Une première approche de l'utilisation des chaînes coréférentielles pour la détection des variantes anaphoriques de termes

Sarah Trichet-Allaire

Université de Nantes, LINA, CNRS
LI - Université de Tours

Résumé

Cet article traite de l'utilité à détecter une chaîne coréférentielle de termes complexes afin d'améliorer la détection de variations de ce même terme complexe. Nous implémentons pour cela un programme permettant de détecter le nombre de variantes anaphoriques d'un terme complexe ainsi que le nombre de variantes anaphoriques de termes dans un texte scientifique. Ces deux fonctionnalités sont développées avec une ancrage dans une chaîne coréférentielle et en dehors de toute chaîne coréférentielle, afin de pouvoir évaluer l'efficacité de cette méthode.

Mots-clés : chaîne coréférentielle, détection automatique, variante de terme, anaphore nominale.

Abstract

This article discusses the usefulness to detect a coreferential chain of complex terms in order to improve detection of variations of this same complex term. We implement a program for this purpose, allowing the detection of the number of anaphoric variants of a complex term, as well as the number of anaphoric variants of terms in scientific texts. These two functions are developed either with ou without the anchoring in a coreferential chain.

Keywords: coreferential chain, automatical detection, term variation, nominal anaphora.

1. Introduction

L'école de Vienne définit un terme comme une étiquette unique sur un concept défini a priori (l'approche onomasiologique, qui part du concept pour aller vers le signe). Les informaticiens linguistes ont cependant mis l'accent sur l'importance de la variation terminologique dans un texte, et donc sur la nécessité de partir d'une classe de termes pour aller vers un concept (approche sémasiologique, qui part du signe pour aller vers le concept). La détection des variantes anaphoriques de termes correspond alors à la détection d'une classe de termes désignant un concept unique.

La détection des anaphores nominales est un sujet qui peut trouver des applications autant en recherche d'information (RI) que dans les systèmes de questions-réponses (SQR) ou bien en traduction et résumé automatiques. La construction d'ontologies basées sur les variantes nominales peut en effet permettre de trouver des termes synonymes en RI, utiliser des mots-clés plus diversifiés en SQR, éviter les répétitions en traduction et résumé automatiques, voire, dans ce domaine, trouver des termes plus pertinents.

La détection de termes anaphoriques, plus précise que la détection de variantes nominales, est la

recherche des termes désignant un concept commun dans un contexte particulier. Ainsi, le terme « outil » peut faire référence à « moteur de recherche » dans un contexte tel que les textes scientifiques, mais sera plus difficilement considéré comme une variante nominale hors contexte.

La recherche sur l'anaphore, et plus globalement sur la coréférence, est un domaine étudié depuis une trentaine d'années. Les conférences MUC (Message Understanding Conference) 6 et 7 ont souligné l'importance de la détection de la coréférence, et ont entraîné ainsi une recrudescence de la recherche depuis une dizaine d'années.

Cependant, si beaucoup de travaux ont été effectués sur la résolution de l'anaphore pronominale (Hobbs, 1978 ; Mitkov, 2002), la résolution ainsi que la détection de l'anaphore nominale ont vu peu de résultats concrets. Les principaux articles les étudiant se situent plutôt dans le domaine de la linguistique et permettent de donner un cadre d'application, mais qui reste encore vague et sans application concrète. Seuls (Gelbukh et Sidorov, 1999) proposent une méthode pour rechercher des variantes nominales : l'élaboration d'un lexique permettant la résolution des anaphores indirectes.

Les recherches en linguistique proposent principalement de se baser sur la saillance, phénomène cognitif qui désigne le terme mis en avant dans une partie d'un discours. Ainsi, pour la résolution de l'anaphore pronominale, le terme le plus saillant d'une partie du discours (proposition, phrase ou paragraphe) est le candidat le plus pertinent pour un pronom non résolu.

