UNIVERSITÉ DU QUÉBEC À MONTRÉAL

OPTIMISATION DU TRAITEMENT AUTOMATIQUE DES RELATIONS ANAPHORIQUES PRONOMINALES PAR L'UTILISATION DES RELATIONS D'ASYMÉTRIE

THÈSE

PRÉSENTÉE

COMME EXIGENCE PARTIELLE DU

DOCTORAT EN INFORMATIQUE COGNITIVE

PAR

CALIN BATORI

JUIN 2015

REMERCIEMENTS

Je tiens à remercier Mme Anna Maria Di Sciullo d'avoir accepté de diriger cette thèse. Du moment où elle m'a accueilli au sein du Laboratoire de recherche sur les asymétries d'interface elle a toujours su m'apporter ses conseils et guider mes questionnements de manière à rendre ma recherche plus fructueuse. Merci pour votre soutien indéfectible et pour votre patience.

Je remercie également M. Ghislain Lévesque, le codirecteur de cette thèse, de m'avoir constamment assuré que j'étais sur le bon chemin et de m'avoir encouragé de poursuivre.

Je tiens également à remercier les membres du jury pour leurs commentaires et suggestions.

Je remercie de plus mes collègues du Laboratoire de recherche sur les asymétries d'interface et du Programme de doctorat en informatique cognitive. Leur implication altruiste et leurs suggestions m'ont grandement aidé à préciser ma position.

Pour terminer, je remercie de tout mon cœur ma famille de m'avoir encouragé tout le long de mon cheminement doctoral. Je leur dédie cette thèse.

TABLE DE MATIÈRES

LISTE DES FIGURES

LISTE DES TABLEAUX

LISTE DES ABRÉVIATIONS

aLAD	module de résolution d'anaphores pronominales intégré à l'analyseur LAD
AU	Analyseur universel
CALN	Compréhension automatique du langage naturel
CP	Projection maximale du complémenteur
DD-Linking	Liage dans le domaine du discours
DP/NP/VP/PP	Projection maximale d'un déterminant/nom/verbe/préposition
EAPOS	Ensemble d'antécédents possibles
EAPR	Ensemble d'anaphores pronominales à résoudre
F	Fenêtre d'attention utilisée par aLAD
FDG	Functional Dependency Grammar
GU	Grammaire universelle
HHP	hypothèse de la hiérarchie des projections homogènes
IA	Intelligence Artificielle
IUA	Interprétation Sous Asymétrie
LAD	Analyseur syntaxique développé au *L*aboratoire de recherche sur les *A*symétries *D*'interfaces
LL(1)	Analyseur qui procède de manière descendante et produit une dérivation à gauche avec 1 mot d'avance

MARS	Mitkov's Anaphora Resolution System - algorithme de Mitkov 2002
MUC	Message Understanding Conference
NLP	Natural Language Processing
POS	Part of Speech Tagger
RAP	Resolution of Anaphora Procedure - algorithme de Lappin et Leass 1994
TALN	Traitement automatique du langage naturel
TP	Projection maximale du temps

LISTE DE TERMES CLÉS

Analyse avec des connaissances linguistiques superficielle versus profondes : Selon la quantité d'information linguistique utilisée pendant l'analyse, les analyseurs peuvent être *profonds* (deep parsers) ou *superficiels* (shallow parsers). Un analyseur superficiel est rapide et robuste et utilise souvent des heuristiques et techniques probabiliste pour obtenir une structure simple des relations existantes entre les termes analysés. Un analyseur profond utilise une description grammaticale détaillée de la langue traitée pour obtenir une description riche des relations syntactico-sémantiques présentes dans le texte analysé. Autres termes utilisés : approches pauvres versus riches en connaissances linguistiques.

Anaphore : Un procédé de reprise de l'information antérieurement énoncée dans lequel un élément qui n'est pas autonome du point de vue référentiel acquière sa référence en établissant une relation, dite anaphorique, avec son antécédent.

Anaphore pronominale : Une anaphore qui est exprimée par un pronom

Antécédent : Élément indépendant du point de vue référentiel qui contribue à l'interprétation de l'anaphore

Cataphore (backwards anaphora) : Une relation *anaphorique* dans laquelle l'ordre des éléments est inversé – l'élément qui n'est pas indépendant du point de vue référentiel devance celui qui contribue son interprétation référentielle.

Coréférence : Relation qui s'établit entre deux ou plusieurs termes qui ont le même référent et qui peuvent apparaitre dans la même proposition ou dans des propositions différentes. Dans ce cas, les termes sont co-indexés et ont les mêmes indices référentiels.

Liage : Spécifie les conditions de bonne formation des relations de dépendance référentielle entre les anaphores et leurs antécédents dans un domaine restreint à la proposition.

Grammaire universelle : Un ensemble d'opérations et de principes innés qui permettent aux humains d'apprendre une langue et de générer des phrases potentiellement infinies à partir d'un nombre fini d'éléments.

Analyseur universel : Un analyseur postulé dans la ligne de recherche de la grammaire générative et qui analyse les expressions linguistiques en interprètent des grammaires qui instancient les opérations et les principes de la grammaire universelle. (voir Chapitre 2)

Représentation : Dans les sciences cognitives un objet de la cognition est *re-crée* par un processus de *re-présentation* à l'intérieur de l'esprit connaisseur. La représentation est la *re-présence* dans l'esprit de l'objet connu.

Résolution des anaphores (anaphora resolution) : Est la sélection de l'antécédent d'une anaphore suite à la résolution d'éventuelles ambiguïtés référentielles. La résolution de l'anaphore pronominale consiste à trouver ce à quoi réfère l'anaphore réalisée par un pronom, soit son antécédent.

Structure d'arguments de la phrase : Description des propriétés syntaxiques du sujet (argument externe) et de l'objet (argument interne)

RÉSUMÉ

Le traitement automatique des relations anaphoriques est nécessaire pour résoudre plusieurs problèmes actuels du TALN comme, entre autres, la recherche et l'extraction d'information, la synthèse et les réponses aux questions, la production automatique des résumés, la traduction automatique et l'interprétation des dialogues.

Un grand nombre d'applications informatiques qui ont été créées pour le traitement automatique du langage naturel analysent l'information linguistique de surface, évitant ainsi les problèmes associés à la complexité de calcul générée par l'information riche sous-jacente aux expressions linguistiques. Pour la résolution des anaphores, ce choix aboutit à des logiciels plus robustes, mais avec des performances limitées.

Nous développons un modèle plus adéquat du traitement humain des relations anaphoriques en général et des relations anaphoriques pronominales en particulier. Nous soutenons que toute simulation d'une faculté cognitive humaine doit avoir à la base un modèle qui intègre nos connaissances de cette Faculté. Dans notre thèse nous utilisons les relations d'asymétries, centrales à la faculté du langage, pour proposer un système automatique de résolution des anaphores pronominales. Ce système serait plus performant que les systèmes appelés robustes.

Nous avons étendu la couverture d'un analyseur morphosyntaxique qui récupère les relations d'asymétrie entre les constituants et nous avons implémenté un module de résolution anaphorique qui utilise l'information fournie par l'analyseur. Nous avons ensuite fait des tests pour évaluer la performance du système ainsi obtenu. L'examen comparatif des résultats appuie la validité de notre recherche.

INTRODUCTION

L'apparition des ordinateurs répond à la nécessité d'effectuer automatiquement des opérations de calcul complexes, exigeantes pour l'esprit humain. J'évoque l'image du fermier qui, devant un seau plein de blé, se demande combien de grains le seau peut contenir. Des approximations volumétriques lui allègeront la tâche de quantifier sa récolte. Mais ce qu'il fait avant le calcul, est un travail de sélection des données d'entrée et des opérations à appliquer afin d'obtenir la précision et la granularité des résultats désirés.

En fait il n'y a pas d'évaluation plus exacte de la récolte que l'ensemble formé par la totalité des grains en question. Tout comme il n'y a pas de représentation plus exacte de la dynamique de la météo que l'ensemble formé par la totalité des phénomènes atmosphériques. Dans ce même sens, nous pouvons voir le système formé par la racine d'un arbre et le sol l'entourant comme étant le calculateur le mieux équipé pour déterminer la dynamique de développement et la forme de la racine en fonction de l'humidité et de la résistance du sol, du besoin de soutien du tronc, etc. Il s'agit ici de *calculateurs* que j'appellerais " lourds ", qui n'assureront pas la reproductibilité des expériences et qui auront des résultats d'une portée limitée. Ce ne sont pas des ordinateurs construits pour un besoin quelconque de l'humain, mais des systèmes implémentés naturellement. Si l'humain veut avoir une représentation utilisable d'un tel phénomène en son entièreté ou en partie, il doit

d'abord procéder à un travail d'extraction de ce qui est significatif dans cette trop lourde implémentation.

La création d'une représentation d'un objet de la cognition est une entreprise risquée, car il s'agit d'un processus subjectif toujours biaisé par les connaissances disponibles au moment de l'observation et par les compétences du chercheur. Dans ce sens, avec du recul, nous pouvons dire que le géocentrisme ptoléméen était bien fondé, mais que l'héliocentrisme copernicien a tout simplement bénéficié d'un cadre de recherche plus généreux.

Supposons, comme Platon autrefois, l'existence d'un œil curieux qui après avoir vu le monde s'interroge sur sa propre apparence. S'il trouve un moyen de se voir, que ce soit son reflet dans un autre œil, comme Platon le suggère, ou n'importe quel autre miroir, la réflexion est une entité essentiellement différente de l'original. Toutefois, pour que l'observation soit possible, le miroir doit reproduire, dans une certaine mesure, une fonction semblable à celle de l'œil - celle de permettre la (re)production des images.

Faisons l'exercice d'établir un parallèle en ce qui concerne la cognition. Quel serait le miroir approprié dans lequel l'esprit curieux pourrait trouver le reflet de ses processus cognitifs? Il n'y a pas moyen de se représenter directement les détails de la cognition. Il existe malgré cela des moyens de décrire se qui se passe à l'intérieur de l'esprit connaisseur : le langage, la peinture, la musique, etc.

En linguistique, depuis les recherches d'Humboldt on ne voit pas le langage et la langue comme des outils destinés à satisfaire la nécessité de communication, mais plutôt, comme Herder, un préalable indispensable à la condition humaine, le générateur même de la pensée. Le langage sert comme outil pour l'investigation de soi même. On se tourne vers le langage pour connaître les détails de l'esprit dans tous les domaines de la cognition.

La faculté du langage est intimement impliquée dans la cognition, fait communément accepté par les chercheurs en sciences cognitives et en linguistique.

Cependant, l'importance que les différents paradigmes de recherche entendent accorder au rôle joué par le langage dans la cognition varie considérablement. Toutefois, il reste incontestable que pour comprendre un certain nombre de processus cognitifs nous avons besoin d'une compréhension approfondie des phénomènes linguistiques qui les accompagnent. Les relations anaphoriques reliant des expressions pronominales à des expressions nominales douées de référence sont des manifestations de tels processus. Les relations anaphoriques sont aussi un bon exemple de la création et de l'articulation du sens dans le langage avec des moyens d'expression minimaux.

Notre travail de recherche bénéficie de la perspective singulière qui s'offre aux explorations poursuivies à la frontière de plusieurs disciplines. Nous nous proposons de contribuer à l'optimisation de la simulation des fonctions cognitives langagières au sein du traitement automatique du langage naturel (TALN), en implémentant un modèle linguistique du traitement de l'anaphore pronominale basé sur les asymétries d'interface. La théorie de l'asymétrie formulée dans (Di Sciullo 1999, 2005a) permet une modélisation dans laquelle les caractéristiques particulières de la grammaire, comme l'ordre des mots et la structure d'arguments de la phrase, sont dérivées à partir d'un ensemble limité de relations asymétriques. Cette même théorie précise les conditions optimales d'interprétation des constructions linguistiques, des conditions qui font partie de la Grammaire Universelle (décrite en 2.1), et donc applicables à toutes les langues.

Selon une définition de la simulation informatique "a computer simulation is not merely a calculation of variables but rather an imitation of a model's behavior in space and time" (Küppers 2005). Le modèle proposé sera plus proche de la modalité dont le cerveau humain traite le langage - comme le confirment les données qui contribuent à soutenir cette hypothèse, données provenant des recherches en linguistique théorique et appliquée, en biologie, en psychologie et en informatique.

Les logiciels en TALN qui intègreront cette simulation optimisée pourront traiter des faits linguistiques jusque-là interprétés erronément par les systèmes existants et ainsi dépasser des seuils de performance en traitement des relations pronominales anaphorique et s'approcher davantage de la performance humaine.

CHAPITRE I

PROBLÉMATIQUE ET OBJECTIFS DE RECHERCHE

Notre recherche porte sur le traitement automatique des relations anaphoriques pronominales. Elle nous a conduit à proposer une solution informatique d'analyse de ces relations, dont la performance dépasse celle qui est obtenue par les systèmes informatiques existants. Ce chapitre précise la problématique de notre recherche et présente ses objectifs.

1.1 Problématique de la recherche

Nos questions centrales sont les suivantes : comment les relations anaphoriques entre un pronom et un antécédent sont-elles traitées par les humains et comment peuvent-elles être traitées efficacement par des systèmes informatiques?

Nous ciblons donc une modélisation informatique de la faculté humaine en mesure de traiter les relations anaphoriques pronominales. Le sujet de notre recherche appartient autant au domaine de l'étude du langage humain qu'à l'informatique. Par conséquent, dans ce chapitre nous décrirons d'abord les propriétés des relations entre pronoms et antécédents du point de linguistique, et nous présenterons ensuite les points principaux de la problématique dans le cadre du traitement automatique des langues naturelles et en particulier du traitement automatique des relations anaphoriques.

1.1.1 L'anaphore – description linguistique

Du point de vue descriptif, l'anaphore est un procédé de reprise d'information antérieurement énoncée. Deux éléments concourent à la réalisation de ce procédé : un qui n'est pas autonome du point de vue référentiel, *l'anaphore*, et un deuxième, *l'antécédent*, qui est une expression autonome du point de vue référentiel et qui participe à l'interprétation sémantique et référentielle du premier. L'anaphore a le même référent que son antécédent. La relation entre ces deux éléments est appelée une relation anaphorique. Les relations anaphoriques sont des reprises sémantiques qui contribuent de façon très importante à l'interprétation des phrases. En effet, l'anaphore et son antécédent entrent dans une relation de dépendance syntaxique et sémantique qui est essentielle à l'interprétation de la phrase dont ils font partie et à la cohérence textuelle. Plus les relations entre les anaphores et leurs antécédents sont transparentes, plus la cohérence d'un texte est solide.

