurer l'avoir vu sans nécessairement être en mesure d'expliquer la trace qu' il sujet laisse chez le spectateur . |
| | Le réalisateur sujet a soigneusement gommé toutes les références à une époque précise et nous pourrions tout aussi bien être en 1960 qu' en 2000 . |
| Léo | La caméra sujet est sage , mis à part une envolée particulièrement belle , celle où Léo sujet observe la fillette , un instant qui nous marque autant que lui . |

| | |
|---|---|
| | Il est rare de voir des films qui[sujet] abordent , de l'intérieur , l'écriture ou la genèse d' une oeuvre[coi] , et encore plus rare que la tentative se fasse sans tomber dans le cliché du poète torturé ou dans la chronologie des événements biographiques d' un auteur célèbre . |
| Léo | Le seul cliché qu' on pourrait reprocher à Hébert[coi] , et ce , même si la quête[sujet] du personnage[sujet] principal passe par les sens , c' est d'abuser un peu trop de la chemine ouverte sur la poitrine nue de Léo[coi] . |
| | On avait déjà compris que les poètes sont irrésistibles . . . |

# Annexe G

# Fréquence des noms communs

Ceci est la liste de noms communs qui ont été étiquetés manuellement dans le corpus d'entraînement. Ces mots ont été classés selon le type de texte dans lequel ils se trouvent, ainsi que selon le nombre d'occurrences (noté en gras) dans chacun des sous-corpus étudiés. Il est à noter que cette liste a été générée automatiquement. C'est pourquoi les caractères accentués sont ordonnés après l'alphabet normal.

## G.1 Fusions de compagnies

**21 :** fusion

**19 :** milliards

**17 :** offre

**14 :** millions

**10 :** compagnie

**7 :** compagnies ; proposition

**6 :** acquisition ; organismes

**5 :** actionnaires

**4 :** 2003

**3 :** 1984 ; accord ; côté ; entente ; groupe ; secteur ; sociétés ; succursales

**2 :** 1996 ; assureur ; autorités ; chef ; conseil ; directeur ; fabricant ; marché ; projet ; règles ; société

**1 :** 1994 ; 1997 ; 1998 ; 2002 ; 2004 ; 2026 ; actif ; actions ; activités ; administrateurs ; analystes ; année ; audiences ; automne ; avis ; avoirs ; banque ; bureau ; bénéfice ; choix ; clients ; compte ; conseillers ; conseils ; considérations ; consolidation ; coupures ; création ; côte ; de ; dernière ; dirigeant ; débentures ; dépôts ; effets ; employés ; entité ; entreprises ; experts ; feu ; file ; filiale ; fin ; financier ; financières ; firme ; fournisseurs ; fusions ; groupement ; groupes ; guichet ; général ; hypothèque ; identité ; imprimantes ; initiales ; intégration ; jour ;

leader ; maison ; maisons ; mariage ; milliard ; monopole ; nom ; options ; organisme ; placement ; poupée ; primes ; programme ; président ; rapprochement ; recette ; rendement ; rivale ; semaine ; système ; taille ; tentative ; titre ; transaction ; transparence ; tribunal ; voisins ; vert ; échéancier ; émission ; éventualité

## G.2 Critiques de films

**41 :** film

**12 :** personnages

**11 :** homme

**10 :** fils ; histoire

**8 :** scénario

**7 :** femme ; père ; réalisatrice ; récit

**5 :** famille ; héros ; réalisateur ; scènes

**4 :** dernier ; dialogues ; films ; fois ; groupe ; intrigue ; millions ; mère ; vie

**3 :** acteurs ; caméra ; couple ; dernière ; dynamique ; en ; images ; mari ; moine ; personnage ; tout

**2 :** acteur ; actrice ; amant ; catastrophe ; cinéaste ; collection ; compagnons ; dame ; de ; documentaire ; enfant ; enquête ; fiancée ; fille ; finale ; flic ; gens ; hommes ; long ; metteur ; musique ; ménage ; métrage ; oeuvre ; partition ; pertinence ; petit-fils ; plans ; productrice ; projet ; président ; pères ; questions ; quête ; rencontre ; rêves ; rôle ; scène ; spectateur ; séquence ; temps ; traits ; vision ; volet ; écrivain ; équipe ; êtres

**1 :** % ; 1945 ; 1964 ; 1983 ; 1985 ; 4 ; accident ; action ; ado ; allure ; amalgame ; ami ; amie ; amis ; amitié ; androgyne ; anniversaire ; ans ; appart ; approche ; architecte ; arrivée ; as ; assassin ; assistance ; attitude ; attraction ; auteure ; avant-première ; avocat ; bac ; beau ; blessée ; blockbuster ; boxeur ; brigand ; bronzés ; cadres ; cas ; catastrophes ; chefs-d'oeuvre ; chien ; chinoise ; chose ; chute ; château ; chèque ; cinéma ; cinémascope ; classique ; clients ; collègues ; coloc ; commandant ; communauté ; commune ; compagnie ; compassion ; complicité ; comportement ; condamné ; consentement ; contribution ; copie ; coproducteur ; correspondance ; couleur ; coup ; cousins ; critique ; culture ; cynique ; célibataires ; côté ; danseuse ; demi-sous-sol ; descendants ; dessinateur ; discours ; douce ; début ; décor ; décors ; détenteur ; effet ; effets ; efforts ; enfants ; engin ; entreprise ; envolées ; ersatz ; euro ; exercice ; exploit ; faiblesses ; façon ; fer ; festival ; fidélité ; figurants ; filiations ; filles ; fin ; foi ; folie ; fonctions ; franchise ; français ; frères ; futilité ; féru ; gadjo ; gaillards ; garçon ; genre ; gestes ; gouvernante ; goût ; grand-mère ; grand-père ; griefs ; guerre ; génie ; génération ; hamburger ; hauteur ; heure ; hypothèse ; idoles ; immersion ; immeuble ; immigrant ; individu ; initiales ; innovation ; instants ; instructeur ; instruments ; interprète ; interprètes ; interprétation ; jeu ; jeunes ; jours ; justicier ; leader ; lenteur ; liberté ; lien ; liens ; maestro ; magie ; maisonnée ; maladresse ; maman ; marque ; membres ; menteur ; mentor ; mer ; message ; mesure ; minois ; mission ; mobilité ; modèle ; monde ; morale ; mort ; motel ; mots ; muse ; musiciens ; mère-grand ; narrateur ; nature ; nazi ; non-dits ; nouvelles ; oeuvres ; opinions ; opus ; papa ; pare-balles ; parents ; pari ; paroles ; part ; parti ; partie ; patriarche ; perles ; pesanteur ; petite ; petite-fille ; petits ; petits-fils ; philosophie ; phénomène ; plage ; plaisir ; point ; portrait ; potentiel ; potes ; poudings ; premier ; première ; prestation ; pris ; prise ; prochain ; professeur ; progression ; propension ; prostituée ; protagonistes ; pré-ado ; préoccupation ; pseudo-tuerie ; public ; péril ; question ;