L'utilisation du concept de saillance permet de diviser le problème de la résolution de la référence en deux étapes. La première étape consiste en la détection d'un terme, l'anaphore, qui fait référence à un concept déjà évoqué dans le texte. La deuxième étape, une fois l'anaphore détectée, est de trouver l'antécédent, qui est le terme référent de l'anaphore. Cet antécédent est le terme le plus saillant de la partie du discours où apparaît l'anaphore.

Une analyse plus fine par le discours est proposée par (Jacques, 2003) afin de détecter la réduction de terme, qui correspond à une part importante des variantes anaphoriques nominales. Cette approche discursive suggère l'utilisation de marques du discours, telles que la typographie et la disposition des paragraphes afin d'élaborer des macro et micro structures dans le texte. Cette structuration permettrait ainsi de mettre en avant les éléments saillants du discours.

Le but de ce travail est d'exploiter le phénomène de la saillance afin de détecter les variantes anaphoriques dans les textes scientifiques. En effet, la réduction d'un travail de recherche à un genre particulier (textes scientifiques, journalistiques, littéraires) permet de cerner des aspects discursifs spécifiques des textes, comme ici la présence de mots-clés et de résumés.

2. Cadre de travail

Tous les exemples suivis d'un astérisque sont issus du corpus TAL(N).

Nous étudierons plus spécifiquement dans cet article les anaphores mettant en jeu les termes complexes. Les variantes anaphoriques de termes peuvent être classées en deux catégories : les anaphores fidèles et infidèles.

La première catégorie, les anaphores fidèles, concerne des termes simples (exemple 1 (Kraif, 2000)*) ou complexes (exemple 2 (D'Alessandro, 2001)*) reprenant la tête d'un terme complexe.

1 « Les bi-textes sont *des corpus bilingues parallèles*$_1$, généralement segmentés et alignés au niveau des phrases. Une des applications les plus directes de *ces corpus*$_1$ consiste à en extraire

automatiquement des correspondances lexicales. »

2 « Au niveau international, c'est également en 1968 que sont publiés les travaux sur *les deux premiers systèmes automatiques de synthèse*$_1$ pour l'anglais : *le système américain par diphone*$_1$ de Dixon et Maxey, démontré en 67 et publié en 1968, et *le système par règles japonais*$_1$ de Umeda, Matsui, Teranishi et Suzuki, publié aussi en 1968. »

Pour cette catégorie, le repérage de la tête du terme complexe donne un indice sur la détection de reprises anaphoriques, mais ce n'est pas non plus systématiquement fiable.

La deuxième catégorie, les anaphores infidèles, correspond à des reprises dont les têtes sont différentes de celles de leurs antécédents. L'extension peut être identique (exemple 3 (D'Alessandro, 2001)*) ou pas (exemple 4 (Tanguy et Hathout, 2002)*).

3 « Le vocodeur à canaux a été remplacé par *une technique de synthèse*$_1$ de meilleure qualité : la prédiction linéaire. *Le système de synthèse*$_1$ ainsi réalisé a donné lieu à des développements commerciaux importants, sous forme de logiciels et de cartes informatiques spécialisées. »

4 « *Webaffix*$_1$ (1) utilise un moteur de recherche pour collecter des formes candidates qui contiennent un suffixes graphémique donné, (2) prédit les bases potentielles de ces candidats et (3) recherche sur le Web des co-occurrences des candidats et de leurs bases prédites. *L'outil*$_1$ a été utilisé pour enrichir Verbaction, un lexique de liens entre verbes et noms d'action ou d'événement correspondants. »

La difficulté est ici de bien repérer l'anaphore. La simple détection de l'extension dans les cas similaires à l'exemple 3 est un indice de reprise mais que nous n'utiliserons pas ici car il peut engendrer un nombre important de faux positifs. Les cas similaires à l'exemple 4 sont les plus complexes, car ils doivent se passer entièrement de la détection des composantes du terme complexe.