Les relations anaphoriques peuvent être intra-phrastiques, comme c'est le cas en (1), où l'anaphore pronominale, le pronom réfléchi *himself*, et son antécédent *Einstein* figurent à l'intérieur de la même proposition. Elles peuvent aussi être inter-phrastiques ou discursives, comme c'est le cas en (2), où l'anaphore peut être séparée par une ou plusieurs propositions de son antécédent, le pronom *his* dans ce cas.

(1) *Einstein* didn't consider *himself* to be a great mathematician.

(2) I heard that a mathematician found mistakes in *Einstein's* proof. But while the mathematics was somewhat flawed, *his* physical intuition was correct.

L'antécédent doit nécessairement être une expression autonome du point de vue référentiel. Fréquemment, dans le but d'alléger un texte, après une première énonciation de l'antécédent, qui est la source de la référence, la reprise est faite par

des pronoms anaphoriques. L'antécédent et les pronoms anaphoriques reliés constituent ainsi une chaîne référentielle, comme c'est le cas en (3).

(3) *Einstein* had in 1905 a miracle year; *he* published a number of groundbreaking papers and obtained *his* PhD. *His* most famous paper of that year is the "On the Electrodynamics of Moving Bodies".

L'anaphore peut prendre la forme de plusieurs catégories grammaticales, incluant des pronoms, comme dans les exemples (1)-(3), des constituants nominaux, verbaux ou adverbiaux. Elle peut même être nulle, c'est-à-dire qu'elle n'a pas de contenu phonétique, comme en

(4) par opposition à (5) ou le pronom anaphorique a un contenu phonétique manifeste.

(4) Worries go down better with soup than without Ø. (proverbe juif)

(5) Worries go down better with soup than without *it*.

La notion d'anaphore pronominale a été étudiée dans différents cadres théoriques, incluant la linguistique structurale et la grammaire générative. En outre, les travaux dans le domaine ont ciblé des propriétés morphologiques, syntaxiques, sémantiques et pragmatiques des relations anaphoriques pronominales. Étant donné que ce n'est pas l'objectif de cette thèse de faire une revue exhaustive de la littérature dans le domaine, nous présenterons ici les brièvement l'approche de Lyons dans le cadre de la syntaxe structurale, celle de Reboul dans le cadre de la pragmatique, et celle de Reinhart et de Partee en sémantique.

Dans le cadre de la linguistique structurale, Lyons (1977) propose une analyse de la référence anaphorique qui met en relation la deixis[1] et l'anaphore dans le contexte

[1] Termes dont la signification dépend entièrement du contexte immédiat dans lequel ils sont produits. Un synonyme est *indexicalité*.

de l'univers-du-discours. Il fait une distinction entre deux définitions de la référence anaphorique : une dans laquelle le pronom réfère à son antécédent et l'autre, et laquelle il adopte, où le pronom réfère au référent de son antécédent.

Le texte construit l'univers-du-discours qui contient un ensemble structuré de référents possibles. Pendant que la deixis est un moyen d'ajouter des entités à cet ensemble, l'anaphore présuppose que le réfèrent est déjà connu. Dépendamment du contexte, la référence d'un pronom peut être déictique ou anaphorique.

Ainsi, Lyons considère, comme Bühler (1982 p21), que l'anaphore est un type d'indexicalité où l'indication se fait vers le contenu du contexte discursif, mais que l'anaphore est dérivée de la deixis, cette dernière étant est plus fondamentale. "Anaphora presuposes that the referent should already have its place in the universe-of-discourse. Deixis does not; indeed deixis is one of the principal means open to us of putting entities into the universe-of-discourse so that we can refer to them susequently."

Lyons précise que l'ensemble de référents est disponible pendant la durée de la conversation (ou lecture du texte) et que la sélection d'un référent est sujette à des contraintes sur la mémoire de travail; il remarque aussi que d'autres conditions doivent être utilisées pour optimiser l'identification.

Nous ne considèrerons pas d'approche de la linguistique structurale pour la résolution de l'anaphore pronominale que nous formulons dans cette thèse.

Dans le cadre de la pragmatique, plusieurs travaux soulignent le rôle du contexte extralinguistique dans le traitement de l'anaphore pronominale, incluant des notions de relevance, de cohésion du discours et de structure discursive. Mentionnons. à titre d'exemple, les travaux de Sperber et Wilson (1986), Ducrot et Todorov 1972, Halliday et Hasan (1976), Hobbs (1979), Miller (1977), Fradin (1984, 1986), Danlos (2001), Danlos et Gaiffe (2000), Roussarie et al (2007).

Ainsi Reboul (1989) propose une description pragmatique de l'anaphore pronominale, au sens de la Théorie de la pertinence de Sperber et Wilson (1986). Cette théorie, comme celle de Miller (1977), fait la différence entre connaissance lexicales et connaissances pragmatiques. Reboul critique la définition traditionnelle de Ducrot et Todorov (1972) selon laquelle «Un segment de discours est dit anaphorique lorsqu'il est nécessaire, pour lui donner une interprétation (même simplement littérale) de se reporter à un autre fragment du même discours» (Ducrot et Todorov 1972, 358). Elle suggère plutôt que : "[...] les anaphoriques ont souvent cette particularité que l'attribution de leurs référents ne se fait pas tant à travers le discours préalable (ou dans le cas des cataphoriques, ultérieur) mais plutôt à travers la représentation mentale que se fait l'interlocuteur d'un individu ou d'un objet, représentation qui inclut les changements éventuels qui les concernent". (Reboul 1989: 84)

Reboul souligne que les inférences ont un rôle à jouer non seulement dans les problèmes traditionnellement reconnus comme non strictement linguistiques, en particulier les problèmes d'ordre pragmatiques, mais qu'elles sont beaucoup plus présentes dans le traitement morpho-syntaxique et sémantique. Toutefois, elle mentionne que si l'attribution de réfèrent à certaines anaphores pronominales passe par le recours à des connaissances non strictement linguistiques, ceci soulève la question de la légitimité d'une approche non linguistique d'exemples linguistiques.

L'approche pragmatique à l'anaphore pronominale a été mise en cause notamment dans Evans (1977, 1980), Geach (1962), Partee (1997), Sag (1976) Soames (1994) et dans Williams (1977) et Higginbotham (1980). Selon Neale (2005) : "At first blush, the pragmatic theory of pronouns is economical, relying, as it does, on a single reading of say 'his'. It is, however, forced to posit a second bound reading of pronouns, to be employed when they are hooked up to quantified expressions, an heroic effort would be needed to reduce binding to indexicality. More

importantly, the pragmatic theory just makes the wrong empirical predictions in very many case, unilluminating of the way pronouns are used." (Neale 2005, p. 205)

Nous ne considérerons pas l'approche pragmatique à la résolution de l'anaphore pronominale dans cette thèse.

Dans le cadre de la sémantique, plusieurs travaux ont traité de l'interprétation des pronoms et des relations anaphoriques dont ils font partie. Ainsi, selon les travaux de Reinhart (1976, 1983b, 1983a), Grodzinsky et Reinhart (1993, 1993) et les travaux des Bach et Partee (Partee 1978, 1980, 2004; Partee et Bach 1981) il existe une différence fondamentale entre l'anaphore coréférentielle et l'anaphore comme variable liée. Alors que pour Montague (1973) tous les pronoms sont des variables liées, selon Partee (1978, 2004) la coréférence n'est qu'un cas particulier du phénomène plus général d'anaphore pragmatique : "Where I do want to draw a sharp line is between the bound variable use and the pragmatic use of pronouns. The bound variable use is best described at the level of syntactic form and semantic interpretation of single sentences, and the relevant question is not what the pronoun refers to, but what quantifier phrase is binding it. The pragmatic use is best described at the pragmatic level, where the full context of the sentence in use is considered; on the syntactic level, these pronouns are really no different from proper names, and at the semantic level, they can be viewed as free variables or as dummy names." (Partee 1978, p.112)

Ces études ont données lieu à plusieurs travaux à l'interface syntaxe-sémantique, dont les travaux de Reinhart (2006) et de Reuland (2011). Nous inclurons des aspects de la sémantique des pronoms dans l'implantation computationnelle de la résolution de l'anaphore pronominale que nous présenterons au Chapitre 3.

Dans le cadre de la grammaire générative, plusieurs travaux couvrant les aspects syntaxiques de l'anaphore ainsi que les questions associées à leur interprétation à l'interface syntaxe-sémantique se sont développés. En particulier dans le cadre de la théorie chomskyenne qui décrit une différence entre les pronoms simples tels que *him* and *her* et les pronoms réfléchis tels que *himself* et *herself* et qui s'appuie sur les travaux de Reinhart (1999) qui montre le rôle central des relations syntaxiques pour l'identification des relations entre pronoms et leurs antécédents locaux.

Les conditions de la Théorie du Liage proposée par Chomsky (1981) gouvernent les relations anaphoriques entre pronoms et leurs antécédents dans des domaines locaux tels que les propositions. Ces conditions sont universelles et sujettes à des paramètres de variation, qui sont discutés notamment dans les travaux de Rappaport (1986) pour ce qui est des différences entre l'anglais et le russe pour la distribution des pronoms réciproques.

Nous situons notre analyse linguistique des relations anaphoriques pronominales dans le cadre de la grammaire générative (Chomsky 1981 et travaux reliés). Nous excluons donc les approches non génératives au traitement de l'anaphore.

Dans le cadre de la grammaire générative il existe deux analyses majeures du phénomène. La première est de nature représentationnelle, soit la co-indexation, c'est à dire l'assignation du même indice de référence à l'antécédent et à l'anaphore (Chomsky 1981), voir (6). La seconde est dérivationnelle, soit le déplacement de l'antécédent dans une position syntaxique supérieure à l'anaphore. Le constituant déplacé et l'anaphore font partie de la même chaîne dérivationnelle après le déplacement (Kayne 2001). En (7), la position d'origine du constituant déplacé est occupée par une copie silencieuse de ce constituant.

(6) [the students]$_i$ trust themselves$_i$

(7) [the students] trust [~~the students~~ themselves]

Nous adoptons une analyse représentationnelle des relations de liage entre antécédent et anaphore, que nous étayerons au Chapitre 2. Cette analyse ne fait pas appel au déplacement de l'antécédent, mais s'appuie sur la position et les traits l'antécédent et l'anaphore. Le choix d'une analyse représentationnelle plutôt qu'une analyse dérivationnelle des relations anaphoriques réside dans notre objectif de développer un traitement computationnel unifié des relations anaphoriques dans le domaine de la phrase et dans le domaine du discours. Les opérations de déplacement sont limitées au domaine de la phrase, ces opérations ne peuvent s'appliquer dans le domaine du discours. Nous ciblerons les cas de relations anaphoriques ou l'antécédent précède le pronom anaphorique, que ce soit dans le domaine de la phrase que dans le domaine du discours. L'implémentation computationnelle que nous formulons pour le traitement automatique des anaphores pronominales se base sur la Théorie du liage et son extension dans le domaine du discours qui repose sur la relation d'accord asymétrique défini dans la Théorie de l'asymétrie (Di Sciullo 2005a).

1.1.2 Délimitation du domaine empirique de la thèse

Nous illustrons le domaine empirique identifié dans le cadre de cette thèse par les exemples suivants :

(8) These students appreciate themselves.

(9) These students appreciate them.

(10) The students know that they appreciate themselves.

(11) The students know that they appreciate them.

(12) These students are good friends. They appreciate themselves.

(13) These students are good friends. They appreciate them.

Les exemples (8)-(13) illustrent le fait que les pronoms réfléchis, tels que *themselves*, ont des propriétés différentes de celles des pronoms personnels, tels que le pronom *them*, en ce qui concerne l'anaphore, que ce soit dans des phrases matrices (8) et (9), dans des phrases enchâssées, (10) et (11), ou encore que ce soit dans le domaine du discours, et donc dans le domaine discursif minimal constitué de deux phrases adjacentes, (12) et (13).

Nous excluons donc de notre étude les cas des relations cataphoriques, c'est-à-dire, les cas ou l'anaphore précède et annonce l'antécédent, (14)-(18). Ces relations sont exclues du domaine empirique de cette thèse, qui cible les propriétés syntaxiques des relations anaphoriques les plus courantes. Les relations cataphoriques sont plus rares. Ce sont des figures de rhétorique, qui mettent en jeu des règles stylistiques d'antéposition, qui ne font pas partie de notre sujet d'étude.

(14) He is really interesting, this student of physics.

(15) If you see him, tell John that his student is interesting.

(16) He started talking on astrophysics. The speaker was really interesting.

(17) After he arrived, John saw Paul.

(18) John looked at him. Paul was watering the plants.

Nous excluons de plus les relations anaphoriques entre deux constituants nominaux (DP) non-pronominaux, tels que le DP *a genius* dans les exemples (19)-(22), ainsi que les relations anaphoriques qui impliquent des constituants propositionnel, (23), prédicatifs, (24) et adverbiaux, (25). Ainsi en (23) l'antécédent du pronom clitique *le* est la proposition enchâssée *Marie exagère*; en (24) l'antécédent du pronom clitique *l*(e) est la proposition *Paul est intelligent*; en (25) l'antécédent du pronom clitique prépositionnel *y* est le constituent *à Paris*.

(19) John saw Paul. He thinks that he is a genius.

(20) John saw Paul. The genius did not recognize him.

(21) John thinks that he is a genius.

(22) John, the genius, just came in.

(23) Il pense que Marie exagère. Paul le pense aussi.

(24) Paul est intelligent. Marc l'est aussi.

(25) Pierre va à Paris, Luc y va aussi.

Le domaine empirique de notre étude se limite donc à l'anaphore pronominale autant dans le domaine de la phrase que dans le domaine du discours où le pronom est un pronom personnel, possessif ou réfléchi de la troisième personne.

1.1.3 Le traitement automatique du langage naturel

Le traitement automatique des relations anaphoriques fait partie du domaine plus large du traitement automatique du langage naturel (TALN). Le TALN est souvent abordé sans que référence soit faite à l'approche utilitariste choisie dans les années 1960 quand les professionnels de l'Intelligence Artificielle (IA) se sont penchés sur des problèmes reliés au traitement automatique des langues.

Ainsi, à l'époque de la Guerre froide (1947-1989), qui se caractérise par une période de confrontation idéologique et de tension entre les États Unis et l'URSS, de grosses sommes sont investies dans des recherches dans le domaine de la traduction automatique, qui promettent la réalisation d'un traducteur automatique du russe vers l'anglais et vice-versa. L'objectif était, dans ce contexte compétitif, l'applicabilité immédiate des résultats de recherche. Mais les obstacles étaient nombreux : comment

choisir le sens du mot à traduire, où segmenter une expression, comment ordonner correctement les séquences dans le texte sortant ? Il est vite devenu clair qu'une compréhension plus approfondie du langage était nécessaire avant de passer à l'automatisation.