quotidien ; radar ; rang ; rapper ; rappeur ; rebelle ; recrue ; remarques ; repères ; roman ; réalisation ; réalités ; référence ; réussite ; rôles ; sbires ; scénariste ; sein ; semaine ; signification ; site ; situations ; société ; soldats ; solitaire ; solitude ; sonore ; sous-produit ; spéciaux ; succès ; superhéros ; supérieurs ; survivants ; talent ; tenancier ; thème ; tirade ; ton ; touches ; trame ; travail ; travelling ; trio ; télescopage ; témoignage ; témoignages ; témoin ; usine ; vacances ; vengeance ; version ; vieillesse ; virages ; volonté ; vérité ; yeux ; âge ; écrans ; écurie ; élus ; élève ; élèves ; être

## G.3 *HOWTO Linux*

**101 :** fichier

**72 :** système

**57 :** pilote

**54 :** fichiers

**44 :** répertoire

**39 :** disque

**34 :** de

**30 :** commande ; partitions

**27 :** partition

**26 :** programme

**21 :** serveur

**19 :** lecteur

**18 :** disquette ; noyau

**16 :** interface

**15 :** installation ; section ; type ; utilisateur

**14 :** machine

**13 :** client

**12 :** carte ; disquettes ; distribution ; dur ; répertoires ; terminal

**11 :** document ; données ; page ; systèmes

**10 :** script ; version

**9 :** base ; lecteurs

**8 :** matériel ; nom ; options ; paquetage

**7 :** 0x340 ; applications ; archive ; cas ; contrôleurs ; distributions ; programmes ; utilitaire ; étape

**6 :** configuration ; contrôleur ; langage ; ligne ; message ; paquetages ; problème ; variable ; étapes

**5 :** adresse ; commandes ; disques ; démon ; groupe ; place ; périphérique ; périphériques ; site ; sources ; versions

**4 :** article ; bibliothèque ; bibliothèques ; documentation ; format ; informations ; lien ; logiciels ; modules ; noyaux ; outils ; paramètre ; paramètres ; sous-répertoires ; station ; touche ; travail ; utilitaires

**3 :** 0 ; 1 ; 5 ; archives ; bus ; clients ; code ; endroit ; forum ; gestionnaire ; images ; lignes ; logiciel ; messages ; mode ; modèle ; méthode ; numéro ; option ; ordinateur ; pages ; pilotes ; port ; session ; souris ; utilisation ; écran

**2 :** 2 ; 3 ; 4 ; application ; auteur ; besoins ; binaires ; canal ; choix ; compilation ; compte ; dernier ; emplacement ; entrée ; erreurs ; exemple ; façon ; forums ; frontières ; genre ; hôtes ; image ; installations ; liste ; machines ; mails ; mandataire ; menu ; modem ; moment ; méthodes ; noms ; notation ; opération ; processeurs ; processus ; protocole ; questions ; réponse ; serveurs ; services ; sites ; test

**1 :** #1 ; #2 ; (dés-)installation ; 0x ; 0x230 ; 0x320 ; 10 ; 1024 ; 11 ; 150-152 ; 203 ; 204 ; 21 ; 22 ; 26 ; 27 ; 28 ; 3. ; 3.5 ; 4096 ; 8 ; 82 ; accès ; accélérateur ; adaptateur ; affirmative ; aides ; annonce ; arborescence ; architecture ; auteurs ; autochargeur ; avertissement ; bases ; broches ; bug ; but ; canaux ; cartes ; cavaliers ; champ ; chemin ; chip ; cibles ; communauté ; compatibilité ; complications ; composant ; configurations ; connecteur ; connexions ; console ; contenu ; contributions ; contôleurs ; convertisseur ; copyright ; câbles ; d' ; d'affichage ; d'environnement ; daemon ; daemons ; dernière ; diffusion ; disque(s) ; disquette(s) ; distributeurs ; documentations ; documents ; domaines ; démarrage ; dépendance ; déplacement ; désirs ; détection ; enregistrement ; entourage ; entrées ; erreur ; essai ; exemples ; existence ; exploitation ; expérience ; extension ; extensions ; exécutables ; exécution ; faille ; faire ; fenêtres ; fois ; fonction ; formats ; gestion ; gourous ; groupes ; guide ; géométrie ; humeur ; hôte ; identificateur ; identité ; indications ; information ; interfaces ; jours ; jumpers ; kit ; librairie ; magazine ; mail ; manuel ; marchent ; marcher ; matériels ; million ; modification ; modifications ; moniteur ; moniteurs ; mots ; mécanisme ; mémoire ; navigateurs ; news ; niveau ; norme ; occupation ; optimiseur ; ordre ; outil ; pare-feu ; part ; particularité ; partie ; parties ; partition(s) ; passe ; pays ; performances ; plateforme ; points ; portion ; portmapper ; processeur ; procédure ; procédé ; produit ; projet ; propos ; pseudo-racine ; repartitionnement ; restriction ; revendeur ; réglage ; réseaux ; résultat ; shell ; simplicité ; solution ; son ; sous ; sous-répertoire ; spécificités ; stations ; style ; suivantes ; super-utilisateur ; superutilisateur ; surprise ; swap ; séquence ; table ; temps ; terme ; terminaux ; thèse ; tout ; type(s) ; usage ; utilisateurs ; valeurs ; vie ; vitesse ; vérifications ; web ; widgets ; zone ; écrans ; écritures ; éloge