Un autre système de filtrage est alors nécessaire afin de détecter ce type de variantes.

3. Méthode

Tous les exemples suivis d'un astérisque sont issus du corpus TAL(N).

Notre méthode de détection de variantes anaphoriques se déroule en deux étapes. La première étape consiste à repérer une chaîne coréférentielle de termes complexes afin que la détection soit ancrée sur une partie plus restreinte du texte. Cette partie plus restreinte, de la première référence jusqu'à la dernière référence plus 2 phrases, est alors extraite afin d'être analysée plus finement lors de la deuxième étape où sont recherchées les variantes anaphoriques.

La détection d'une chaîne coréférentielle permet de centrer la recherche de variantes anaphoriques sur des passages plus saillants du texte (Jacques, 2003). Cette méthode a déjà été utilisée par (Dupont, 2002) dans le modèle des attentes du lecteur afin de résoudre les anaphores pronominales.

La définition la plus simple d'une chaîne coréférentielle est une suite de plusieurs références dans un texte. En théorie, la distance séparant deux références n'a de limite que le texte, surtout lorsque la référence désigne une entité nommée. Cependant, nous choisirons une définition plus restrictive de la chaîne coréférencielle. En effet, la plupart des références peuvent être trouvées dans une fenêtre de 2 phrases (Mitkov, 2002). Nous reprenons cette donnée concernant les anaphores pronominales pour les anaphores nominales, en l'augmentant d'une phrase. Cette augmentation a été testée de manière expérimentale afin d'augmenter le nombre de chaînes

coréférentielles détectées.

Nous considérerons donc comme chaîne coréférentielle une suite de trois termes complexes identiques, la variation morphologique mise à part : un antécédent suivi d'au moins deux anaphores, chacun de ces termes étant distants de 2 phrases maximum l'un de l'autre (exemple 1 (D'Alessandro, 2001)*).

1 « Les *systèmes* qui ont existé mais dont il n'a pas été possible de retrouver d'exemples sonores, sont les suivants : *le système de synthèse* par règles de l'ICP, décrit dans [BAI88] (*le système actuel* de ce laboratoire figure ici), *les premiers systèmes par diphones* de la société Electrel (dont figure ici *le système KALI*), *le système Mockingboard* commercialisé dans les années 80 par BIP, qui était relié à un ordinateur Apple II, *le prototype de synthèse par diphones* développé par Telic-Alcatel dans les années 80 autour d'un microprocesseur Nec7720, *le système par diphones* ICO85 de Vecsys.[...] La troisième partie du disque contient des exemples de *systèmes de synthèse* contemporains, qui ne correspondent pas à un article du volume. Ici encore, l'auteur a sollicité des exemples de tous les *systèmes* actuels parlant français dont il a pu avoir connaissance, en ce début du 21ème siècle. »

Une fois la chaîne coréférentielle extraite, nous recherchons les variantes anaphoriques du terme complexe référencé. Dans une première étape, la tête syntaxique du terme complexe est recherchée entre chaque occurrence du terme complexe. Si elle est détectée au sein de la chaîne coréférentielle, elle est alors considérée comme une variante anaphorique fidèle. Ensuite, si aucune variante fidèle n'est détectée, nous recherchons les noms ayant un déterminant démonstratif. Si un tel déterminant est détecté, le nom est considéré comme une variante infidèle. Si aucune variante n'est détectée à cette étape, nous recherchons un nom avec un déterminant défini, puis un déterminant indéfini. Le processus est ensuite réitéré entre chaque occurrence du terme complexe à partir de la recherche de la tête syntaxique.

La détection de la tête syntaxique au sein d'une chaîne coréférentielle permet de vérifier si cette même tête peut être une reprise anaphorique. Nous ne cherchons pas pour l'instant à résoudre l'anaphore, mais seulement à la détecter. Nous n'avons donc pas à nous pencher sur le problème de références croisées comme dans l'exemple 2 donné dans la section 2. Ainsi, la reprise de la tête syntaxique au sein d'une chaîne coréférentielle est systématiquement analysée comme une reprise anaphorique.