Le langage est une faculté dont la complexité se manifeste à plusieurs niveaux : morpho-lexicale, syntaxique, sémantique, pragmatique. Ces niveaux incluent des unités et des relations spécifiques entre ces unités. Dans une situation normale de communication, les humains possèdent les connaissances nécessaires pour traiter les unités, et les relations entre les unités, connaissances qui leur permettent de comprendre des textes. Les applications courantes en TALN, sont programmées pour traiter le langage à tous ces niveaux. Mais, contrairement aux langages de programmation, le langage naturel utilise très souvent des expressions potentiellement ambiguës, c'est-à-dire des expressions qui sont associées à plus d'une interprétation sémantique, incluant le sens et la référence (Frege 1892/1952). L'humain emploie naturellement les stratégies appropriées pour préciser ou reconstruire le sens et la référence des expressions là où ils ne sont pas explicits et réussit, habituellement, à trouver un moyen de poursuivre la communication. Cependant, cela devient beaucoup plus compliqué quand un système informatique doit construire une représentation à partir d'une expression ambiguë. Le traitement de l'ambigüité est un des problèmes les plus difficiles à traiter en TALN.

Les exemples dans le Tableau 1 illustre par des exemples du français des ambiguïtés décelables à différents niveaux.

lexico-morphologique	*Ton* instrument a changé de *ton*.
syntaxique	Je préfère le gâteau au café.
sémantique	Le pompier sent la fumée.
pragmatique	Le pilote devrait faire une bonne course.

Tableau 1 - Types d'ambigüité

Un texte est un produit d'opérations cognitives, incluant celles de la faculté du langage. L'interprétation des expressions linguistiques qui le compose se construit dynamiquement. Les liens morpho-syntaxiques entre les parties des expressions linguistiques contribuent localement à l'interprétation des constituants syntaxiques et des propositions. La cohérence du texte entier s'appuie sur les relations de référence et de co-référence qui s'établissent entre ses éléments. Une expression référentielle, une fois énoncée, peut être évoquée ultérieurement par une autre expression référentielle. Si les références à une même entité se répètent à plusieurs reprises, nous parlons, comme nous l'avons indiqué plus haut, de la formation d'une *chaîne référentielle*. L'usage adéquat des relations de référence peut alléger la compréhension d'un texte et le rendre plus cohérent. Mais, quand une référence est ambiguë, le sens du texte est miné et, si l'humain peut attendre la suite pour que des nouveaux éléments désambiguïse le texte, pour un système automatique l'analyse est, dans la plupart des cas, compromise.

Les éléments qui participent à l'élaboration de la cohérence d'un texte sont multiples : les catégories grammaticales, l'ordre des mots et des phrases, l'intonation, les relations de référence et de coréférence. Nous avons défini l'anaphore comme une expression référentielle qui renvoie à une entité qui a été énoncée antérieurement dans le discours, qui est son antécédent. La *résolution anaphorique* (anaphora resolution) est le processus d'identification de l'antécédent auquel est liée une anaphore.

Dans les situations réelles de communications, les relations anaphoriques sont rarement ambiguës. Ce n'est pas le cas dans des expressions telles que (26), ou l'antécédent du pronom réfléchi ne peut être que *Marie*. La difficulté s'accroit néanmoins si l'on substitue un pronom personnel au pronom réfléchi, comme en (27). Dans ce cas l'antécédent du pronom personnel n'est certainement pas *Marie*, mais il est impossible de l'identifier hors contexte. La difficulté s'accroît évidemment avec

des phrases complexes, telles que (28) et (29). En (28), le pronom *it* peut prendre n'importe quel antécédent et aussi *the last slice of cake*. En (29) le pronom *it* peut également prendre n'importe quel antécédent et aussi *the plate*. Dans ce cas, la résolution n'est pas trop laborieuse pour un humain, mais elle peut l'être pour un ordinateur s'il n'a pas accès à des éléments de connaissance du monde réel. L'exemple en (30) est d'une grande difficulté même pour les humains.

(26) Mary saw herself in the mirror.

(27) Mary saw her in the mirror.

(28) The boy took the last slice of cake from the plate and ate it hastily.

(29) The boy took the last slice of cake from the plate and washed it hastily.

(30) *Annie* mistakenly sent *Brenda Cathy*'s picture of *herself* and now *she*'s angry with *them* because of *it*.

L'être humain relie des pronoms à leurs antécédents de manière naturelle, puisque les êtres humains sont naturellement prédisposés au langage. Or la nature des systèmes informatiques est différente. Il ne faut pas oublier que les ordinateurs et les principes de la programmation ont été conçus pour répondre à des besoins de calcul balistique et cryptographique. Les calculs mathématiques à eux seuls ne peuvent résoudre des relations anaphoriques telles que celles qui sont illustrées dans les exemples (28)-(30).

L'information pertinente pour l'identification de l'antécédent dans une relation anaphorique provient de plusieurs niveaux linguistiques. Le processus de traitement automatique[2] des anaphores dans le domaine du discours requiert minimalement que

[2] En fait, une précision que nous devons faire est qu'il est rare que le processus de résolution des anaphores se fasse de façon totalement automatique. Il existe peu de systèmes, du moins à notre connaissance, qui puissent faire sans intervention humaine toutes les phases de traitement de l'entrée du texte jusqu'à l'identification des antécédents.

les anaphores et leurs antécédents présumés soient identifiés et qu'une sélection soit effectuée afin de repérer, dans l'ensemble de ses antécédents possibles, le meilleur candidat pour chaque anaphore.

L'information disponible au niveau morpho-lexical, traits Φ (personne, nombre, genre) et catégoriels, est souvent suffisante pour l'interprétation de bon nombre d'anaphores, comme celle en (31), ou le trait de genre du pronom *elle* permet d'identifier *Virginie*, plutôt que *Paul*, comme un antécédent possible, puisque ces deux expressions ont le même trait ϕ [féminin].

(31) *Virginie* aperçut Paul et, tout à coup, *elle* se mit à pleurer.

Il est important de noter que l'accord grammatical ne se réduit pas à l'identité des traits, comme c'est souvent mentionné (cf. Chomsky 1995). Les constituants qui font partie de relations d'accord se distinguent non seulement par les positions qu'ils occupent dans les expressions linguistiques, mais aussi par l'ensemble de traits qui les définissent. Les relations d'identité partielle entre les traits des pronoms anaphoriques et les traits de leurs antécédents nous ferons adopter la notion d'accord asymétrique (Di Sciullo 2005a), détaillée au Chapitre 2. Une telle notion permet de traiter les relations anaphoriques pronominales domaine de la phrase et dans le domaine du discours.

Au niveau syntaxique d'autres informations viennent s'ajouter aux traits morpho-lexicaux pour identifier les constituants, les limites des propositions et les contraintes sur la coréférence qui dépendent de la structure des expressions linguistiques. Par exemple, les relations d'enchâssement de constituent auront des effets sur les relations anaphoriques, qui ne sont pas soumises au contexte du discours. Par exemple, en (32) *Amanda* peut être l'antécédent du pronom *her*, mais pas du pronom *she*. En (33), *the women* peut être l'antécédent du pronom *her*, mais pas de *Amanda.*

(32) Amanda thinks [that she does not like her].

(33) The women [that Amanda met yesterday] does not like her.

De telles relations asymétriques, que nous définirons au Chapitre 2, sont cruciales dans la dérivation des expressions linguistiques, incluant celles qui incluent des relations anaphoriques pronominales.

Il existe des cas où les informations morpho-lexicales et syntaxiques ne suffisent pas pour identifier l'antécédent d'un pronom dans le domaine du discours, comme c'est le cas dans les exemples (28)-(30) plus haut ainsi que dans les exemples (34) et (35) ci-dessous. Sans utiliser des connaissances sémantiques et du monde réel, un système automatique ne peut qu'approximer l'antécédent de *they*. C'est aussi le cas en (36) où des informations non linguistiques sont nécessaires pour qu'un système automatique puisse relier *the Police* à *they* et *the pickpockets* à *them*. Toutefois, nous ne couvrirons pas cet aspect du traitement automatique des anaphores.

(34) Each child ate *strawberries*. *They* were delicious.

(35) *Each child* ate a strawberry. *They* were delighted.

(36) When the passengers brought *the pickpockets*$_i$ to *the Police*$_j$, *they*$_j$ arrested *them*$_i$.

Les stratégies utilisées dans l'identification de l'antécédent d'une anaphore varient beaucoup d'un chercheur à l'autre. Cependant, il y a une tendance à utiliser l'information provenant des niveaux d'analyse linguistique (morpho-lexical et syntaxique dans le Tableau 1) dans la formulation des contraintes qui peuvent éliminer des antécédents possibles. En (37), des trois candidats possibles, *the man* est exclu automatiquement sur la base de la contrainte d'accord en nombre.

(37) *The man* gave *the children some candies*$_i$. *They*$_i$ were delicious.

D'autres informations serviront à la création des préférences qui, au lieu d'écarter des éléments de l'ensemble d'antécédents possibles, vont plutôt les évaluer en leur attribuant un poids plus ou moins important, selon leur probabilité d'accéder au titre d'antécédent. En (37) les propriétés sélectionnelles des deux candidats restant après l'élimination de *the man* seront interprétées de sorte que le poids final de *some candies* sera plus élevé que celui de *the children.*

Toutefois, les traits sémantiques inhérents aux expressions morpho-lexicales, tels que les traits [humain] et [animé] dans le cas de l'exemple (37) permet également d'éliminer *the chidren* des antécédents possibles du pronom *they*. Les propriétés sélectionnelles de l'adjectif *delicious* n'incluent pas ces traits. Les constituants en question ne s'accordent pas du point de vue de leurs traits sémantiques. La relation d'accord est centrale dans les relations anaphoriques. Elle l'est également dans le système de traitement automatique des anaphores pronominales que nous avons développé et que nous décrirons au Chapitre 4.

1.2 Objectifs de recherche

La résolution des relations anaphoriques est nécessaire pour résoudre plusieurs problèmes actuels du TALN comme, entre autres, la recherche et l'extraction d'information, la synthèse et les réponses aux questions, la production automatique des résumés, la traduction automatique et l'interprétation des dialogues.

La performance des systèmes actuels en TALN qui font de la résolution des relations anaphoriques est limitée par les résultats des composantes de la phase de prétraitement, c'est à dire l'analyse automatique de la syntaxe des expressions linguistiques. Mais plutôt que de se confronter à l'amélioration de la phase de prétraitement, la plupart des systèmes informatiques adoptent des approches pauvres

en connaissances linguistiques (syntaxique et sémantique). En conséquence, les systèmes de résolution des relations anaphoriques utilisant des connaissances linguistiques pauvres ne reflètent pas la complexité des processus cognitifs de l'humain et, par conséquent, n'arrivent pas à traiter adéquatement une partie significative des relations anaphoriques présentes dans un texte.

L'objectif central de cette thèse est de développer un modèle plus adéquat du traitement humain des relations anaphoriques en général et des relations anaphoriques pronominales en particulier.

Selon Agre (1997) et Küppers (2005), la simulation informatique devrait être orienté vers une reproduction de la dynamique originelle de la fonction cognitive recherche, et non seulement vers l'obtention des résultats finaux. Par conséquent, notre recherche investiguera d'abord le traitement humain des relations anaphoriques.

Nous assumerons que les relations asymétriques, définies dans le Chapitre 2, sont fondamentales aux opérations de la faculté du langage. L'hypothèse centrale de cette thèse est que ces relations asymétriques doivent être disponibles au moment du traitement automatique des anaphores pronominales afin de simuler le traitement de l'anaphore pronominale par les humains et maximiser ainsi l'efficacité de traitement computationnel des relations anaphoriques.

Notre deuxième objectif est d'implémenter un système entièrement autonome de résolution des relations anaphoriques pronominales qui simule le traitement humain des relations anaphoriques. Pour ce faire, nous augmenterons les capacités d'analyse d'un parseur syntaxique, le parseur LAD, présenté au Chapitre 3. Ceci requiert:

- l'insertion de nouveaux traits (morphologiques (Φ), et sémantiques) ;
- l'extension de la couverture computationnelle du parseur pour permettre le traitement des catégories vides et l'analyse du discours;

- l'implémentation des fonctions d'analyse syntaxique basée sur les relations asymétriques : l'accord et de la c-commande asymétriques.

Ensuite, un nouveau module, aLAD, sera ajouté au parseur syntaxique qui permettra ainsi d'identifier automatiquement les antécédents des anaphores pronominales dans les phrases et dans le discours. Le résultat sera une solution optimisée de traitement automatique des anaphores pronominales dans une implémentation fidèle au traitement humain des relations anaphoriques.

Finalement, nous ferons une évaluation du aLAD et nous utiliserons une série de tests afin de vérifier sa performance ainsi que la validité de notre hypothèse centrale.

1.3 Signification des notre recherche

Notre travail de recherche est une contribution à l'informatique cognitive, elle relie le TALN aux propriétés de la faculté du langage pour l'optimisation de la résolution anaphorique pronominale. Il permet d'intégrer de nouvelles hypothèses dans le cadre du TALN.

Les applications logicielles qui résulteront de notre travail pourrons être utilisées pour améliorer la performance de systèmes de traitement de l'anaphore pronominale, comme nous l'illustrerons au Chapitre 4 et au Chapitre 5 sur la base de l'analyse de la performance de deux systèmes bien connus dans le domaine : MARS (Mitkov's Anaphora Resolution System) et RAP (Lappin 1994). Du point de vue des sciences cognitives, notre recherche apportera des validations computationnelles aux théories récentes sur la faculté du langage, un aspect central de la cognition humaine.

Dans ce chapitre, nous avons présenté la problématique, le domaine empirique et les hypothèses qui sous-tendent notre recherche des points de vue linguistique et

informatique. Nous avons précisé l'enjeu cognitif ainsi que la contribution de la présente étude à l'avancement des connaissances dans le domaine de l'informatique cognitive. Dans le prochain chapitre nous présenterons le cadre théorique que nous avons choisi et ses prédictions pour le traitement de l'anaphore pronominale.