## Annexe H

# Fréquence des syntagmes nominaux dans les chaînes coréférentielles

# H.1 Fusions de compagnies

## H.1.1 S.N. appartenant à une chaîne coréférentielle (de longueur > 1)

TAB. H.1 – S.N. APPARTENANT À UNE CHAÎNE CORÉFÉRENTIELLE – FUSIONS DE COMPAGNIES

| | Tête = nom | | Tête = Npr | | Tête = pro | | Tous | |
|---|---|---|---|---|---|---|---|---|
| **Lexical** | Domaine 221<br>Autre 250 | | Humain 61<br>Entité 304<br>Lieu 64<br>Abstrait 0 | | Défini 2<br>Démonstratif 25<br>Possessif 1<br>Personnel 72 | | 372 | |
| **Micro** | Simp. | Comp. | Simp. | Comp. | Simp. | Comp. | Simp. | Comp. |
| Défini | 145 | 110 | 61 | 22 | 0 | 3 | 206 | 124 |
| Indéfini | 13 | 48 | 1 | 7 | 0 | 3 | 13 | 45 |
| Démonstratif | 46 | 7 | 0 | 0 | 0 | 0 | 46 | 7 |
| Possessif | 57 | 34 | 1 | 0 | 0 | 0 | 58 | 32 |
| Interrogatif | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Aucun | 9 | 1 | 304 | 33 | 89 | 5 | 402 | 38 |
| Total | 271 | 200 | 367 | 62 | 89 | 11 | 726 | 246 |
| **Macro** | | | | | | | | |
| Sujet | | 153 | | 166 | | 63 | | 379 |
| C.O.D | | 118 | | 9 | | 10 | | 126 |
| C.O.I | | 33 | | 13 | | 14 | | 59 |
| C.C. | | 47 | | 52 | | 4 | | 101 |
| Attribut | | 19 | | 2 | | 0 | | 18 |
| C. nom | | 91 | | 158 | | 1 | | 244 |
| Non réf. | | 2 | | 3 | | 0 | | 5 |
| Autre | | 3 | | 1 | | 2 | | 5 |
| Apposition | | 5 | | 25 | | 6 | | 35 |
| **Total** | | 471 | | 429 | | 100 | | 972 |

### H.1.2 Têtes de chaînes coréférentielles

TAB. H.2 – TÊTES DE CHAÎNES CORÉFÉRENTIELLES – FUSIONS DE COMPAGNIES

| | Tête = nom | | Tête = Npr | | Tête = pro | | Tous | |
|---|---|---|---|---|---|---|---|---|
| **Lexical** | Domaine 29<br>Autre 51 | | Humain 22<br>Entité 33<br>Lieu 22<br>Abstrait 0 | | Défini 0<br>Démonstratif 0<br>Possessif 0<br>Personnel 0 | | 149 | |
| **Micro** | Simp. | Comp. | Simp. | Comp. | Simp. | Comp. | Simp. | Comp. |
| Défini | 15 | 44 | 18 | 8 | 0 | 0 | 33 | 47 |
| Indéfini | 2 | 16 | 0 | 0 | 0 | 0 | 2 | 13 |
| Démonstratif | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Possessif | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Interrogatif | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Aucun | 3 | 0 | 35 | 16 | 0 | 0 | 38 | 16 |
| Total | 20 | 60 | 53 | 24 | 0 | 0 | 73 | 76 |
| **Macro** | | | | | | | | |
| Sujet | | 30 | | 19 | | 0 | | 48 |
| C.O.D | | 23 | | 2 | | 0 | | 20 |
| C.O.I | | 5 | | 1 | | 0 | | 6 |
| C.C. | | 8 | | 13 | | 0 | | 21 |
| Attribut | | 1 | | 0 | | 0 | | 1 |
| C. nom | | 13 | | 28 | | 0 | | 39 |
| Non réf. | | 0 | | 0 | | 0 | | 0 |
| Autre | | 0 | | 0 | | 0 | | 0 |
| Apposition | | 0 | | 14 | | 0 | | 14 |
| **Total** | | 80 | | 77 | | 0 | | 149 |

### H.1.3 Syntagmes anaphoriques

TAB. H.3 – SYNTAGMES ANAPHORIQUES – FUSIONS DE COMPAGNIES

| | Tête = nom | | Tête = Npr | | Tête = pro | | Tous | |
|---|---|---|---|---|---|---|---|---|
| **Lexical** | Domaine 151<br>Autre 169 | | Humain 2<br>Entité 1<br>Lieu 0<br>Abstrait 0 | | Défini 2<br>Démonstratif 25<br>Possessif 1<br>Personnel 72 | | 414 | |
| **Micro** | Simp. | Comp. | Simp. | Comp. | Simp. | Comp. | Simp. | Comp. |
| Défini | 113 | 55 | 1 | 0 | 0 | 3 | 114 | 54 |
| Indéfini | 3 | 6 | 0 | 1 | 0 | 3 | 3 | 7 |
| Démonstratif | 46 | 7 | 0 | 0 | 0 | 0 | 46 | 7 |
| Possessif | 54 | 34 | 1 | 0 | 0 | 0 | 55 | 32 |
| Interrogatif | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Aucun | 2 | 0 | 0 | 0 | 89 | 5 | 91 | 5 |
| Total | 218 | 102 | 2 | 1 | 89 | 11 | 309 | 105 |
| **Macro** | | | | | | | | |
| Sujet | | 109 | | 1 | | 63 | | 172 |
| C.O.D | | 76 | | 0 | | 10 | | 85 |
| C.O.I | | 19 | | 0 | | 14 | | 33 |
| C.C. | | 34 | | 0 | | 4 | | 37 |
| Attribut | | 12 | | 1 | | 0 | | 10 |
| C. nom | | 64 | | 1 | | 1 | | 64 |
| Non réf. | | 0 | | 0 | | 0 | | 0 |
| Autre | | 3 | | 0 | | 2 | | 4 |
| Apposition | | 3 | | 0 | | 6 | | 9 |
| **Total** | | 320 | | 3 | | 100 | | 414 |

### H.1.4 Syntagmes coréférentiels

TAB. H.4 – SYNTAGMES CORÉFÉRENTIELS – FUSIONS DE COMPAGNIES

| | Tête = nom | | Tête = Npr | | Tête = pro | | Tous | |
|---|---|---|---|---|---|---|---|---|
| **Lexical** | Domaine 33<br>Autre 28 | | Humain 36<br>Entité 252<br>Lieu 40<br>Abstrait 0 | | Défini 0<br>Démonstratif 0<br>Possessif 0<br>Personnel 0 | | 378 | |
| **Micro** | Simp. | Comp. | Simp. | Comp. | Simp. | Comp. | Simp. | Comp. |
| Défini | 13 | 11 | 39 | 14 | 0 | 0 | 52 | 23 |
| Indéfini | 7 | 24 | 1 | 6 | 0 | 0 | 7 | 23 |
| Démonstratif | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Possessif | 3 | 0 | 0 | 0 | 0 | 0 | 3 | 0 |
| Interrogatif | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Aucun | 1 | 1 | 251 | 17 | 0 | 0 | 252 | 17 |
| Total | 25 | 36 | 291 | 37 | 0 | 0 | 315 | 63 |
| **Macro** | | | | | | | | |
| Sujet | | 12 | | 133 | | 0 | | 144 |
| C.O.D | | 16 | | 5 | | 0 | | 16 |
| C.O.I | | 7 | | 10 | | 0 | | 16 |
| C.C. | | 4 | | 38 | | 0 | | 41 |
| Attribut | | 6 | | 1 | | 0 | | 7 |
| C. nom | | 14 | | 127 | | 0 | | 139 |
| Non réf. | | 1 | | 2 | | 0 | | 3 |
| Autre | | 0 | | 1 | | 0 | | 1 |
| Apposition | | 1 | | 11 | | 0 | | 11 |
| **Total** | | 62 | | 328 | | 0 | | 378 |