De plus, la détection de déterminants est un indice simple nous permettant de vérifier si l'analyse des chaînes coréférentielles est pertinente afin de repérer des variantes anaphoriques. L'utilisation d'autres indices tels que le genre ou le nombre mettrait de côté un certain nombre de variantes anaphoriques infidèles, dont les racines peuvent être différentes de l'antécédent. Ainsi, nous préférons gardons un seul indice que sont les déterminants, plutôt que d'avoir de multiples indices qui pourraient nous empêcher de détecter les variantes anaphoriques correctes. L'influence de cet indice assez simple est donc analysé dans un premier temps, sachant que cette analyse sera affinée à l'avenir.

Le choix du déterminant comme seul indice est piloté par le fait qu'un déterminant démonstratif est un indice de saillance du nom, tandis qu'un déterminant défini sera un indice de moindre saillance : une graduation est ainsi mise en place entre les candidats termes.

Nous allons à présent dérouler l'algorithme 1 sur l'exemple 2 (D'Alessandro, 2001)* :

2 « 33 ans de *synthèse de la parole*$_1$ à partir du texte : une promenade sonore (1968-2001). *Cet article*$_2$ présente un disque compact de 69 exemples sonores de *synthèse de parole*$_1$.

Algorithm 1 détection de variantes anaphoriques de termes complexes

```
while ∃ une chaine coreferentielle do
  Entre chaque terme complexe t
  if ∃ reprise r de la tete syntaxique then
    variations[t] = r ;
  else if ∃ determinant demonstratif pour un nom n1 then
    variations[t] = n1 ;
  else if ∃ determinant defini pour un nom n2 then
    variation[t] = n2 ;
  end if
  fin entre
end while
```

> Des exemples de 25 systèmes de *synthèse automatique*$_3$ à partir du texte, principalement en français, sont décrits, avec 54 voix différentes. [...] Ensuite, des parcours d'écoute sont proposés au lecteur/auditeur, afin d'apprécier divers aspects de la *synthèse de parole*$_1$: par types de synthétiseurs, sur le calcul de la prosodie, par type d'unités de *synthèse*$_4$ ».

L'algorithme détecte tout d'abord la chaîne coréférentielle formée par les trois occurrences du terme complexe *synthèse de la parole.* Ensuite, le terme *synthèse* est recherché entre les deux premières occurrences du terme complexe. N'étant pas détecté, un déterminant démonstratif est recherché. L'algorithme détecte alors *cet article* en tant que variante anaphorique (ce qui est faux). La recherche continue ensuite entre la deuxième et la troisième occurrence. Le terme *synthèse* est détecté, et le terme complexe *synthèse automatique* est alors enregistré en temps que variante fidèle. La recherche continue jusqu'à 2 phrases après le dernier terme complexe. Le terme *synthèse* est également répertorié en tant que variante anaphorique.

4. Analyse et évaluation

Le travail d'analyse et d'évaluation se fait entièrement à partir de données réelles à l'aide d'un corpus spécialisé autour de la communauté TAL. Le corpus étudié, appelé corpus TAL(N) (Dubreil, 2004), rassemble un million d'occurences, sans compter la ponctuation, ce qui est la taille optimale pour un corpus spécialisé. Il est constitué d'un ensemble de 193 articles scientifiques écrits en français, rassemblés de manière équilibrée entre articles de revue et articles de conférence à des vue d'analyse sémantico-discursive. Le corpus est ensuite traité avec ACABIT, extracteur de termes complexes (Daille, 2003). La détection des termes complexes se fait à l'aide de la sortie d'ACABIT, un regroupement des termes complexes qui sont ensuite indexés. Ces données sont ensuite couplées avec la sortie du lemmatiseur FLEMM afin de reconnaître les noms et les types des déterminants.