CHAPITRE II

CADRE THÉORIQUE LINGUISTIQUE

Nous avons choisi de mener notre étude sur le type d'anaphore qui est le plus commun, soit l'anaphore pronominale. Le besoin pressant d'un niveau constant de performance et la difficulté de la tache de résolution de ce type d'anaphore a poussé les chercheurs dans les années 1990 vers l'emploi des solutions de traitement, dits "robustes". Ce type de traitement orienté par la constance des résultats favorise l'utilisation d'une analyse linguistique superficielle du texte. Or le phénomène à traiter, soit l'anaphore pronominale, est fondamentalement d'ordre linguistique.

Nous croyons que, pour aboutir à une solution informatisée qui veut simuler le traitement que l'humain fait de l'anaphore, il est nécessaire de choisir un cadre théorique adéquat sur lequel se basera notre traitement computationnel du phénomène. Dans le présent chapitre nous précisons le cadre théorique linguistique que nous choisissons pour notre étude.

2.1 Approches adoptées

Nous adhérons étroitement à l'hypothèse que la faculté du langage est déterminée génétiquement et que les propriétés des représentations linguistiques sont attribuables

à l'interaction de la faculté du langage avec les autres facultés de la cognition. Rappelons que notre objectif dans cette thèse est de développer un modèle computationnel plus adéquat du traitement humain des relations anaphoriques pronominales en s'approchant le plus possible de la manière dont l'humain traite ces relations.

L'appareil théorique que nous assemblons ci-dessous devra ainsi nous permettre d'identifier les éléments fondamentaux qui font l'objet de la computation, de décrire le déploiement du processus computationnel et de spécifier les moyens nécessaires à sa mise en œuvre.

a. La théorie chomskyenne de la faculté du langage

Chomsky postule l'existence d'une capacité innée, inscrite dans le code génétique, de réaliser, récursivement, des combinaisons potentiellement infinies à partir d'un nombre fini d'éléments. Des structures linguistiques communes à toutes les langues définissent une Grammaire Universelle caractéristique à la cognition humaine. Nous adoptons le cadre théorique de la grammaire générative chomskyenne sur la faculté du langage incluant les développements subséquents dans la théorie des principes et paramètres, soit la théorie Gouvernement Liage (1981) et le programme minimaliste (1995).[3]

b. La théorie du liage

Les relations de dépendance référentielle qui lient les pronoms et les anaphores à leurs antécédents sont régies par la théorie du liage. Nous adoptons la théorie du

---

[3] Le programme minimaliste vise la formulation d'un modèle de grammaire basse sur la simplicité des opérations qui font partie de la faculté du langage, c'est-à-dire de la faculté cognitive dédiée au langage, tout en offrant une explication unifiée des phénomènes linguistiques.

liage[4] telle qu'énoncée par Chomsky 1981 (38):

(38) A. Une anaphore doit être *liée* dans son domaine de liage.

B. Un pronom ne peut pas être *lié* dans son domaine de liage.

C. Une expression référentielle doit être *libre.*

Le domaine local de liage que nous ciblerons est le domaine propositionnel. Dans ce domaine, les pronoms réfléchis, tels que *himself*, doivent êtres liées, et les pronoms non anaphoriques, tels que *him*, doivent être libres; de plus les expressions référentielles telles que *John* doivent être libres. Les notions *lié* et *libre* en (39) sont définis à l'aide de la relation de c-commande asymétrique. La relation de *c-commande asymétrique* en (40) est un cas particulier de la relation de c-commande. Enfin, la relation de *c-commande* fait appel à la relation plus basique de dominance (41).

(39) A. α est lié par β ssi :

α et β sont co-indexés[5], et

α c-commandes asymétriquement β

B. α est libre ssi α n'est pas lié

---

[4] Les indices de référence et la notion de co-indexation sont éliminés dans le Programme Minimaliste, dont un objectif est de réduire l'appareil technique au minimum. Nous les utiliserons uniquement pour simplifier la lecture des relations anaphoriques. En outre, nous distinguons le liage de la coréférence. Le liage est limité au domaine de la phrase et est soumis à la condition de c-commande asymétrique. La coréférence peut se faire dans le domaine de la phrase et aussi entre les phrases sans que la c-commande asymétrique soit en jeu.

[5] Deux éléments sont co-indexés s'ils partagent le même index référentiel.

(40) A. relation de *c-commande*

α c-commande β , ssi:

tout γ qui domine α domine β et

α ne domines pas β

B. relation de *c-commande asymétrique*

α c-commande asymétriquement β ssi :

α c-commande β et

β ne c-commande pas α

(41) α c-commande β ssi :

α ne domine pas β , et

le premier nœud à ramification qui domine α domine aussi β

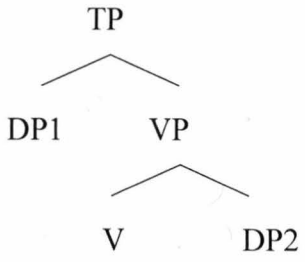
Une conséquence empirique de (38) et (39) est qu'un pronom ne peut pas c-commander son antécédent dans le domaine de la proposition. Les exemples en (42) illustrent ce fait et dans la structure partielle en (43), commune aux exemples en (42), le sujet (DP1) c-commande asymétriquement l'objet (DP2). Nous utilisons des indices ($_{i, j}$, etc.) pour identifier les constituants en relation anaphorique et le symbole # pour les structures qui ne sont pas interprétables selon la coindexation désignée.

(42) a. #He$_i$ likes John$_i$.

b. #She$_i$ hates Mary$_i$.

c. #They$_i$ met the students$_i$.

(43)



Selon le Principe B de la Théorie du Liage, l'antécédent d'un pronom ne peut être un constituant nominal qui le c-commande asymétriquement. Ainsi les exemples en (44) illustrent le fait qu'un pronom en position objet ne peut avoir le constituant en position sujet comme antécédent.

(44) a. #John$_i$ likes him$_i$.

b. #Mary$_i$ hates her$_i$.

c. #The students$_i$ met them$_i$.

Par contre, dans le domaine local d'une proposition, un pronom peut faire partie d'une relation anaphorique avec un constituant qui ne le c-commande pas asymétriquement. Nous désignerons ces cas par l'expression 'anti-c-commande'. La représentation arborescente en (46) illustre cette relation. Ainsi, dans les exemples en (45) les pronoms et les constituants nominaux qui ne les c-commandent pas asymétriquement peuvent être co-référents. En (45a) la relation de coréférence peut être établie entre *him* et *John*, en (45b) entre *her* et *Lucy* et en (45c) entre *them* et the *professors*. La structure partielle en (46) illustre le fait que le pronom *him* n'est pas commandé asymétriquement par le DP *John*, puisque la première projection maximale qui domine *John* ne domine pas *him*, et donc ce pronom n'est pas soumis

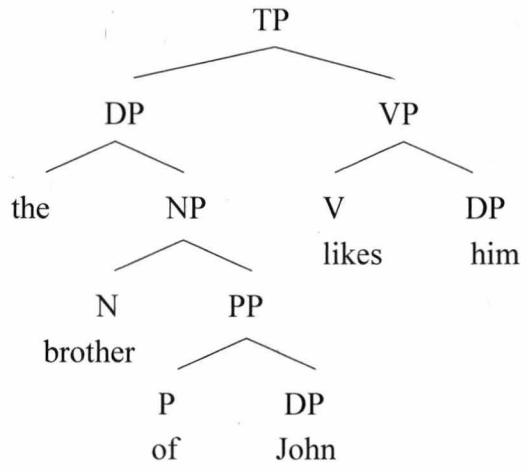
au Principe B de la théorie du Liage. Il peut ainsi faire partie d'une relation de coréférence avec *John*.

(45) a. The brother of John$_i$ likes him$_i$.

b. The sister of Lucy$_i$ hates her$_i$.

c. The students of these professors$_i$ like them$_i$.

(46)



Enfin, un pronom dans une phrase enchâssée peut prendre comme antécédent un pronom dans la phrase matrice. Les exemples en (47) illustrent le fait. En (47a) *him* peut prendre *John* comme antécédent, en (47b), *her* peut prendre *Mary* comme antécédent et en (47c), *the students* peut être l'antécédent de *them*. La relation de c-commande asymétrique est limitée au domaine de la proposition, et donc le sujet d'une phrase enchâssée ne peut être l'antécédent d'un pronom qui occupe la position objet. Mais rien n'empêche qu'une relation de coréférence soit établie entre le sujet d'une phrase matrice et un pronom qui occupe une position sujet ou objet dans la phrase enchâssée.

(47) a. John$_i$ thinks that Paul likes him$_i$.

b. Mary$_i$ thinks that Lucy hates her$_i$.

c. The students$_i$ think that they like them$_i$.

Les exemples ci-dessus présentent tous les cas de figure qui seront couverts par le traitement automatique de l'anaphore pronominale qui sera proposée au Chapitre 4 à l'exception de l'anaphore pronominale discursive, qui sera discutée à la section 4.1 dans le cadre de la discussion sur la relation d'accord dans le domaine discursif.

c. La théorie de l'asymétrie

Nous adoptons en outre le cadre théorique de la Théorie de l'asymétrie proposée par Di Sciullo (1999, 2005a). La Théorie de l'asymétrie offre une modélisation de la Faculté du Langage qui s'appuie sur des notions qui ont un rôle central en linguistique, mais aussi en biologie et en physique. Les opérations de la Faculté du langage génèrent des relations structurales incluant les relations asymétriques, telles que les relations de préséance, de dominance et la c-commande asymétrique, définie en (40b) plus haut. Selon la théorie de l'asymétrie, les relations asymétriques sont privilégiées dans la dérivation des expressions linguistiques, comme c'est le cas par exemple pour l'ordre linéaire des constituants linguistiques et les structures d'arguments. "[…] the elemental properties of grammar at the initial state are relations and not singular features or categories and that the grammar grows on the basis of the development of elemental relations onto more complex ones…" (Di Sciullo 2003b p.8.)

La Théorie de l'asymétrie prédit correctement plusieurs propriétés des langues naturelles. Pour les propriétés morphologiques, elle prédit correctement que les constituants des expressions morphologiques sont asymétriques ne peuvent être inversés, (48). En outre, elle prédit également que les opérateurs qui font partie des expressions morphologiques, tels que la négation et les modaux ont une portée fixe.

Par exemple en (49), l'expression complexe *unthinkable*, incluant le morphème de négation (NEG) *un-* et le morphème modal (MOD) *–able* est non ambiguë. C'est à dire que la négation aura toujours portée large sur le modal. Ainsi (a) aura l'interprétation de (b) et non de (c).

(48) a. anti-constitu-tion-nal

b. * tion-nal-constitu-anti

c. * constitu-anti-tion-nal

(49) a. unthinkable

b. NEG MOD (impossible de pouvoir penser)

c. MOD NEG (#possible de ne pas pouvoir penser)

Cette Théorie a des conséquences pour les propriétés des structures morphologiques (Di Sciullo 1996, 1999, 2000, 2004, 2014), ainsi que pour les relations entre les traits dans constituants morphologiques (Di Sciullo 1997, 2009; Di Sciullo et Tenny 1997; Di Sciullo et Slabakova 2005; Di Sciullo et D'Alessandro 2008; Di Sciullo et Landman 2009).

Pour les propriétés syntaxiques, la Théorie de l'asymétrie permet d'expliquer la présence des asymétries entre argument interne et argument externe, illustrées en (50) et entre argument interne et adjoint, illustrée en (51), ou les constituent rayés sont des copies silencieuses des constituants déplacés (Di Sciullo, Paul, et Somesfalean 2003; Di Sciullo 2006; Di Sciullo et Isac 2008b).

(50) Asymétrie Argument Interne / Argument Externe

a. ?What do you recall [~~what~~ whether [Bill bought ~~what~~]].

b. *Who do you recall [~~who~~ whether [~~who~~ bought a book]].

c. What do you think [~~what~~ that [Mary left ~~what~~ on the table]].

d. *Who do you think [~~who~~ that [~~who~~ left the book on the table]].

(51) Asymétrie Argument Interne / Adjoint

a. Which problem did you wonder [how ~~which problem~~ [PRO to solve [~~which problem~~] [~~how~~]]].

b. *How did you wonder [which problem [PRO to solve [~~which problem~~] [~~how~~]]].

c. ??Which letter did you meet [a man who worded [~~which letter~~] carefully].

d. *How carefully did you meet the man [that worded the letter [~~how carefully~~]].

En outre, la Théorie de l'asymétrie suppose l'existence de la Grammaire Universelle (GU), qui définit notre connaissance innée du langage. Pour que le contenu de la Grammaire Universelle soit accessible au traitement, la théorie de l'asymétrie postule l'existence de l'Analyseur Universel (AU) qui a le rôle de récupérer ces connaissances et de les rendre interprétables. Selon la théorie de l'asymétrie, les relations entre la grammaire et l'analyseur prennent la forme de l'hypothèse GU/AU en (52).

(52) Hypothèse GU/AU

La Grammaire Universelle (GU) est conçue tel qu'elle assure l'analyse optimale des expressions linguistiques en termes des relations asymétriques ; l'Analyseur Universel (AU) est conçu pour la récupération optimale des expressions linguistiques en termes des relations asymétriques (Di Sciullo 1999)

Une conséquence de cette hypothèse est que les points de symétrie résultent dans des correspondances non optimales entre la grammaire et l'analyseur, ce qui donne

lieu à dans des interprétations non optimales des expressions linguistiques. De plus, des expérimentations ont été faites pour étudier si la perception humaine diffère pour les structures avec des arguments et des modificateurs. Les résultats obtenus indiquent que:

- le traitement cognitif est plus long pour les structures verbales avec préfixes externes ou internes (Tsapkini, Jarema et Di Sciullo 2004)
- le traitement cognitif est plus long pour les composés adjoint-verbe que pour les composés objet-verbe (Di Sciullo et Tomioka 2008)
- le temps de réaction pour structures moyennes est plus long que pour les structures passives (Di Sciullo, et al. 2008)

Les structures dérivées par la grammaire sont très complexes et leur interprétation optimale nécessite le concours de l'Analyseur pour assurer une mise en forme canonique, compréhensible des expressions linguistiques. Di Sciullo postule l'Interprétation Sous Asymétrie (53) comme une condition d'accessibilité aux données. L'AU assure la récupération et le traitement optimal des relations asymétriques sous le contrôle d'IUA. (Di Sciullo 2000, p. 6)

(53) Interpretation Under Asymmetry (IUA)

a. an interpretation is optimally obtained under a unique local asymmetrical relation.

b. a local asymmetrical relation optimally supports a unique interpretation.