## H.2 Critiques de films

### H.2.1 S.N. appartenant à une chaîne coréférentielle (de longueur > 1)

TAB. H.5 – S.N. APPARTENANT À UNE CHAÎNE CORÉFÉRENTIELLE – CRITIQUES DE FILMS

| | Tête = nom | | Tête = Npr | | Tête = pro | | Tous | |
|---|---|---|---|---|---|---|---|---|
| **Lexical** | Domaine 223<br>Autre 301 | | Humain 194<br>Entité 91<br>Lieu 10<br>Abstrait 1 | | Défini 1<br>Démonstratif 53<br>Possessif 2<br>Personnel 150 | | 1012 | |
| **Micro** | Simp. | Comp. | Simp. | Comp. | Simp. | Comp. | Simp. | Comp. |
| Défini | 133 | 85 | 29 | 14 | 0 | 12 | 159 | 104 |
| Indéfini | 30 | 52 | 4 | 13 | 3 | 3 | 37 | 67 |
| Démonstratif | 36 | 27 | 4 | 2 | 0 | 0 | 40 | 29 |
| Possessif | 104 | 43 | 1 | 1 | 0 | 0 | 105 | 42 |
| Interrogatif | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Aucun | 6 | 8 | 188 | 40 | 177 | 14 | 371 | 58 |
| Total | 309 | 215 | 226 | 70 | 180 | 29 | 712 | 300 |
| **Macro** | | | | | | | | |
| Sujet | | 151 | | 111 | | 103 | | 359 |
| C.O.D | | 116 | | 24 | | 40 | | 179 |
| C.O.I | | 55 | | 25 | | 31 | | 110 |
| C.C. | | 59 | | 21 | | 11 | | 90 |
| Attribut | | 18 | | 5 | | 0 | | 22 |
| C. nom | | 95 | | 79 | | 7 | | 179 |
| Non réf. | | 4 | | 12 | | 1 | | 16 |
| Autre | | 10 | | 7 | | 6 | | 23 |
| Apposition | | 16 | | 12 | | 10 | | 34 |
| **Total** | | 524 | | 296 | | 209 | | 1012 |

### H.2.2 Têtes de chaînes coréférentielles

TAB. H.6 – TÊTES DE CHAÎNES CORÉFÉRENTIELLES – CRITIQUES DE FILMS

| | Tête = nom | | Tête = Npr | | Tête = pro | | Tous | |
|---|---|---|---|---|---|---|---|---|
| **Lexical** | Domaine 21<br>Autre 48 | | Humain 36<br>Entité 16<br>Lieu 4<br>Abstrait 1 | | Défini 0<br>Démonstratif 3<br>Possessif 0<br>Personnel 0 | | 127 | |
| **Micro** | | | | | | | | |
| | Simp. | Comp. | Simp. | Comp. | Simp. | Comp. | Simp. | Comp. |
| Défini | 12 | 22 | 8 | 7 | 0 | 1 | 20 | 29 |
| Indéfini | 10 | 22 | 0 | 4 | 0 | 0 | 10 | 25 |
| Démonstratif | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| Possessif | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Interrogatif | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Aucun | 1 | 1 | 24 | 14 | 0 | 2 | 25 | 17 |
| Total | 23 | 46 | 32 | 25 | 0 | 3 | 55 | 72 |
| **Macro** | | | | | | | | |
| Sujet | | 18 | | 13 | | 0 | | 31 |
| C.O.D | | 10 | | 5 | | 1 | | 16 |
| C.O.I | | 10 | | 8 | | 0 | | 18 |
| C.C.. | | 6 | | 6 | | 0 | | 12 |
| Attribut | | 2 | | 1 | | 0 | | 2 |
| C. nom | | 16 | | 18 | | 1 | | 34 |
| Non réf. | | 1 | | 1 | | 0 | | 2 |
| Autre | | 3 | | 1 | | 0 | | 4 |
| Apposition | | 3 | | 4 | | 1 | | 8 |
| **Total** | | 69 | | 57 | | 3 | | 127 |

### H.2.3 Syntagmes anaphoriques

TAB. H.7 – SYNTAGMES ANAPHORIQUES – CRITIQUES DE FILMS

| | Tête = nom | | Tête = Npr | | Tête = pro | | Tous | |
|---|---|---|---|---|---|---|---|---|
| **Lexical** | Domaine 169<br>Autre 225 | | Humain 6<br>Entité 1<br>Lieu 0<br>Abstrait 0 | | Défini 1<br>Démonstratif 49<br>Possessif 2<br>Personnel 150 | | 595 | |
| **Micro** | Simp. | Comp. | Simp. | Comp. | Simp. | Comp. | Simp. | Comp. |
| Défini | 115 | 58 | 5 | 0 | 0 | 11 | 117 | 64 |
| Indéfini | 11 | 7 | 0 | 0 | 3 | 3 | 14 | 10 |
| Démonstratif | 36 | 26 | 1 | 0 | 0 | 0 | 37 | 26 |
| Possessif | 93 | 40 | 0 | 1 | 0 | 0 | 93 | 39 |
| Interrogatif | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Aucun | 4 | 4 | 0 | 0 | 177 | 11 | 181 | 14 |
| Total | 259 | 135 | 6 | 1 | 180 | 25 | 442 | 153 |
| **Macro** | | | | | | | | |
| Sujet | | 119 | | 5 | | 103 | | 222 |
| C.O.D | | 91 | | 1 | | 39 | | 131 |
| C.O.I | | 40 | | 0 | | 31 | | 70 |
| C.C. | | 49 | | 0 | | 11 | | 59 |
| Attribut | | 6 | | 0 | | 0 | | 6 |
| C. nom | | 69 | | 0 | | 5 | | 73 |
| Non réf. | | 2 | | 1 | | 1 | | 3 |
| Autre | | 6 | | 0 | | 6 | | 12 |
| Apposition | | 12 | | 0 | | 9 | | 19 |
| **Total** | | 394 | | 7 | | 205 | | 595 |