4.1. Données d'évaluation

L'évaluation est effectuée à partir d'une première analyse manuelle du corpus TAL(N) traité par ACABIT. Sept termes complexes sont analysés à travers différentes caractéristiques : l'antécédent, la structure morpho-syntaxique de la référence, le type de préposition, le nombre de phrases séparant l'anaphore et l'antécédent, le type du déterminant, le respect de l'accord en genre et en nombre et l'appartenance à une chaîne coréférentielle.

Ces sept termes sont : *système de synthèse, analyse syntaxique, moteur de recherche, corpus bilingue, entité nommée, modèle statistique et modèle de langue.*

Parmis les 87 références détectées, 45 % font partie d'une chaîne coréférentielle. Si l'ensemble des références est donc loin d'être couverte par les chaînes coréférentielles, celles-ci concernent cependant un nombre non négligeable d'anaphores.

4.2. Expérimentations

Deux versions de l'algorithme sont expérimentées : une première version détecte les variantes au sein des chaînes coréférentielles en suivant l'algorithme donné dans la section 3. La deuxième version se base seulement sur les critères syntaxiques (reprise de la tête et type du déterminant) afin de vérifier la pertinence de la détection des chaînes coréférentielles pour trouver les variantes anaphoriques de termes.

5. Résultats

Tous les exemples suivis d'un astérisque sont issus du corpus TAL(N).

Deux types d'évaluation sont effectuées : une première évaluation est effectuée à partir d'un texte dans son ensemble. Dans une deuxième évaluation, un unique terme complexe est étudié à travers l'ensemble des textes du corpus, de la même manière que l'évaluation manuelle.

5.1. Analyse sur un texte

La première méthode d'évaluation se concentre sur un texte de (Meunier, 1999)* et analyse toutes les références trouvées, et pas seulement les sept termes complexes de l'analyse manuelle. L'expérimentation est réalisée dans une première phase avec la prise en compte des chaînes coréférentielles, puis hors chaîne coréférentielle dans une deuxième étape afin de vérifier la pertinence de ce critère.

5.1.1. Avec les chaînes coréférentielles

Lors de l'expérimentation faite dans le cadre de chaînes coréférentielles, onze chaînes coréférentielles sont détectées, dont plusieurs se chevauchent, c'est-à-dire que pour une chaîne avec les termes T1 T2 et T3, la chaîne suivante pourra comporter les termes T2 et T3.

Le terme « entrée lexicale » est détecté à plusieurs reprises, ainsi que sa variante fidèle « entrée ».

Sur les 41 variantes anaphoriques détectées, 13 correspondent à des variantes fidèles, et toutes sont correctes. Deux variantes anaphoriques infidèles sont également détectées : les termes « lemme » (exemple 1) et « famille », qui font référence aux entités désignées par les entrées lexicales. Nous pouvons donc considérer « entrée », « lemme » et « famille » comme des variantes anaphoriques infidèles, dans la mesure où la même entité est désignée (ce qui compose les entrées du tableau).

1 « Ainsi, il n'est pas possible d'identifier une *entrée lexicale* par le *lemme* qu'elle utilise. En effet, prenons l'exemple du *verbe* voler [...]. Nous ne pouvons donc identifier ces deux *entrées lexicales* par le verbe voler. »

Avec 20 variantes correctes détectées pour 41 variantes au total, nous arrivons avec une précision de 48,8 %, ce qui est un score relativement faible, si nous gardons en tête que la méthode que nous proposons utilise de faibles contraintes syntaxiques.