La théorie de l'asymétrie offre un modèle minimaliste de la GU dans lequel les dérivations morphologiques et syntaxiques se font en parallèle. Les deux opérations génériques de la GU, en (54) et (55), s'appliquent, si la condition (56) est remplie, dans diverses dérivations produisant des objets linguistiques morphologiques ou syntaxiques. (Di Sciullo 2005a, pp. 29-30)

(54) *Shift*(α , β): Given two objects α , β , Shift (α , β) derived a new object δ projected from α .

(55) *Link*(α , β): Given two objects α and β , α sister-containing β , Link (α , β) derives the object (α , β), where α and β are featurally related

(56) *Agree*(j1, j2)[6]: Given two sets of features j1 and j2, Agree holds between j1 and j2, iff j1 properly includes j2, and the node dominating j1 sister-contain the node dominating j2 .

La relation *Agree* en (56), ou la relation d'*accord asymétrique*, défini l'accord entre deux constituants en termes d'asymétrie de traits (relation d'inclusion propre) et d'asymétrie configurationnelle ('sister contain' ou c-commande asymétrique). La relation d'accord asymétrique est également en jeu dans les relations anaphoriques et nous l'utiliserons dans le traitement des relations d'anaphore pronominale que nous présenterons dans le Chapitre 4.

d. La théorie de la modularité de l'esprit

Pour l'aspect architecture cognitive de cette thèse, nous adoptons le cadre théorique de la modularité de l'esprit tel que formulé par Jerry Fodor (1983). Selon la théorie de la modularité de l'esprit, les différentes fonctions cognitives, dont le traitement syntaxique et sémantique des expressions linguistiques, sont réalisées par des modules différents.

Fodor emploie la métaphore de l'esprit qui fonctionne comme un ordinateur, l'architecture mentale étant présente depuis le départ. Chaque fonction du cerveau se

---

[6] *Agree* (j1, j2), ou l'accord asymétrique, requiert qu'une relation de sous-ensemble propre existe entre les éléments qui subissent Shift et Link.

réalise grâce à des modules spécialisés, implémentés dans des réseaux neuronaux identifiables, qui cueillent l'information sensorielle et la traitent de manière indépendante en la canalisant vers les systèmes centraux qui sont la mémoire et la pensée.

e. La théorie du liage discursif

L'interprétation des relations anaphoriques dans le domaine intrapropositionnel est régie par les conditions A, B et C de la Théorie du Liage, énoncés en (38) plus haut, qui identifie ainsi les relations anaphoriques permises et celles interdites dans le domaine intrapropositionnel. Ces principes ne sont pas adaptés au domaine interpropositionnel. La c-commande asymétrique ne s'applique pas dans le domaine du discours. Afin de traiter les relations anaphoriques pronominales dans le domaine du discours nous adoptons la notion de liage discursif (57) telle qu'énoncée par Di Sciullo (2005b, p. 332). Dans le domaine discursif, l'opération de liage est toujours conditionnée par l'existence de la relation d'accord asymétrique.

(57) A pronominal must be linked in the Domain of the Discourse.

Étant donnée la notion de liage discursif, (57), l'accord asymétrique, (56) et l'ensemble de traits en (58), les relations anaphoriques en (59) peuvent être identifiées. En effet, comme le montre Di Sciullo (2005c), pour qu'une relation anaphorique soit bien formée, les traits formels et sémantiques des antécédents et des pronoms doivent être en relations de sous-ensemble propre. L'ensemble de traits de l'antécédent d'un pronom est le super-ensemble alors que l'ensemble de traits du pronom lié par cet antécédent est le sous-ensemble. En (59a) nous avons une relation anaphorique bien formée, les traits de l'anaphore *himself* étant inclus dans l'ensemble de traits de son antécédent, qui le c-commande asymétriquement (liée dans son domaine de liage). En (59b) le pronom *him*, bien qu'en accord asymétrique avec son antécédent local (ses traits son un sous-ensemble des traits de son antécédent), ne peut pas entrer dans une relation anaphorique avec *the chief officer* parce que, selon le

principe B de la Théorie du liage (38), un pronom ne peut pas être lié dans son domaine de liage. L'analyse est comparable en (59c) avec l'observation que *him* peut chercher un antécédent à l'extérieur de son domaine local (la phrase enchâssée). Le cadre discursif est encore plus large en (59d), où le pronom *him* ne peut pas avoir un antécédent dans son domaine local, mais est libre de chercher dans le domaine discursif un antécédent avec lequel il est en relation d'accord asymétrique.

(58) Traits formels et sémantiques[7] (Di Sciullo 2005c)

	Traits formels			Traits sémantiques		
	pers	num	gen	Ir	ani	w
DPro fort	+	+	+	+/-	+/-	+/-
DPro faible	-	+	+	+/-	+/-	+/-
DP	3rd	+	+	+	+/-	+/-

(59) (Di Sciullo 2005c)

a. [the chief officer] trusts [himself]

{+Ir, +ani, +w} { -Ir, +ani, +w}
{+3rdpers,+ sing, +mass} {+3rdprs,+ sing, +masc}

b. [the chief officer] trusts [him]

{+Ir, +ani, +w} {-Ir, +ani , -w}
{+3rdprs,+ sing, +masc} {+3rdprs,+ sing, +masc}

c. [the president] thinks [that [the chief officer] trusts [him]]

{+Ir, +ani, +w} {-Ir, +ani, +w}
{3rdprs,+ sing, +masc} {+3rdprs,+ sing, +masc}

[7] Les traits sémantiques sont : référence indépendante [±Ir], animé [±ani], partie/tout [±w].

d. [the president] talked to [the members of the company]
| └..... | .....
{+Ir, +ani, +w} {+Ir, +ani, +w}
{+3rdpers,+ sing, +masc} {3rdpers,+ plur, +masc}

today. [The reactions of the shareholders] were unequal.
……..... | .....
{+Ir, -ani, +w }
{+3rdpers,+ plur, +masc}

[The minutes of the meeting] indicate [that [the chief officer]
……..... | ..... | .....
{+Ir, -ani, +w} {+Ir, +ani, +w}
{+3rdpers, + plur, +neut} {+3rdpers,+ sing, +masc)

trusts [him]
…….. | ┘
{-Ir , +ani, +w}
{+3rdpers,+ sing, +masc}

Nous n'adopterons pas l'analyse des relations pronominales dérivées par déplacement de constituant, telle que proposée dans Kayne (2001). Selon cette analyse l'antécédent et le pronom sont d'abord engendrés dans un même constituant, puis l'antécédent se déplace dans une position supérieure. Étant donné que le déplacement de constituant est limité au domaine de la phrase, et que nous voulons proposer un traitement computationnel de l'anaphore pronominale qui s'étend au discours, nous écartons une analyse selon laquelle les relations anaphoriques sont dérivées par déplacement.

Notre implantation computationnelle de la résolution de l'anaphore pronominale est basée sur les asymétries morpho-syntaxiques, telle que la c-commande asymétrique et l'accord asymétrique. Nous exclurons donc les analyses sémantiques,

telle que celle de Reinhart et Reuland (1993) et Heim (1982) par exemple, ou des analyses pragmatiques de l'anaphore pronominale, telle que celle de Levinston (2000) et Liu (2010).

Dans ce chapitre, nous avons motivé le choix du cadre théorique linguistique et cognitif que nous utilisons dans notre recherche. Dans le prochain chapitre, nous justifierons l'approche que nous adoptons du point de vue informatique.

CHAPITRE III

TRAITEMENT COMPUTATIONNEL DES ANAPHORES

Dans ce chapitre, nous présentons les propriétés des applications informatiques dédiées à la résolution des anaphores, plus particulièrement des anaphores pronominales. Nous discutons des conséquences que la richesse des informations linguistiques peut avoir sur le résultat du processus de résolution. Nous motivons également le choix de l'analyseur syntaxique que nous utiliserons par la suite et nous décrivons les propriétés de ce dernier.

3.1 Le traitement automatique des relations anaphoriques – état de l'art

Selon Hirst (1981), le traitement automatique des relations anaphoriques a débuté dans les années '60 avec le système STUDENT (Bobrow 1964). Par la suite d'autres systèmes ont été développés, dont SHRDLU (Winograd 1972), LSNLIS (Woods, Kaplan and Weber 1972), MARGIE (Rieger 1975), le *case-driven parser* (Taylor 1975), le système *naïve* (Hobbs 1976), et le système utilisant des préférences lexicales (Wilks 1975) (Jackson et Moulinier 2002, p.189).

Ces systèmes utilisent, la plupart du temps, des méthodes de vérification de contraintes syntaxiques et des heuristiques combinées avec des opérations de synthèse et d'inférence. La résolution procède par élimination, en écartant chaque antécédent qui viole une contrainte. Les résultats de ces systèmes sont souvent insatisfaisants, mais il y a aussi des exceptions, comme l'algorithme de Hobbs, dont la performance est comparable aux résultats des systèmes les plus récents. Quelque peu singulière dans cette période est l'approche de Klappholz (Klappholz et Lockman 1975). Ils sont les premiers à envisager une solution complète au problème de la résolution des relations anaphoriques. Selon eux, il est nécessaire d'opérer sur une représentation globale du sens, représentation formée d'un ensemble de connaissances du monde réel et une mémoire contenant le *focus* des propositions antérieures.

La structure du discours gagne de l'importance dans l'implémentation des systèmes TALN après les études de Kantor et Grosz sur le rôle du focus dans la compréhension des textes en général, et dans la résolution des relations anaphoriques en particulier. Des systèmes qui utilisent ces concepts sont proposés par Sidner (1978) et Webber (1978). Grosz et Sidner poursuivent le travail sur la théorie du focus et développent *centering theory* (centrage d'attention sur un fragment ou un autre du texte), utilisée dans la résolution automatique des anaphores par Brennan, Friedman et Pollard en 1987, Carter en 1987, Walker en 1989, Hahn et Strube en 1997 et Tetreault en 1999. (Tetreault 2001)

Un exemple intéressant d'intégration des contraintes et des préférences dans la résolution des relations anaphorique est le système LUCY (Rich et LuperFoy 1988), qui présente une structure distribuée dans laquelle une multitude de facteurs négocient le poids final d'un antécédent possible. Selon ces auteurs: "there exists no single, coherent theory upon which an anaphora resolution system can be built, there are many partial theories each of which accounts for a subset of phenomena that influence the use and interpretation of pronominal anaphora."

Carbonell et Brown (1988) adoptent aussi une stratégie multi-approches, c'est-à-dire une stratégie qui utilise une combinaison de contraintes et de préférences d'ordre syntaxique, dont les restrictions « sélectionnelles », ainsi que des contraintes de préférences reliées à la structure du discours et aux connaissances du monde réel.

Une caractéristique commune aux systèmes de résolution utilisant des connaissances linguistiques riches est le fait qu'ils visent des cadres d'application limités et qu'ils sont très dépendants de l'information initialement codée. Il va sans dire qu'un système conçu pour traiter le discours technique et qui est muni de connaissances détaillées sur les règles qui régissent seulement ce type de discours ne peut pas traiter convenablement le discours narratif usuel. Dans le but de construire des systèmes avec des performances plus constantes par rapport aux différents types de textes à analyser, les chercheurs entendent faire des concessions et une nouvelle vague de systèmes apparaît après 1990, qui utilisent des stratégies moins coûteuses du point de vue computationnel et plus robustes. Les analyseurs syntaxiques partiels, les ontologies, l'apprentissage machine et le traitement probabiliste font partie des systèmes de nouvelle génération.

Dans le domaine spécifique de la résolution des relations anaphoriques, cette révolution commence par la proposition de Dagan et Itai (1991) qui utilise des connaissances linguistiques *pauvres* : un analyseur syntaxique partiel et une base de données avec des statistiques sur des colocations, c'est-à-dire la fréquence de co-occurrence de certains mots. L'anaphore est substituée à chacun de ses antécédents possibles, et, chaque fois qu'une substitution est opérée, une évaluation de la structure nouvellement obtenue est faite en utilisant les bases de données de colocations. Le but est de trouver l'antécédent qui, une fois substitué à l'anaphore, abouti à la structure dont le taux d'occurrence le plus élevé dans le texte. Cette approche a été bien reçue, et elle a ouvert la porte à d'autres systèmes dits robustes, orientés performance.

Un de ces systèmes parmi les plus discutés est le RAP (voir Chapitre 5 pour une description), proposé par Shalom Lappin et Herbert Leass (Lappin et Leass 1994). Il utilise la sortie d'un parseur syntaxique et une modélisation élémentaire d'un mécanisme d'attention, sans employer de conditions sémantiques, ni d'informations concernant la structure du discours ou du monde réel. Ses performances le situent parmi les systèmes les plus réussis. De bons résultats sont aussi obtenus par le système de CogNIAC (Baldwin 1997), qui utilise un parseur superficiel et seulement six règles de sélection d'antécédents qui sont basées sur des heuristiques.

En outre, Ruslan Mitkov propose un système automatique de résolution des anaphores qui inclut plusieurs stratégies robustes et qui nécessitent peu de connaissances syntaxiques et sémantiques. MARS (Mitkov 2002), bénéficie du soutien d'un analyseur syntaxique performant, le FDG Parser (Tapanainen and Jarvinen 1997). Cependant, le parseur FDG n'analyse pas des informations syntaxiques importantes telles que la relation entre l'argument externe et l'argument interne[8], voir Figure 1, ce qui a des conséquences négatives pour l'efficacité de la résolution anaphorique. Nous allons le montrer au Chapitre 5 et contraster l'analyse de MARS avec celle que nous avons développée en utilisant un parseur basé sur des connaissances linguistiques riches. Ainsi, dans l'analyse de la phrase "Mary sees Paul." , FDG ne récupère pas les relations d'asymétrie que le prédicat verbal a avec, d'une part, son argument interne (l'objet) et, d'une autre, avec son argument externe (le sujet). L'analyse de FDG utilise plutôt une relation de c-commande symétrique entre l'argument interne et l'argument externe d'un prédicat, comme illustre la Figure 1, ou *Mary* et *Paul* sont directement dominés par *sees*.

---

[8] La structure d'arguments est l'information présente dans une entrée lexicale d'un prédicat précisant les relations syntaxiques et sémantiques que le prédicat peut avoir avec ses arguments. La structure d'argument des propositions est dérivée par les opérations syntaxiques et reconstruite par l'analyseur syntaxique. Voir la section 3.4 pour le fonctionnement de l'analyseur syntaxique LAD.