### H.2.4 Syntagmes coréférentiels

TAB. H.8 – SYNTAGMES CORÉFÉRENTIELS – CRITIQUES DE FILMS

| | Tête = nom | | Tête = Npr | | Tête = pro | | Tous | |
|---|---|---|---|---|---|---|---|---|
| **Lexical** | Domaine 32<br>Autre 21 | | Humain 144<br>Entité 63<br>Lieu 6<br>Abstrait 0 | | Défini 0<br>Démonstratif 1<br>Possessif 0<br>Personnel 0 | | 264 | |
| **Micro** | Simp. | Comp. | Simp. | Comp. | Simp. | Comp. | Simp. | Comp. |
| Défini | 3 | 2 | 15 | 6 | 0 | 0 | 18 | 8 |
| Indéfini | 8 | 22 | 3 | 8 | 0 | 0 | 11 | 30 |
| Démonstratif | 0 | 0 | 3 | 2 | 0 | 0 | 3 | 2 |
| Possessif | 11 | 3 | 1 | 0 | 0 | 0 | 12 | 3 |
| Interrogatif | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Aucun | 1 | 3 | 150 | 25 | 0 | 1 | 151 | 26 |
| Total | 23 | 30 | 172 | 41 | 0 | 1 | 195 | 69 |
| **Macro** | | | | | | | | |
| Sujet | | 12 | | 89 | | 0 | | 100 |
| C.O.D | | 11 | | 16 | | 0 | | 27 |
| C.O.I | | 3 | | 17 | | 0 | | 20 |
| C.C. | | 4 | | 15 | | 0 | | 19 |
| Attribut | | 10 | | 1 | | 0 | | 11 |
| C. nom | | 10 | | 60 | | 1 | | 71 |
| Non réf. | | 1 | | 1 | | 0 | | 2 |
| Autre | | 1 | | 6 | | 0 | | 7 |
| Apposition | | 1 | | 8 | | 0 | | 7 |
| **Total** | | 53 | | 213 | | 1 | | 264 |

## H.3 *HOWTO Linux*

### H.3.1 S.N. appartenant à une chaîne coréférentielle (de longueur > 1)

TAB. H.9 – S.N. APPARTENANT À UNE CHAÎNE CORÉFÉRENTIELLE – HOWTO LINUX

| | Tête = nom | | Tête = Npr | | Tête = pro | | Tous | |
|---|---|---|---|---|---|---|---|---|
| **Lexical** | Domaine 705<br>Autre 356 | | Humain 0<br>Entité 580<br>Lieu 0<br>Abstrait 430 | | Défini 1<br>Démonstratif 177<br>Possessif 0<br>Personnel 265 | | 2468 | |
| **Micro** | Simp. | Comp. | Simp. | Comp. | Simp. | Comp. | Simp. | Comp. |
| Défini | 317 | 181 | 88 | 35 | 0 | 8 | 395 | 215 |
| Indéfini | 129 | 91 | 31 | 16 | 0 | 1 | 154 | 97 |
| Démonstratif | 188 | 14 | 9 | 1 | 0 | 0 | 197 | 15 |
| Possessif | 72 | 17 | 13 | 1 | 0 | 0 | 81 | 18 |
| Interrogatif | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| Aucun | 37 | 14 | 747 | 69 | 416 | 18 | 1200 | 95 |
| Total | 744 | 317 | 888 | 122 | 416 | 27 | 2028 | 440 |
| **Macro** | | | | | | | | |
| Sujet | | 203 | | 148 | | 221 | | 566 |
| C.O.D | | 394 | | 313 | | 145 | | 834 |
| C.O.I | | 61 | | 50 | | 28 | | 138 |
| C.C. | | 167 | | 136 | | 18 | | 314 |
| Attribut | | 9 | | 23 | | 5 | | 34 |
| C. nom | | 164 | | 240 | | 6 | | 404 |
| Non réf. | | 45 | | 50 | | 0 | | 93 |
| Autre | | 14 | | 26 | | 13 | | 53 |
| Apposition | | 4 | | 24 | | 7 | | 32 |
| **Total** | | 1061 | | 1010 | | 443 | | 2468 |

### H.3.2 Chaînes coréférentielles

TAB. H.10 – CHAÎNES CORÉFÉRENTIELLES – HOWTO LINUX

| | Tête = nom | | Tête = Npr | | Tête = pro | | Tous | |
|---|---|---|---|---|---|---|---|---|
| **Lexical** | Domaine 168<br>Autre 84 | | Humain 0<br>Entité 44<br>Lieu 0<br>Abstrait 85 | | Défini 0<br>Démonstratif 0<br>Possessif 0<br>Personnel 0 | | 369 | |
| **Micro** | | | | | | | | |
| | Simp. | Comp. | Simp. | Comp. | Simp. | Comp. | Simp. | Comp. |
| Défini | 73 | 76 | 10 | 5 | 0 | 0 | 81 | 76 |
| Indéfini | 29 | 47 | 4 | 4 | 0 | 0 | 32 | 48 |
| Démonstratif | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| Possessif | 8 | 2 | 0 | 0 | 0 | 0 | 8 | 2 |
| Interrogatif | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| Aucun | 10 | 5 | 89 | 17 | 0 | 0 | 99 | 21 |
| Total | 122 | 130 | 103 | 26 | 0 | 0 | 222 | 147 |
| **Macro** | | | | | | | | |
| Sujet | | 45 | | 14 | | 0 | | 58 |
| C.O.D | | 96 | | 32 | | 0 | | 126 |
| C.O.I | | 12 | | 10 | | 0 | | 21 |
| C.C. | | 40 | | 20 | | 0 | | 59 |
| Attribut | | 5 | | 9 | | 0 | | 11 |
| C. nom | | 35 | | 21 | | 0 | | 53 |
| Non réf. | | 11 | | 10 | | 0 | | 20 |
| Autre | | 6 | | 9 | | 0 | | 15 |
| Apposition | | 2 | | 4 | | 0 | | 6 |
| **Total** | | 252 | | 129 | | 0 | | 369 |

### H.3.3 Syntagmes anaphoriques

TAB. H.11 – SYNTAGMES ANAPHORIQUES – HOWTO LINUX

| | Tête = nom | | Tête = Npr | | Tête = pro | | Tous | |
|---|---|---|---|---|---|---|---|---|
| **Lexical** | Domaine 115<br>Autre 155 | | Humain 0<br>Entité 2<br>Lieu 0<br>Abstrait 5 | | Défini 1<br>Démonstratif 177<br>Possessif 0<br>Personnel 265 | | 1012 | |
| **Micro** | Simp. | Comp. | Simp. | Comp. | Simp. | Comp. | Simp. | Comp. |
| Défini | 13 | 4 | 1 | 1 | 0 | 8 | 14 | 13 |
| Indéfini | 4 | 4 | 0 | 0 | 0 | 1 | 4 | 5 |
| Démonstratif | 186 | 14 | 1 | 0 | 0 | 0 | 187 | 14 |
| Possessif | 36 | 9 | 1 | 0 | 0 | 0 | 35 | 9 |
| Interrogatif | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Aucun | 0 | 0 | 2 | 1 | 416 | 18 | 418 | 19 |
| Total | 239 | 31 | 5 | 2 | 416 | 27 | 658 | 60 |
| **Macro** | | | | | | | | |
| Sujet | | 73 | | 0 | | 221 | | 294 |
| C.O.D | | 79 | | 2 | | 145 | | 224 |
| C.O.I | | 11 | | 0 | | 28 | | 39 |
| C.C. | | 51 | | 1 | | 18 | | 70 |
| Attribut | | 0 | | 0 | | 5 | | 5 |
| C. nom | | 53 | | 3 | | 6 | | 62 |
| Non réf. | | 1 | | 0 | | 0 | | 1 |
| Autre | | 2 | | 0 | | 13 | | 15 |
| Apposition | | 0 | | 1 | | 7 | | 8 |
| **Total** | | 270 | | 7 | | 443 | | 418 |