5.1.2. Hors chaîne coréférentielle

L'analyse des termes complexes du texte (Meunier, 1999) donne un résultat de 164 variantes détectées. Le taux de précision est de 31,7 %, soit un résultat nettement inférieur aux résultats trouvés à l'aide des chaînes coréférentielles. Il faut noter cependant que le texte de (Meunier, 1999) est particulièrement riche en termes complexes comme « entrée lexicale », et il se trouve peu d'ambiguïtés désignant des entrées lexicales de nature différente, et pouvant créer des erreurs au niveau des variantes anaphoriques fidèles.

Nous voyons donc que la détection de termes complexes au sein des chaînes coréférentielles donne un résultat au pire équivalent en précision à la détection des variantes sans chaînes coréférentielles.

5.2. Analyse sur un terme complexe : « analyse syntaxique »

La deuxième méthode d'évaluation consiste à étudier un terme particulier (ici, « analyse syntaxique ») à travers l'ensemble du corpus d'étude.

Lors de l'évaluation manuelle sur le corpus TAL(N), 24 références ont été détectées, que ce soit hors ou dans une chaîne coréférentielle. Les anaphores infidèles détectées lors de l'analyse manuelle sont multiples (« première partie du traitement », « opération de composition syntaxique », « étape », « filtrage »,« résultat », etc.). Douze de ces termes appartiennent à une chaîne coréférentielle, répartis dans trois chaîne coréférentielles distinctes.

Nous pouvons comparer ce résultat avec celui du premier type d'expérimentation réalisé avec la détection de chaînes coréférentielles afin d'effectuer un calcul de rappel.

5.2.1. Avec les chaînes coréférentielles

L'évaluation automatique recense seulement quatre variantes anaphoriques différentes du terme « analyse syntaxique ». Aucune des variantes infidèles ne fait partie de l'évaluation manuelle, tandis que neuf variantes fidèles ont été détectées à la fois par l'évaluation manuelle et l'évaluation automatique.

Dans les variantes fidèles, par contre, le terme « analyse » est détecté correctement à dix reprises. En revanche, le terme « analyse sémantique » est incorrectement détecté comme étant une variante anaphorique d'« analyse syntaxique » (exemple 1 (Biskri et Delisle, 1999)*). De plus, le terme « analyse » dans ce contexte fait référence à « analyse sémantique » (exemple 2 (Biskri et Delisle, 1999)*), ce qui est également faux.

1 « L'*analyse syntaxique*$_1$ nous permet d'accéder aux constituants syntaxiques dont nous avons besoin pour l'*analyse sémantique*$_2$. »
2 « La partie centrale de l'*analyse sémantique*$_1$ est l'*analyse*$_2$ casuelle semiautomatique et interactive de HAIKU qui utilise un système de cas général, indépendant de tout domaine particulier (Barker et al. 1997). »

Dans les variantes infidèles, les termes détectés sont « système » (exemple 3 (Biskri et Delisle, 1999)*), « complément » (exemple 4 (Fabre et Frérot, 2002)*) et « OCR » (exemple 5 (Aloulou *et al.*, 2000)*).

3 « En l'absence de connaissances sémantiques a priori, une *analyse syntaxique*$_1$ détaillée et indépendante du domaine est le seul guide fiable vers le sens des énoncés d'un corpus. La

partie centrale de l'analyse sémantique est l'analyse casuelle semiautomatique et interactive de HAIKU qui utilise un *système*$_2$ de cas général, indépendant de tout domaine particulier. »

4 « Notre cadre de travail est la construction d'un outil d'*analyse syntaxique*$_1$ de corpus, Syntex (Bourigault, Fabre, 2000). Dans ce contexte, nous nous intéressons ici à la question du rattachement prépositionnel, et cherchons à déterminer une stratégie qui permette d'aller au-delà de la décision simple de rattachement (à quel recteur se rattache une préposition) pour rendre possible une typologie des *compléments*$_2$. »

5 « Dans cet article, nous présentons une approche originale d'*analyse syntaxique*$_1$ robuste appliquée à l'arabe et basée sur l'architecture multiagent. Comme première application de notre approche, notre système sera couplé avec un système de reconnaissance de l'écriture arabe dans le but d'effectuer, d'une part, la validation linguistique des mots reconnus par l'*OCR*$_2$ (Optical Character Recognition) et d'autre part [...] »