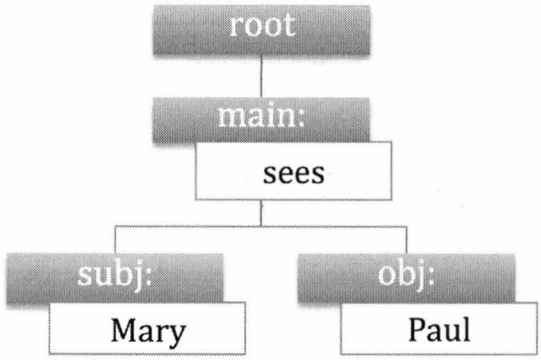


Figure 1 - Exemple de relation symétrique entre les arguments d'un prédicat

Les systèmes basés sur les corpus, ou qui ne font pas appel à des connaissances linguistiques riches, tels que ceux que nous avons mentionnés à la section précédente, n'identifient pas correctement une partie significative des relations anaphoriques. Ces systèmes peuvent obtenir de bons résultats, comme c'est le cas du système de Lappin et Leass (1994), par exemple, mais c'est souvent sur des domaines d'analyse étroits et la modélisation qu'ils proposent du traitement du langage par la cognition humaine est limitée.

Shalom Lappin propose aussi une solution hybride dans Lappin (2003). Nous présentons dans le Tableau 2 un aperçu schématique de son modèle. Ce dernier inclut un algorithme séquentiel qui emploie dans un premier temps, dans le Module 1, des méthodes syntaxiques et des heuristiques moins coûteuses du point de vue computationnel pour éliminer la majorité de cas de relations anaphoriques improbables. Pour les cas ambigus qui restent, dans le Module 2, le traitement utilise des informations sémantiques et des informations provenant du monde réel. La complexité de calculs augmente avec le raisonnement par abduction[9], dans le Module 3, jusqu'à ce qu'un seul antécédent soit sélectionné. Nous mentionnons aussi la

[9] L'abduction est un type de raisonnement qui, par opposition à la déduction et à l'induction, permet de formuler des hypothèses et de les utiliser dans l'explication de certaines observations.

recherche dans des listes Gazeteer[10] (p.ex. Evans et Orasan 2000, utilisée plus tard par MARS) qui rend possible l'utilisation de connaissances externes ou d'ensembles de données annotés pour faciliter la résolution de relations anaphoriques.

Module 1	**Module 2**	**Module 3**
Syntactic Salience & Recency Measures+	*Statistically Determined Lexical Preference Measures* →	*Abductive inference* ⟹
Syntactic & Morphological Filtering →	New Ranked Candidate List →	resolved
Ranked Candidate List →	Confidence Metric 2 →	
Confidence Metric 1 →	correctly resolved;	
correctly resolved;	unresolved ⟹	
unresolved ⟹		

Tableau 2 - L'algorithme séquentiel (Lappin 2003, p.13)

Stuckardt (2004) fait une évaluation de l'algorithme de Lappin, qu'il compare brièvement aux systèmes utilisant un traitement parallèle, avant de se pencher sur une solution séquentielle. Ses résultats montrent que, du point de vue de l'ingénierie linguistique, une approche séquentielle du traitement automatique des anaphores présente des avantages. Stuckardt discute également d'un autre problème de fond dans le TALN : le choix d'architecture du système.

Il est communément admis que pour faire un traitement intelligent du langage humain, un système TALN a besoin d'une architecture appropriée. C'est-à-dire une architecture qui permet de traiter des représentations issues d'une modélisation

[10] Ensemble de listes contenant des noms d'entités (p. ex. villes, organisations, jours de la semaine, etc.)

adéquate du langage naturel. Cependant, une des meilleures méthodes pour vérifier la validité d'un modèle du langage naturel reste son implémentation à l'aide d'un système TALN.

3.2 TALN avec des connaissances linguistiques pauvres

Les premiers ordinateurs utilisés pour le traitement automatique des relations anaphoriques disposaient de ressources computationnelles limitées. Ce fait a eu des conséquences directes sur le choix des informations linguistiques à traiter. Lorsque les ressources computationnelles plus puissantes sont devenues plus accessibles, des mesures de simplification du traitement computationnel ont été prises afin d'éviter une explosion combinatoire de solutions possibles, et pour assurer une certaine constance de la performance pour les différents types de textes à analyser. Dans un souci d'apporter plus de précision et de stabilité dans la performance des systèmes automatiques de traitement du langage naturel, les approches dites robustes proposent des algorithmes qui utilisent surtout des heuristiques, des analyses syntaxiques partielles et un traitement probabiliste au détriment des connaissances linguistiques riches. (Dagan et Itai 1991, Lappin et Leass 1994, Baldwin 1997, Mitkov 2002).

Un défaut majeur des approches robustes en TALN est qu'elles ne considèrent pas la complexité des processus cognitifs humains en jeux dans l'usage du langage en général, et, en particulier, dans le traitement des relations anaphoriques. En conséquence elles n'arrivent pas à identifier une partie significative des relations anaphoriques présentes dans un texte. La plupart du temps, un système robuste analyse une phrase comme une concaténation de symboles sans connaître, par exemple, sa structure d'argument.

En outres, des propriétés essentielles du langage naturel, en particulier les relations d'asymétrie, qui sont propres aux structures d'arguments et aux relations anaphoriques ne sont pas prises en compte par les systèmes dits robustes. Nous contrastons l'analyse partielle privilégiée par les approches *robustes* (Figure 1, ci-dessus) avec une analyse plus riche d'une structure verbale (Figure 2), ou l'argument externe DP1 c-commande asymétriquement l'argument interne DP2. En effet, la catégorie qui domine immédiatement DP1, soit vP domine également DP2, mais ce n'est pas le cas pour DP2, puisque VP domine immédiatement DP2 mais non DP1.

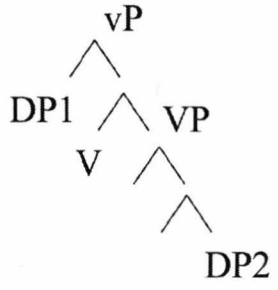


Figure 2 - Exemple de relation asymétrique entre les arguments d'un prédicat

Kennedy et Boguraev (1996) présentent un algorithme de résolution anaphorique qui modifie et étends l'algorithme de Lappin et Leass (1994), qui se ne base pas sur une analyse syntaxique profonde, mais prends uniquement appui sur la catégorisation des constituants des textes soumis à l'analyse, des classes de coréférence ainsi que des contraintes négatives (60), excluant la coréférence pronominale dans certain cas. Cet algorithme atteint une performance de 75%, ce qui constitue une performance inférieure à celle de l'algorithme de Lappin et Leass, qui est de 85%.

(60) Condition 1: A pronoun cannot corefer with a coargument.

Condition 2: A pronoun cannot corefer with a nonpronominal constituent which it both commands and precedes.

Condition 3: A pronoun cannot corefer with a constituent which contains it.

Les relations de dominance et de préséance syntaxiques prises indépendamment ne sont pas suffisantes pour identifier les relations anaphoriques pronominales. Par contre la notion de c-commande asymétrique, subsume ces deux relations structurales et contribue à l'efficacité du traitement des anaphores pronominales. La relation de c-commande asymétrique est limitée au domaine de la proposition, et la notion d'accord asymétrique, nécessaire indépendamment, offre une solution élégante au traitement des anaphores pronominales discursives, comme nous le montrerons au Chapitre 4.

D'autres approches utilisent des ontologies, telles que WordNet, et des heuristiques basées sur des analyses syntaxiques partielles (chunking) identifiant des modèles fréquents (patterns) d'anaphore pronominale. Par exemple, dans Lin et Liang (2011), des stratégies de filtrage sont utilisées pour la reconnaissance de l'anaphore. Ces stratégies éliminent par exemple des pronoms pléonastiques comme antécédents et déterminent si certains pronoms ont plusieurs antécédents, comme c'est le cas de pronoms pluriels tels que *they* et *themselves*. De plus les antécédents possibles sont évalués selon un ensemble de traits syntaxiques et sémantiques qui déterminent leur proéminence. L'évaluation du système a partie d'un ensemble de résumés de court textes biomédicaux (103 résumés MEDLINE contenant 177 paires d'anaphore pronominales) révèle un taux de succès de 78% lorsque des traits syntaxiques, tels que la position de l'antécédent en position de tête de constituant est pris en compte. Cette étude indique que l'intervention de traits syntaxiques et sémantiques dans les systèmes de résolution d'anaphore pronominale augmente leur performance.

Features	Score
Recency	0-2
Subject and Object Preference	1
Grammatical Role Agreement	1
Number Agreement	1
Longest Common Subsequence	0-3
Semantic Type Agreement	-1 if not or +2
Biomedical Antecedent	-2 if not or +2

Tableau 3 - Évaluation de proéminence des antécédents possibles (Lin et Liang 2011)

Les systèmes robustes sont des outils relativement fiables lorsqu'ils s'appliquent dans des domaines restreints (comme le traitement de la documentation technique), mais leurs faiblesses sont connues et reconnues par leurs auteurs même : "In spite of the comparatively high results obtained by some knowledge-poor algorithms and the claim that certain types of anaphor can be successfully resolved without real-world knowledge, the lack of semantic, domain or real-world knowledge imposes serious limitations. [...] Another significant problem for automatic anaphora resolution systems is that the accuracy of the pre-processing is still too low [...] whereas POS[11] taggers are fairly reliable, full or partial parsers are not." (Mitkov 2002, p. 278)

De plus, dans un système de traitement linguistiquement pauvre de type modulaire, par exemple le système de Mitkov, les erreurs d'analyse produites par le module en aval sont amplifiées dans le traitement ultérieur. Dans la Figure 3 présentons deux exemples d'analyse échouée par MARS. Afin d'obtenir les résultats de l'analyse nous avons utilisé l'implémentation (Mitkov 2002) disponible en ligne à l'adresse: clg.wlv.ac.uk/demos/MARS/index.php

An archaeologist discovered a skeleton. He sent it to the museum.

- He appears in paragraph 2, sentence 2, from position 1 to position 1. It is singular. The antecedent is indicated to be a skeleton in paragraph 2, sentence 1, from position 4 to position 5.
- it appears in paragraph 2, sentence 2, from position 3 to position 3. It is singular. The antecedent is indicated to be An archaeologist in paragraph 2, sentence 1, from position 1 to position 2.

Jane's sister arrived. Paul's sister saw him.

- him appears in paragraph 2, sentence 2, from position 4 to position 4. It is singular. The antecedent is indicated to be Jane's sister in paragraph 2, sentence 1, from position 1 to position 2.

Figure 3 - Deux analyses de Mars

[11] Le POS (Part of Speech) Tagger associe une catégorie (Nom, Verbe, etc.) aux items lexicaux qui font partie des expressions linguistiques.

Nous avons essayé de voir si une corrélation existe entre les performances des systèmes TALN et la richesse de l'information linguistique qui leur est disponible. À cette fin nous avons compilé, dans le Tableau 4, le taux de succès de plusieurs systèmes *robustes* qui utilisent un degré variable de connaissances linguistiques.

Bien que les meilleures performances se situent près de 90%, atteintes par des modules spécialisés dans des tâches particulières qui sont appliqués à des discours contrôlés (Dagan et Itai 1991, Chieu et Ng 2002), le taux de succès moyen des systèmes *robustes* ou à connaissances linguistiques pauvres se situe autour de 70%. Si nous considérons la disponibilité de l'information linguistique, nous observons l'existence d'une relation de proportionnalité directe entre cette disponibilité et la performance des systèmes de résolution des anaphores. Ces limites sont dénoncées par des chercheurs en TALN, qui réclament l'utilisation de connaissances linguistique riches: "So the crucial problem here, at the heart of many NLP applications, is the accurate identification of the structure of sentences and entire discourses". (Grishman 2007, p. 18)

Chercheur	Performance	Disponibilité de l'information linguistique
Cardie & Wagstaff	52.80%	apprentissage non supervisé (clustering) sur texte brut
Soon, Ng & Lim	58.90%	apprentissage sur corpus prétraité
Mitkov (MARS)	62.44%	analyse syntaxique superficielle
Lappin & Leass (RAP)	85%	analyse syntaxique riche ; apprentissage sur corpus
Kennedy & Boguraev	75%	corpus prétraité manuellement
Aone & Bennett	76.27%	apprentissage sur corpus prétraité manuellement (analyse du discours)
Baldwin	77.90%	analyse syntaxique et sémantique superficielle ; prétraitement riche du corpus
Ge, Hale & Charniak	82.90%	étiquetage manuel du corpus
McCarthy & Lehnert	85.80%	corpus prétraité manuellement
Dagan & Itai	86.84%	corpus prétraité manuellement

Tableau 4 - Taux de succès de systèmes à connaissances pauvres

Dans cette section, nous avons présenté les débuts du traitement automatique des relations anaphoriques et nous avons identifié les prémisses qui ont mené à l'adoption du TALN avec des connaissances linguistiques pauvres. Nous avons indiqué certaines conséquences négatives associées à ce type de traitement et nous avons fourni des taux de performance pour ce type de systèmes.

3.3 TALN avec des connaissances linguistiques riches

Dans le cadre du TALN, Hirst (1981) définit la relation anaphorique comme référence subséquente *abrégée* à une entité déjà mentionnée, référence faite avec l'attente qu'elle puisse être interprétée correctement par le lecteur. Bien que non définie dans Hirst, par *abrégée* nous entendons « ayant une dimension moins importante du point de vue du coût computationnel ».[12]

Suivant le cadre théorique que nous adoptons, nous observons que l'humain fait un usage optimal de la faculté du langage. Ainsi, lorsqu'un locuteur utilise une anaphore, il s'attend à ce qu'elle soit correctement interprétée par son interlocuteur. Il postule donc la disponibilité pour son interlocuteur des ressources nécessaires à son interprétation. Ainsi, pour qu'un système TALN puisse reproduire cette performance, il devrait bénéficier d'une modélisation des ressources disponibles à l'interlocuteur. Cependant, il reste très difficile d'identifier l'ensemble fini de ressources que l'humain utilise dans le traitement du langage en général, et, plus spécifiquement,

[12] Nous ne définissons pas cette notion en terme de nombre de caractères. Bien qu'une anaphore pronominale contienne généralement un moins grand nombre de caractères que son antécédent DP, ce n'est pas toujours le cas, comme dans : *Jim* saw *himself*.

dans le cas qui nous intéresse, dans le traitement de la résolution de l'anaphore pronominale. De plus, une fois identifiée, toute cette information n'est pas toujours adaptée au traitement par ordinateur. "Whereas it is almost certain that complex semantic and pragmatic knowledge is needed to solve all the well-formed anaphors, it is highly improbable that this would soon be available for a computational system. Even elaborated semantics and complete parse trees aren't yet realistic for unrestricted text processing." (Popescu-Belis and Robba 1997, p. 94)

Parmi les travaux récents qui adoptent une approche dite riche à la résolution anaphorique pronominale, mentionnons les travaux de Refoufi (2014) pour un système résolution pour les textes en français. Il présente un modèle de résolution d'anaphore pronominale qui utilise des documents dont l'analyse est représentée avec des balises XML. Ce modèle, dont l'architecture est présentée à la Figure 4 est basé sur un algorithme conçu pour utiliser des connaissances riches fournies par un parseur basé sur une grammaire de clauses définies[13] (DCG). La représentation XML permet de recueillir différents types d'information, d'ordre morphologique, syntaxique, sémantique et discursif qui seront utilisées par le module de résolution d'anaphores.