### H.3.4 Syntagmes coréférentiels

TAB. H.12 – SYNTAGMES CORÉFÉRENTIELS – HOWTO LINUX

| | Tête = nom | | Tête = Npr | | Tête = pro | | Tous | |
|---|---|---|---|---|---|---|---|---|
| **Lexical** | Domaine 379<br>Autre 99 | | Humain 0<br>Entité 515<br>Lieu 0<br>Abstrait 330 | | Défini 0<br>Démonstratif 0<br>Possessif 0<br>Personnel 0 | | 1293 | |
| **Micro** | | | | | | | | |
| | Simp. | Comp. | Simp. | Comp. | Simp. | Comp. | Simp. | Comp. |
| Défini | 210 | 97 | 75 | 29 | 0 | 0 | 279 | 122 |
| Indéfini | 85 | 33 | 24 | 12 | 0 | 0 | 104 | 37 |
| Démonstratif | 1 | 0 | 8 | 1 | 0 | 0 | 9 | 1 |
| Possessif | 28 | 5 | 12 | 1 | 0 | 0 | 38 | 6 |
| Interrogatif | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Aucun | 14 | 5 | 633 | 50 | 0 | 0 | 647 | 50 |
| Total | 338 | 140 | 752 | 93 | 0 | 0 | 1077 | 216 |
| **Macro** | | | | | | | | |
| Sujet | | 82 | | 134 | | 0 | | 211 |
| C.O.D | | 192 | | 257 | | 0 | | 437 |
| C.O.I | | 36 | | 40 | | 0 | | 76 |
| C.C. | | 71 | | 115 | | 0 | | 180 |
| Attribut | | 4 | | 14 | | 0 | | 18 |
| C. nom | | 74 | | 213 | | 0 | | 284 |
| Non réf. | | 11 | | 36 | | 0 | | 46 |
| Autre | | 6 | | 17 | | 0 | | 23 |
| Apposition | | 2 | | 19 | | 0 | | 18 |
| **Total** | | 478 | | 845 | | 0 | | 1293 |

## H.4 Tous les domaines

### H.4.1 S.N. appartenant à une chaîne coréférentielle (de longueur > 1)

TAB. H.13 – S.N. APPARTENANT À UNE CHAÎNE CORÉFÉRENTIELLE – TOUS LES DOMAINES

| | Tête = nom | | Tête = Npr | | Tête = pro | | Tous | |
|---|---|---|---|---|---|---|---|---|
| **Lexical** | Domaine 1149<br>Autre 907 | | Humain 255<br>Entité 975<br>Lieu 74<br>Abstrait 431 | | Défini 4<br>Démonstratif 255<br>Possessif 3<br>Personnel 487 | | 4452 | |
| **Micro** | Simp. | Comp. | Simp. | Comp. | Simp. | Comp. | Simp. | Comp. |
| Défini | 595 | 376 | 178 | 71 | 0 | 23 | 760 | 443 |
| Indéfini | 172 | 191 | 36 | 36 | 3 | 7 | 204 | 209 |
| Démonstratif | 270 | 48 | 13 | 3 | 0 | 0 | 283 | 51 |
| Possessif | 233 | 94 | 15 | 2 | 0 | 0 | 244 | 92 |
| Interrogatif | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| Aucun | 52 | 23 | 1239 | 142 | 682 | 37 | 1973 | 191 |
| Total | 1324 | 732 | 1481 | 254 | 685 | 67 | 3466 | 986 |
| **Macro** | | | | | | | | |
| Sujet | | 507 | | 425 | | 387 | | 1304 |
| C.O.D | | 628 | | 346 | | 195 | | 1139 |
| C.O.I | | 149 | | 88 | | 73 | | 307 |
| C.C. | | 273 | | 209 | | 33 | | 505 |
| Attribut | | 46 | | 30 | | 5 | | 74 |
| C. nom | | 350 | | 477 | | 14 | | 827 |
| Non réf. | | 51 | | 65 | | 1 | | 114 |
| Autre | | 27 | | 34 | | 21 | | 81 |
| Apposition | | 25 | | 61 | | 23 | | 101 |
| **Total** | | 2056 | | 1735 | | 752 | | 4452 |

### H.4.2 Chaînes coréférentielles

TAB. H.14 – CHAÎNES CORÉFÉRENTIELLES – TOUS LES DOMAINES

| | Tête = nom | | Tête = Npr | | Tête = pro | | Tous | |
|---|---|---|---|---|---|---|---|---|
| **Lexical** | Domaine 218<br>Autre 183 | | Humain 58<br>Entité 93<br>Lieu 26<br>Abstrait 86 | | Défini 0<br>Démonstratif 3<br>Possessif 0<br>Personnel 0 | | 645 | |
| **Micro** | Simp. | Comp. | Simp. | Comp. | Simp. | Comp. | Simp. | Comp. |
| Défini | 100 | 142 | 36 | 20 | 0 | 1 | 134 | 152 |
| Indéfini | 41 | 85 | 4 | 8 | 0 | 0 | 44 | 86 |
| Démonstratif | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 |
| Possessif | 8 | 2 | 0 | 0 | 0 | 0 | 8 | 2 |
| Interrogatif | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| Aucun | 14 | 6 | 148 | 47 | 0 | 2 | 162 | 54 |
| Total | 165 | 236 | 188 | 75 | 0 | 3 | 350 | 295 |
| **Macro** | | | | | | | | |
| Sujet | | 93 | | 46 | | 0 | | 137 |
| C.O.D | | 129 | | 39 | | 1 | | 162 |
| C.O.I | | 27 | | 19 | | 0 | | 45 |
| C.C. | | 54 | | 39 | | 0 | | 92 |
| Attribut | | 8 | | 10 | | 0 | | 14 |
| C. nom | | 64 | | 67 | | 1 | | 126 |
| Non réf. | | 12 | | 11 | | 0 | | 22 |
| Autre | | 9 | | 10 | | 0 | | 19 |
| Apposition | | 5 | | 22 | | 1 | | 28 |
| **Total** | | 401 | | 263 | | 3 | | 645 |