Ces résultats encourageants sont essentiellement dus aux reprises simples de la tête syntaxique, tandis que les variantes infidèles sont toutes incorrectes dans cette expérience. Sur 15 références détectées, 10 sont correctement détectées, soit un taux de précision de 67 %. De plus, sur les 24 références détectées lors de l'évaluation, 9 ont été correctement repérées, soit un taux de rappel de 38 %. La f-mesure (2*Rappel*Précision/(Rappel+Précision)) est alors de 48 %. Cette mesure faible s'explique par le taux de rappel très bas. Le choix était en effet de favoriser la précision au détriment du rappel, ce dernier ne pouvant être élevé, puisque les termes ne faisant pas partie d'une chaîne coréférentielle sont automatiquement exclus.

5.2.2. Hors chaîne coréférentielle

Lors de l'analyse effectuée sans prendre en compte les chaînes coréférentielles, 180 variantes sont détectées. 45 variantes sont détectées correctement et, parmis celles-ci, 8 correspondent à des variantes infidèles. Les variantes fidèles sont des termes simples comme « analyse » ou des termes complexes comme « arbre d'analyse » (exemple 1 (Chappelier et Rajman, 2001)*). Les variantes infidèles sont diverses, et ne peuvent souvent être considérées comme variantes anaphoriques que dans un contexte particulier. Les huit termes détectés sont : « endroit » (exemple 2 (Mertens, 1999)*), « conversion », « troncature », « validation », « ancrage », « solution », « structure » et « processus » (exemple 3 (Perrier, 2002)*).

1 « [...] l'approche Data-Oriented Parsing (DOP) pour l'analyse syntaxique probabiliste a depuis été largement étudiée par diverses équipes de recherche. La principale limitation de cette approche reste cependant le caractère NP-difficile du problème consistant à trouver l'arbre d'analyse le plus probable. »

2 « L'algorithme est intégré dans un système de synthèse de la parole, comportant plusieurs modules (lemmatisation, *analyse syntaxique*, phonétisation, syllabation, génération de l'intonation, et module acoustique), ce qui complique l'évaluation. En effet, telle erreur dans la sortie peut provenir de plusieurs *endroits*. »

3 « Dans cette approche, l'*analyse syntaxique* consiste à chercher des modèles de descriptions d'arbres sous forme d'arbres syntaxiques complètement spécifiés. Dans toute sa généralité, ce *processus* est hautement indéterministe. »

Parmis les 180 variantes détectées, 45 le sont correctement, ce qui correspond à un taux de précision de 25 %, variantes fidèles et infidèles confondues. Cette proportion doit être comparée avec le taux de précision obtenu à l'aide des chaînes coréférentielles : 67 %. Il est donc clair que l'utilisation des chaînes coréférentielles augmente le taux de précision pour la détection des

	Avec chaînes coréférentielles	Sans chaînes coréférentielles
Taux de précision	67%	25%
Taux de rappel	38%	48%
F-Mesure	48%	33%

Figure 1. Résultats de la détection automatique des variantes anaphoriques de termes

anaphores nominales. Cependant, le taux de rappel est de 48 % (11 variantes correctement détectées sur 24 à détecter d'après l'analyse manuelle). La f-mesure pour la détection des variantes anaphoriques de termes est donc de 33 %, soit un taux inférieur de 15 % à celui trouvé pour la détection effectuée dans le cadre des chaînes coréférentielles.

Ces résultats sont résumés dans la figure 1.