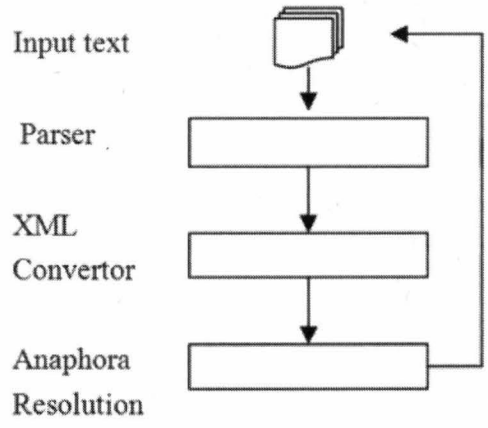


Figure 4 - Architecture globale du système (Refoufi 2014)

[13] Dans cette analyse (Definite Clause Grammar) utilisée dans la programmation logique une grammaire est représentée comme un ensemble de clauses définies du premier ordre.

Les contraintes et préférences suivantes sont utilisées ensuite pour éliminer des candidats possibles:

Contraintes :

- L'accord morphologique : l'anaphore et l'antécédent doivent s'accorder en personne, nombre et genre

- Contraintes syntaxiques sur le domaine local

Préférences :

- Rôle grammatical : les entités en position sujet/objet sont plus saillantes

- Parallélisme : le pronom et l'antécédent appartiennent à la même catégorie syntaxique

- Mention répétée : les entités mentionnées plus fréquemment sont plus saillantes

- Proximité : l'antécédent le plus récent mentionné dans le discours. Le candidat le plus proche de l'anaphore est privilégié.

Les erreurs les plus communes qui influencent le processus de résolution dans ce système sont issues de la délimitation des constituant nominaux. Ceci est une conséquence du fait que les DCG ne tiennent pas en compte des relations de c-commande asymétriques, qui relient par exemple une tête nominale et ses dépendants, comme c'est le cas du parseur LAD. Par contre, les représentations HTML utilisées pour encoder les dépendances référentielles extraites du lexique et de la grammaire pendant l'analyse syntaxique rendent la résolution plus explicite et donc plus facile à exploiter. Selon Refoufi, la performance du système pour la résolution des relations anaphoriques pronominale dans des textes scolaire en français serait proche de 85%.

Nous considérons que, afin de s'approcher davantage du traitement humain des relations anaphoriques, une modélisation informatique devrait s'appuyer sur :

- une théorie de la grammaire qui peut dériver toutes les propriétés des expressions linguistiques ;

- un analyseur morpho-syntaxique qui interprète la grammaire et qui récupère localement la structure des phrases ;
- un module de résolution des anaphores qui utilise l'information récupérée par l'analyseur.

Pour que ce système soit complet devront s'ajouter l'information lexicale, le raisonnement basé sur les connaissances du monde, les contraintes provenant de la modélisation des facultés comme l'attention, la mémoire, etc. Voir la Figure 5 pour le schéma que nous proposons d'un algorithme générique complet de résolution des relations anaphoriques, modélisé sur la manière dont l'humain traite les relations anaphoriques.

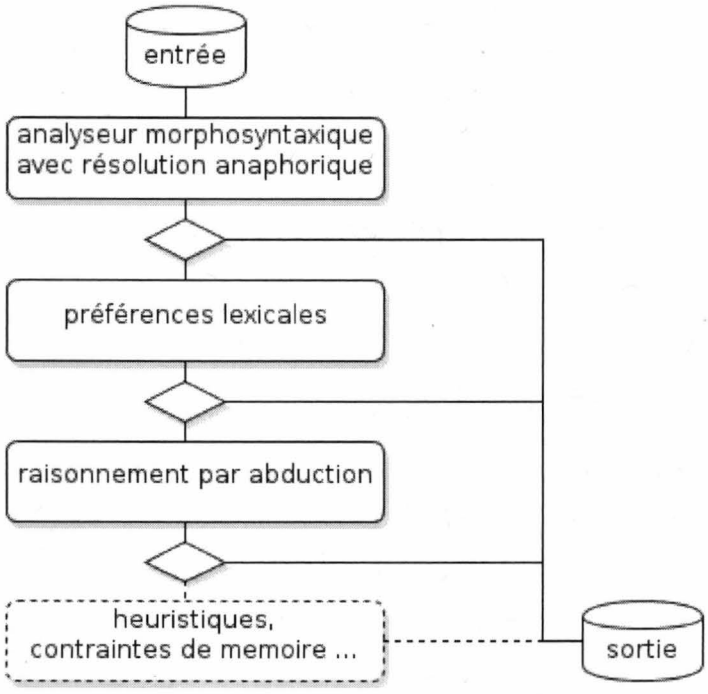


Figure 5 - Composantes d'un algorithme de résolution générique

Dans les paragraphes qui suivent nous présenterons les propriétés générales des analyseurs qui implémentent une théorie de la grammaire, soit la Théorie de l'asymétrie (Di Sciullo 2005a), qui constituent une modélisation de la faculté du

langage, c'est-à-dire des opérations cognitives en jeu dans le traitement du langage par l'humain. Ces opérations sont présentées au Chapitre 4. Les hypothèses sur lesquelles se fonde la Théorie de l'asymétrie, en particulier le fait que les relations dérivées par les opérations de la Faculté du langage sont asymétriques, sont testables expérimentalement, et plusieurs implémentations informatiques ont été réalisées, dont un analyseur morphologique, un analyseur morpho-syntaxique et un analyseur syntaxique. Nous les décrivons brièvement ci-dessous.

L'analyseur morphologique CONCE-MORPHO-PARSE

Le prototype CONCE-MORPHO-PARSE[14] (Di Sciullo 1997a, 1999) analyse les asymétries formelles et sémantiques internes aux mots. "[It] recovers predicate-argument, modifier-modified and operator-variable asymmetries. The first relation is licensed under argument saturation, the second relation is licensed when there is an identification relation, the third relation is licensed when an operator binds a variable." (Di Sciullo 2000) Le prototype récupère les asymétries formelles et sémantiques et utilise ces relations pour produire une analyse fine de la structure catégorielle et sémantique des mots complexes. Une trace de l'analyse du mot *re-enlarge* est présentée dans la Figure 6. (Di Sciullo 1999a, p. 374)

L'analyseur morphologique CONCE-MORPHO-PARSE analyse des mots complexes de bas en haut et identifie les relations structurales entre les constituants, notamment les relations tête-complément (HC) et les relations adjoint-tête (AH) an associant des affixes et des racines dans l'analyse automatique d'un mot (W). L'exemple présenté dans la Figure 6 illustre l'analyse de verbe de l'anglais *reenlarge* (élargir à nouveau) effectué par l'analyseur morphologique.

---

[14] Mise en œuvre réalisé au Laboratoire de recherche sur les asymétries d'interfaces (LAD)

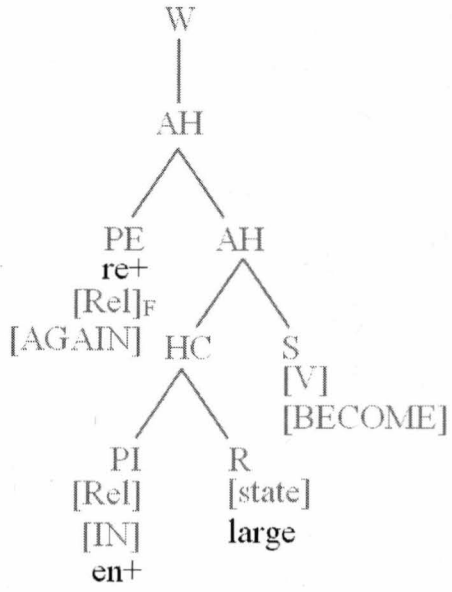


Figure 6 - Analyse dans CONCE-MORPHO-PARSE

Cet analyseur s'appuie sur des entrées lexicales spécifiées pour les traits syntaxiques et sémantiques et un ensemble de règles de production pour construire la structure morpho-conceptuelle sous-jacentes des mots dérivés. L'analyse automatique du verbe causatif *reenlarge* dérive l'association entre le préfixe interne *en+*, qui a le trait syntaxique REL (PREPOSITION RELATIONNELLE) et le trait conceptuel IN (INCHOATIF) et l'adjectif *large*, qui a le trait syntaxique R (REFERENCE) et le trait sémantique STATE. Le résultat de cette analyse est une catégorie AH qui est elle-même associée au verbe (V) inchoatif (BECOME). Le résultat de cette association est une structure AH. Cette structure est elle-même associée au préfixe *re+* qui est un préfixe externe (E) de type relationnel (REL_F) qui a le trait sémantique itératif (AGAIN). Cette dernière association est analysée comme AH et la structure est reconnue par l'analyseur comme un mot (W).

L'analyseur morphosyntaxique PAPPI

L'analyseur PAPPI est une implémentation de la sélection X-barre proposée par Di Sciullo (1995). Il s'agit d'un parseur de type *shift-reduce*[15], qui analyse les expressions morphologiques de gauche à droite et de bas en haut, de manière déterministe. (Di Sciullo et Fong 2001, 2005). Cet analyseur fournit des analyses des structures morphosyntaxiques de l'anglais, telles que *writer* et *writable*, mais il ne fournit aucune analyse pour les structures morphosyntaxiques qui ne correspondent pas à celles de l'anglais, telles que *#arriver* et *#arrivable.*

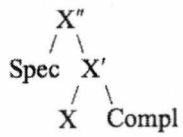
Di Sciullo (1995) propose de traiter les restrictions combinatoires des affixes et des racines en termes de la Théorie X-barre[16] (Chomsky 1970, et travaux reliés) et de liage argumental. L'effet de la sélection X-barre et du liage argumental peut être visualisé par les représentations dans la Figure 7 pour les affixes *–er*, *-able* et les affixes causative et inchoative. Ces affixes sont des têtes morphologiques, c'est-à-dire des éléments qui projettent leurs traits dans leur structure maximale, qui inclut une position de Spécificateur et une position de Complément.

Selon la sélection X-barre, les affixes imposent des contraintes sur le domaine de leur complément, ce qui est illustré dans la Figure 7. Ainsi, le suffixe nominal *–er*, dans *employer* requiert que le Spécificateur de son complément soit une position argumentale (*a+*). L'affixe nominal *–ee* (comme dans *employée*) requiert que le spécificateur et le complément dans le domaine de son complément soient des

---

[15] Un parseur ascendant qui commence l'analyse avec l'entrée (feuilles ou nœuds terminaux) et procède vers le symbole initial de la grammaire (la racine) en utilisant des opérations de décalage et réduction. (shift reduce)

[16] Selon la théorie de la sélection X-barre, une tête morphologique a une sélection configurationnelle représentée par la structure X-barre, (ci-contre). Une structure X-barre contient une tête X, qui projette ses traits aux projections X' et X''. Chaque tête est associée à un complément (Compl) et un Spécificateur (Spec) :



positions *a+*. Les propriétés sélectionnelles du morphème causative abstrait *caus* en Figure 7 différent de l'inchoatif *inc*. Contrairement à l'inchoatif, le causatif requiert une position de spécificateur *a+*. Les morphèmes *–er*, et *inc* lient une position de Spécificateur non argumental (*a-*) dans le domaine de leur complément.

La théorie de la sélection configurationnelle inclut le principe central suivant: Toutes les positions A-barre (*a-*) doivent être liées à des position A (*a+*) dans le domaine de leur complément.

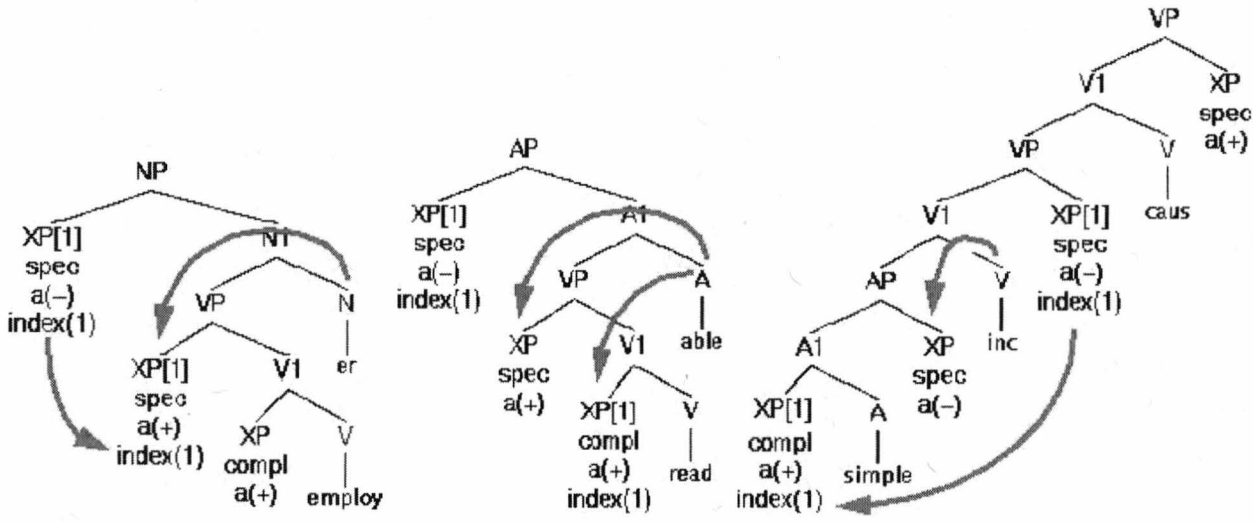


Figure 7 - Exemples de sélection configurationnelle, Di Sciullo et Fong (2002)

L'analyseur permet de retracer la structure argumentale et le liage argumental, comme l'illustrent les traces suivantes pour la dérivation du verbe causative-inchoatif *formalize* à la Figure 8. Di Sciullo et Fong (2001, 2005) montrent également les effets que la variation dans l'ordre des constituants de la structure asymétrique Spécifieur-Tête-Complément peut avoir sur la complexité de traitement. Ainsi 96 actions de l'analyseur sont nécessaires pour l'analyse de *formalize* sur la base d'une structure X-barre ou le Spécificateur est à gauche, alors qu'uniquement 21 actions sont requises

pour l'analyse de cette forme sur la base d'une structure X-barre ou le Spécificateur est à droite.[17]