### H.4.3 Syntagmes anaphoriques

TAB. H.15 – SYNTAGMES ANAPHORIQUES – TOUS LES DOMAINES

| | Tête = nom | | Tête = Npr | | Tête = pro | | Tous | |
|---|---|---|---|---|---|---|---|---|
| **Lexical** | Domaine 435<br>Autre 549 | | Humain 8<br>Entité 4<br>Lieu 0<br>Abstrait 5 | | Défini 4<br>Démonstratif 251<br>Possessif 3<br>Personnel 487 | | 1727 | |
| **Micro** | Simp. | Comp. | Simp. | Comp. | Simp. | Comp. | Simp. | Comp. |
| Défini | 241 | 117 | 7 | 1 | 0 | 22 | 245 | 131 |
| Indéfini | 18 | 17 | 0 | 1 | 3 | 7 | 21 | 22 |
| Démonstratif | 268 | 47 | 2 | 0 | 0 | 0 | 270 | 47 |
| Possessif | 183 | 83 | 2 | 1 | 0 | 0 | 183 | 80 |
| Interrogatif | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Aucun | 6 | 4 | 2 | 1 | 682 | 34 | 690 | 38 |
| Total | 716 | 268 | 13 | 4 | 685 | 63 | 1409 | 318 |
| **Macro** | | | | | | | | |
| Sujet | | 301 | | 6 | | 387 | | 688 |
| C.O.D | | 246 | | 3 | | 194 | | 440 |
| C.O.I | | 70 | | 0 | | 73 | | 142 |
| C.C. | | 134 | | 1 | | 33 | | 166 |
| Attribut | | 18 | | 1 | | 5 | | 21 |
| C. nom | | 186 | | 4 | | 12 | | 199 |
| Non réf. | | 3 | | 1 | | 1 | | 4 |
| Autre | | 11 | | 0 | | 21 | | 31 |
| Apposition | | 15 | | 1 | | 22 | | 36 |
| **Total** | | 984 | | 17 | | 748 | | 1727 |

### H.4.4 Syntagmes coréférentiels

TAB. H.16 – SYNTAGMES CORÉFÉRENTIELS – TOUS LES DOMAINES

| | Tête = nom | | Tête = Npr | | Tête = pro | | Tous | |
|---|---|---|---|---|---|---|---|---|
| **Lexical** | Domaine 444<br>Autre 148 | | Humain 180<br>Entité 830<br>Lieu 46<br>Abstrait 330 | | Défini 0<br>Démonstratif 1<br>Possessif 0<br>Personnel 0 | | 1935 | |
| **Micro** | Simp. | Comp. | Simp. | Comp. | Simp. | Comp. | Simp. | Comp. |
| Défini | 226 | 110 | 129 | 49 | 0 | 0 | 349 | 153 |
| Indéfini | 100 | 79 | 28 | 26 | 0 | 0 | 122 | 90 |
| Démonstratif | 1 | 0 | 11 | 3 | 0 | 0 | 12 | 3 |
| Possessif | 42 | 8 | 13 | 1 | 0 | 0 | 53 | 9 |
| Interrogatif | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Aucun | 16 | 9 | 1034 | 92 | 0 | 1 | 1050 | 93 |
| Total | 386 | 206 | 1215 | 171 | 0 | 1 | 1587 | 348 |
| **Macro** | | | | | | | | |
| Sujet | | 106 | | 356 | | 0 | | 455 |
| C.O.D | | 219 | | 278 | | 0 | | 480 |
| C.O.I | | 46 | | 67 | | 0 | | 112 |
| C.C. | | 79 | | 168 | | 0 | | 240 |
| Attribut | | 20 | | 16 | | 0 | | 36 |
| C. nom | | 98 | | 400 | | 1 | | 494 |
| Non réf. | | 13 | | 39 | | 0 | | 51 |
| Autre | | 7 | | 24 | | 0 | | 31 |
| Apposition | | 4 | | 38 | | 0 | | 36 |
| **Total** | | 592 | | 1386 | | 1 | | 1935 |

# Bibliographie


[Bal97] Breck BALDWIN : CogNIAC : high precision coreference with limited knowledge and linguistic resources. *In Proceedings of ACL/EACL workshop on Operational factors in practical, robust anaphora resolution (ACL97)*, pages 38–45, Madrid, Espagne, 1997.

[Ber97] Sabine BERGLER : Towards reliable partial anaphora resolution. *In Proceedings of ACL/EACL workshop on Operational factors in practical, robust anaphora resolution (ACL97)*, Madrid, Espagne, 1997. Présenté à la «Human Language Technology Conference».

[BM01] Catalina BARBU et Ruslan MITKOV : Evaluation tool for rule-based anaphora resolution methods. *In Proceedings of ACL'01, Meeting of the Association for Computational Linguistics*, pages 34–41, Toulouse, France, 2001.

[BMB+95] Breck BALDWIN, Tom MORTON, Amit BAGGA, Jason BALDRIDGE, Raman CHANDRASEKER, Alexis DIMITRIADIS, Kieran SNYDER et Magdalena WOLSKA : University of Pennsylvania : Description of the University of Pennsylvania system used for MUC-6. *In Proceedings of the Sixth Message Understanding Conference (MUC-6)*, pages 177–191, 1995.

[BWK+03] Sabine BERGLER, Rene WITTE, Michelle KHALIFE, Zhuoyan LI et Frank RUDZICZ : Using knowledge-poor coreference resolution for text summarization. *In Workshop on Text Summarization, Document Understanding Conference (DUC 2003)*, Edmonton, Canada, 2003.

[Cor87] Francis CORBLIN : *Indéfini, défini et démonstratif : constructions linguistiques de la référence*. Librairie Droz, Paris, 1987.

[Cor95] Francis CORBLIN : *Les formes de reprise dans le discours*. Presses Universitaires de Rennes, Rennes, France, 1995.

[Coy02] Frank P. COYLE : *XML, Web Services, and the Data Revolution*. Addison-Wesley, Reading, MA, USA, 2002.

[DGG+94] Jean DUBOIS, Mathée GIACOMO, Louis GUESPIN, Christiane MARCELLESI, Jean-Baptiste MARCELLESI et Jean-Pierre MÉVEL : *Dictionnaire de linguistique et des sciences de langage*. Larousse, Paris, 1994.

[Dup02] Michel DUPONT : Une approche cognitive pour le calcul des chaînes de références. *In TALN2002, 9ème Conférence Annuelle, workshop : Chaînes de références et résolveurs d'anaphores. Nancy du 24 au 27 juin 2002. Tome 2*, pages 193–204, 2002.