6. Conclusion

Nous avons tenté de démontrer dans cet article que la détection de variantes anaphoriques au sein d'une chaîne coréférentielle est plus pertinente que cette même détection en dehors de toute chaîne coréférentielle. Le nombre de chaîne coréférentielles détectées est peu élevé, et ne permet malheureusement pas de donner des résultats fiables, bien que l'analyse d'un terme particulier démontre une meilleure précision avec l'aide d'une chaîne coréférentielle que sans.

Cependant, la recherche des chaînes coréférentielles à partir de trois termes complexes, comme nous le faisons ici, est un processus très limitatif. Le nombre de chaînes coréférentielles détectées est en effet trop restreint pour avoir une analyse fiable. La détection d'une chaîne coréférentielle pourrait donc se limiter à deux termes complexes proches afin d'en augmenter le nombre.

Les termes simples sont également mieux détectés lorsque la recherche est limitée au voisinage d'une chaîne coréférentielle de ce même terme doté d'un complément.

Références

ALOULOU C., BELGUITH L. H. et HAMADOU A. B. (2000). « Vers un système d'analyse syntaxique robuste pour l'Arabe : Application au recouvrement des erreurs de la reconnaissance ». In *Actes de TALN 2000*. Lausanne, France.

BISKRI I. et DELISLE S. (1999). « Un modèle hybride pour le textual data mining : un mariage de raison entre le numérique et le linguistique ». In *Actes de TALN 1999*. Cargèse, France.

CHAPPELIER J.-C. et RAJMAN M. (2001). « Grammaire à substitution d'arbre de complexité polynomiale : un cadre efficace pour DOP ». In *Actes de TALN 2001*. p. 133–142.

DAILLE B. (2003). « Conceptual structuring through term variations ». In *Proceedings ACL 2003 Workshop on Multiword Expressions : Analysis, Acquisition and Treatment*. p. 9-16.

D'ALESSANDRO C. (2001). « 33 ans de synthèse de la parole á partir du texte : une promenade sonore (1968-2001) ». In *T.A.L. : Traitement automatique de la langue*, 42 (1).

DUBREIL E. (2004). « Pourquoi et comment constituer un corpus électronique spécialisé ? Le cas du corpus TAL(N) ». In *Coldoc05 - 2ème Colloque Jeunes Chercheurs en Sciences du Langage*. Nanterre, France.

DUPONT M. (2002). « Une approche cognitive pour le calcul des chaînes de références ». In *Actes de TALN 2002*. Nancy, France.

FABRE C. et FRÉROT C. (2002). « Groupes prépositionnels arguments ou circonstants : vers

un repérage automatique en corpus ». In *Actes de TALN 2002*. Nancy, France.

GELBUKH A. et SIDOROV G. (1999). « A dictionary-based algorithm for indirect anaphora resolution ». In *Proceedings of Vextal*. Venice, Italia.

HOBBS J. R. (1978). « Resolving pronoun references ». In *Lingua*, 44, 311–338.

JACQUES M.-P. (2003). *Approche en discours de la réduction des termes complexes dans les textes spécialisés*. Thèse de doctorat en sciences du langage, Université de Toulouse II.

KRAIF O. (2000). « Extraction automatique de correspondances lexicales évaluation d'indices et d'algorithmes ». In *Actes de TALN 2000*. Lausanne, France, p. 225–236.

MERTENS P. (1999). « Un algorithme pour la génération de l'intonation dans la parole de synthèse ». In *Actes de TALN 1999*. Cargèse, France, p. 233–242.

MEUNIER F. (1999). « Modélisation des ressources linguistiques d'une application industrielle ». In *Actes de TALN 1999*. Cargèse, France.

MITKOV R. (2002). *Anaphora resolution*. Longman, London.

PERRIER G. (2002). « Descriptions d'arbres avec polarités : les Grammaires d'Interaction ». In *Actes de TALN 2002*. Nancy, France.

TANGUY L. et HATHOUT N. (2002). « Webaffix : un outil d'acquisition morphologique dérivationnelle à partir du Web ». In *Actes de TALN 2002*. Nancy, France.