[GP94] Marie-Noëlle GARY-PRIEUR : *Grammaire du nom propre*. Presses Universitaires de France, Paris, France, 1994.

[HC97] Lynette HIRSCHMANN et Nancy CHINCHOR : MUC-7 coreference task definition. *In Message Understanding Conference Proceedings*, 1997. `http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/co%_task.html`.

[HH76] M.A.K HALLIDAY et Ruqaiya HASAN : *Cohesion in English*. Longman, London, 7 édition, 1976.

[Hob78] Jerry R. HOBBS : Resolving Pronoun References. *Lingua*, 44:311–338, 1978.

[Jon94] Kerstin JONASSON : *Le nom propre : constructions et interprétations*. Duculot, Louvain-la-Neuve, 1994.

[Kit82] Richard KITTREDGE : Homogeneity and variation of sublanguages. *In* R. Kittredge & J. LEHRBERGER, éditeur : *Sublanguage : Studies of Language in Restricted Semantic Domains*, pages 107–137, Berlin, 1982. De Gruyter.

[Kit02] Richard KITTREDGE : Paraphrasing for condensation in journal abstracting. *Journal of Biomedical Informatics*, 35(4):265–277, 2002.

[Kle81] Georges KLEIBER : *Problèmes de référence : descriptions définies et noms propres*. Centre d'analyse syntaxique, Metz, 1981.

[Kri80] Saul A. KRIPKE : *Naming and necessity*. Harvard University Press, Cambridge, 1980.

[KTKL96] Leila KOSSEIM, Agnès TUTIN, Richard KITTREDGE et Guy LAPALME : Generating grammatical and lexical anaphora in assembly instructional texts. *In* Giovanni Adorni et MICHAEL ZOCK, éditeur : *Natural Language Generation, An Artificial Intelligence Perspective*, numéro 1036 de Lecture Notes in Artificial Intelligence : Trends, pages 260–275, Berlin, 1996. Springer.

[LL94] Shalom LAPPIN et Herbert J. LEASS : An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20(4):535–561, 1994.

[Mel93] Igor A. MEL'ČUK : *Cours de morphologie générale*, volume 1. Les presses de l'Université de Montréal, CNRS Éditions, Montréal, 1993.

[Mit98] Ruslan MITKOV : Robust pronoun resolution with limited knowledge. *In Proceedings of COLING-ACL*, pages 869–875, Montreal, 1998.

[MS99] Christopher D. MANNING et Hinrich SCHÜTZE : *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts, 1999.

[Pat93] Richard PATRY : L'analyse du discours de niveau discursif en linguistique : cohérence et cohésion. pages 109–143. Delachaux et Niestlé, Neuchâtel, 1993.

[PtMW90] Barbara PARTEE, Alice ter MEULEN et Robert E. WALL : *Mathematical Methods in Linguistics*. Kluwer Academic Publishers, Dordrecht, Netherlands, 1990.

[Sal99] Morris SALKOFF : *A French-English grammar : a contrastive grammar on translational principles*. John Bebjamins Publishing Compagny, Amsterdam/Philadelphia, 1999.

[Sid03] Advaith SIDDHARTHAN : Resolving pronouns robustly : Plumbing the depths of shallowness. *In Proceedings of the Workshop on Computational Treatments of Anaphora, 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2003)*, pages 7–14, Budapest, Hungary, 2003.

[Tel95] Christine TELLIER : *Éléments de syntaxe du français*. Presses de l'Université de Montréal, Montréal, 1995.

[Tro01] François TROUILLEUX : *Identification des reprises et interprétation automatique des expressions pronominales dans des textes en français*. Thèse de doctorat, Université Blaise Pascal, France, 2001.

[Tut92] Agnès TUTIN : *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, 1992.

# Sources des exemples


[1] Robert A. ADAMS (traduction de Pierrette MAYER) : *Calcul différentiel & intégral dans l'espace*. Addison-Wesley, Montréal, page 54, 1988.

[2] Description d'un épisode de l'émission de télévision *Beverly Hills* au réseau de télévision TVA. Disponible à `http://tva.canoe.com`.

[3] Madeleine GUÉNETTE et Lise GUÉNETTE : *La santé ... c'est bon !*, tome 1. Lima, Québec, page 145.

[4] PEYO : La Schtroumpfette. *Les Schtroumpfs*, volume 3. Jean Dupuis, Marcinelle-Charleroi, page 40, 1976.

[5] Claude TURCOTTE : L'Abitibi-Consolidated installera son siège social dans l'édifice Sun Life. *In La Presse Affaires*, le 20 novembre 2002.

[6] Marc-André LUSSIER : Un coup pas très fin ... *In Cyberpresse (La Presse)*, le 3 mai 2003.

[7] Luc PERREAULT : La Turbulence des fluides. *In Cyberpresse (La Presse)*, le 7 septembre 2002.

[8] Chantal GUY : Portrait du poëte en play-boy. *In Cyberpresse (La Presse)*, le 1 mars 2003.

[9] Douglas R. HOFSTADTER : *Gödel, Escher, Bach : an Eternal Golden Braid*, Basic Books, New York, éd. 20ème anniversaire, page 127, 1999.

[10] Chantal GUY : La beauté sans surprises. *In Cyberpresse (La Presse)*, le 10 mai 2003.

[11] L'expansion a desservi ABB. *In La Presse Affaires*, page D9, le 20 novembre 2002.

[12] Aleksi K. LEPAGE : Chris Rock, président des États-Unis. On peut toujours rêver. *In Cyberpresse (La Presse)*, le 29 mars 2003.

[13] David S. LAWYER (traduit par Olivier THARAN) : *Terminal Texte pour Linux*, version 1.06, juin 1999. Disponible à `http://www.freenix.org/unix/linux/HOWTO/Text-Terminal-HOWTO.html`.

[14] Superior Propane restera séparée, pour le moment, d'IGG PC Calgary. *In Le Soleil Questions d'argent*, le 12 décembre 1998.

[15] Martin HAMILTON (traduit par Laurent MARTIN) : *RedHat Linux KickStart HOWTO*, version 0.1, 28 septembre 1998. Disponible à `http://www.freenix.org/unix/linux/HOWTO/KickStart-HOWTO.html`.

[16] Luc PERREAULT : Invasions barbares. *In Cyberpresse (La Presse)*, le 10 mai 2003.

[17] Paul HAIGH (adaptation française par Stéphane LEE CHIP HING) : *Oracle Database HOWTO*, version 1.2, 4 août 1998. Disponible à `http://www.freenix.org/unix/linux/HOWTO/Oracle-HOWTO.html`.