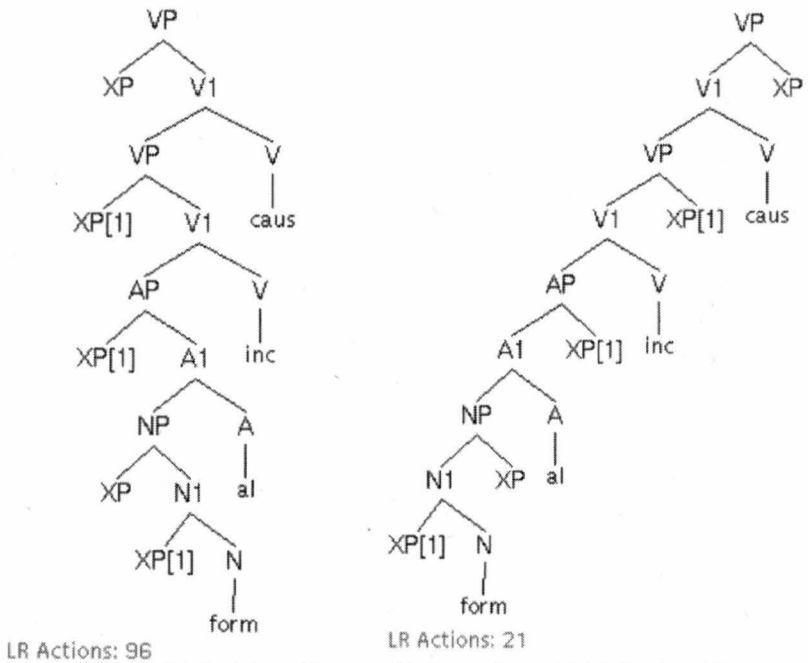


Figure 8 - Analyse de *formalize* dans PAPPI (Di Sciullo et Fong 2005, p. 13)

L'analyseur syntaxique LAD

Une autre implémentation informatique de la Théorie de l'asymétrie est le parseur syntaxique LAD (Di Sciullo et al. 2006, 2010), qui simule le traitement des relations asymétriques et réalise la structure de la phrase. Nous réservons la section suivante pour la description détaillée de l'analyseur syntaxique LAD, car nous utiliserons cet

[17] Di Sciullo et Fong (2001, 2005) pour le détail du traitement computationnel et ses conséquences pour la réduction de la complexité.

analyseur pour notre implémentation informatique afin d'optimiser le traitement automatique des relations anaphoriques.

3.4 Approche adoptée. Application informatique : l'analyseur LAD

L'analyseur LAD[18] est une implémentation informatique de la Théorie de l'asymétrie présentée au Chapitre 2, un parseur LL(1)[19] qui analyse la structure de la phrase de manière incrémentale en récupérant les relations asymétriques et produit à la sortie un arbre syntaxique. Chaque nouveau mot analysé peut seulement étendre l'arbre partiel obtenu jusque là, sans qu'un retour en arrière soit permis. L'analyseur implémenté en Python reçoit en entrée une définition d'une grammaire et génère un ensemble de sous-programmes qui constituent le parseur qui interprète cette grammaire. Un aperçu du code généré est présenté en (60) Le parseur impose des restrictions afin que les opérations de la grammaire soient appliquées dans des domaines locaux sous la condition d'accord asymétrique.

(60)

```
def __CG__grp_CPrel():
if tracing('all'): trace.trace('[ grp_CPrel\n')
xp = new_domain('CP')
if heads_group('WHP'):
  res = __CG__grp_WHP()
  xp = __CG__CmaxP_s(xp, res)
elif heads_group('PP'):
  res = __CG__grp_PP()
  xp = __CG__CmaxP_s(xp, res)
else: trap()
```

---

[18] L'analyseur syntaxique *LAD* a été implémenté au Laboratoire de recherche sur les asymétries d'interfaces. (Di Sciullo et autres 2006, 2010)

[19] L'analyseur LL(k) fait une analyse descendante de l'entrée de gauche à droite dans une passe en regardant un nombre de *k* mots en avant.

```
  if heads_group('DP'):
    if heads_group('DP'):
      res = __CG__grp_DP()
      xp = __CG__TP_s(xp, res)
    else:
      trap()
    if word.haswordcat('V'):
      xp = __CG__vP_h(xp, word)
      sel = word.selects()
      next()
      if sel:
        if heads_group('DP'):
          res = __CG__grp_DP()
          xp = __CG__VP_c(xp, res)
        elif heads_group('PP'):
          res = __CG__grp_PP()
          xp = __CG__VP_c(xp, res)
    else:
      trap()
  elif word.haswordcat('V'):
    if word.haswordcat('V'):
      xp = __CG__vP_h(xp, word)
      sel = word.selects()
      next()
      if sel:
        if word.haswordcat('Adv'):
          xp = __CG__FP_s(xp, word)
          next()
    else:
      trap()
    if heads_group('DP'):
      res = __CG__grp_DP()
      xp = __CG__VP_c(xp, res)
    elif heads_group('PP'):
      res = __CG__grp_PP()
      xp = __CG__VP_c(xp, res)
    elif heads_group('CPrel'):
      res = __CG__grp_CPrel()
      xp = __CG__VP_c(xp, res)
  else: trap()
  close_domain()
  if tracing('all'): trace.trace('grp_CPrel ]\n')
  return xp


...


def __CG__FP_h(xp, word):
  proj, pos = 'FP', 'h'
  link((proj, pos), ('CmaxP', 'H'))
```

```
  return xp

def __CG__FP_s(xp, word, flip=0):
  proj, pos = 'FP', 's'
  set(proj, 'flip', flip)
  shift(word, proj, pos)
  return xp
```

Le parseur interprète la grammaire en appliquant des règles dans des domaines locaux. Les règles des domaines spécifient la réalisation maximale[20] des projections syntaxiques, alors que les instanciations potentielles des domaines sont précisées de manière exhaustive par les règles des groupes. Un aperçu de la grammaire décrivant les domaines et les groupes est présenté en (61)

(61)

```
Dom CP
  Proj CmaxP
        Spec [getw(word, 'cat') == 'WHadj']
              : shift, link(FPP.Spec);
        H    [word.get_cat() == 'Vaux']
              : shift, link(FauxP.h);
        Cmpl shift;

...
Grp DP
  ( PName <DdefP.h> | Dpron <DpronP.h> ... AP <MP.Spec> ...
```

Le domaine le plus inclusif est le domain CP qui est ouvert par défaut au début de l'analyse. Les deux domaines, CP et DP, peuvent être ouvertes alternativement à

---

[20] Les descriptions étendues des domaines CP et DP sont les suivantes :

CmaxP > FP > TP > FnegP > FmodP > FauxP > FPP > vP > VP

DmaxP > FP > QP> Num > ADJ > nP >NP

l'intérieur des groupes (Grp CP, Grp DP, Grp PP, etc.) Les domaines CP et DP sont générés par défaut à leur extension maximale, mais une position devient manifeste à l'intérieur d'un domaine seulement lorsque'elle est occupée par un item lexical provenant de l'entrée. Chaque emplacement dans un domaine correspond à une projection syntaxique contenant trois positions : *Spec*, *H* et *Cmpl*, respectivement spécificateur, tête et complément[21]. Les définitions des domaines incluent ce qui est connu en informatique comme « actions sémantiques ». Pour chaque position, des conditions d'application des opérations *Shift*, *Link* et *Flip*[22] peuvent être spécifiés. Dans le cas où certaines conditions (comme *Agree*) ne sont pas remplies, l'analyse est interrompue avant la fin de la phrase.

Le lexique est implémenté selon les spécifications de la théorie de l'asymétrie, qui postule que la configuration des traits provenant du lexique est telle qu'elle peut être reconnue directement par les opérations *Shift* et *Link*. La théorie de l'asymétrie régit des relations entre des ensembles de traits. Ces traits font partie des spécifications des items lexicaux. Certains items lexicaux ont des traits morphologiques, soit nombre, personne et genre. Ce sont des noms, pronoms personnels et démonstratifs, quantifieurs, verbes et auxiliaires. Des traits sémantiques, par exemple les traits +/- partie, +/- animé, et autres, font aussi partie des spécifications lexicales. Ces traits sont décrits dans (Di Sciullo, 2005c).

---

[21] *Spec, head* et *comp* (ou *specifier, head* et *complement*) sont des positions décrites par la théorie X-bar (Chomsky 1970)

[22] *Shift* est une opération qui assemble des éléments.

Link est une opération de mouvement et de liage.

Flip est une opération qui contribue à la linéarisation. Elle s'applique à un arbre minimal et produit une image en miroir de celui-ci. En syntaxe *Flip* dérive l'ordre des modificateurs « lourds » générés dans le spécificateur des projections fonctionnelles (à gauche de la tête fonctionnelle) pour créer une image en miroir de l'arbre. (Di Sciullo 2005a, p. 29)

Description de l'analyse

La phrase en entrée est chargée dans la mémoire de travail et l'analyse commence avec le premier mot de cette phrase. Une fois la catégorie du premier mot récupérée dans le lexique, elle est comparée avec la catégorie initiale du premier groupe du domaine courant. Si aucune correspondance n'est trouvée, l'analyse est interrompue. Si une correspondance est trouvée le mot est associé à cette position. L'analyse continue avec la récupération dans le lexique de la catégorie du deuxième mot, qui est comparée avec la catégorie initiale de la position suivante disponible dans le groupe. Le processus continue jusqu'à ce que un mot soit trouvé qui ne peut pas être inséré dans le groupe courant[23]. Le groupe est alors fermé et l'analyse continue avec le mot courant et le groupe suivant disponible dans la définition du domaine. Ainsi, un nom commun est reconnu comme une tête et il est assemblé avec un déterminant pour former un DP. Ensuite le DP est placé dans la position *Spécificateur* du TP.

Le parseur analyse deux types d'argument du verbe, interne et externe Figure 9. L'asymétrie de structure d'arguments, soit la relation de c-commande asymétrique entre l'argument externe (le sujet logique) et l'argument interne (l'objet logique), est récupérée par l'analyseur dans des propositions affirmatives et interrogatives, ainsi que dans les constructions passives.

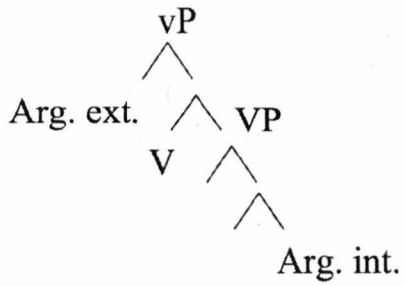


Figure 9 - Structure verbale incluant l'argument externe et l'argument interne

[23] Il est possible qu'un nouveau domaine soit ouvert à l'intérieur du groupe courant pour permettre l'insertion d'un certain mot.

Dans le cas de la proposition affirmative dans la Figure 10 l'analyseur place correctement l'argument externe *the student* dans le *spec.TP*, la position de sujet surface, et identifie le lien vers *spec.vP*, la position de base de l'argument externe. L'argument interne *the book* est correctement généré dans la position *comp.VP*. Comme nous l'avons mentionné plus haut, les éléments barrés sont des copies indiquant la position d'origine du constituant déplacé. L'asymétrie argument/modificateur est également récupérée, puisque le modificateur *in the morning* est généré dans le domaine fonctionnel et il c-commande asymétriquement l'argument interne *the student*.

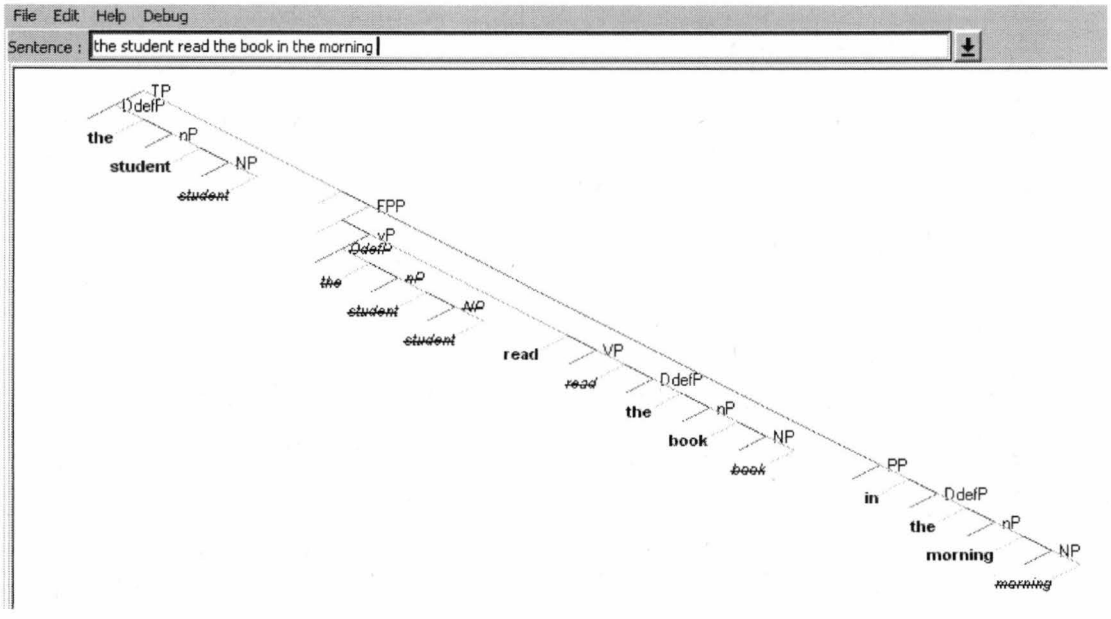


Figure 10 - Analyse de la proposition affirmative *The student read the book in the morning.*

Les asymétries de structure argument / non-argument sont aussi récupérées, par exemple dans les cas où l'argument interne est dans une position non canonique, comme c'est le cas des structures passives. Les actions de l'analyseur sont déterminées par la présence ou l'absence d'éléments lexicaux dans certaines

positions, par les traits lexicaux et, selon le cas, par la satisfaction de certaines conditions grammaticales, comme l'accord.

Dans le cas des constructions passives, la présence du verbe auxiliaire BE et la présence d'un trait participe passé sur le verbe déterminera la création d'un lien entre l'argument interne et le sujet de surface, selon le code en (62) et la 110.

(62)
```
Proj vP
Spec shift;
H [check_feat(FauxP.h, 'be') & check_feat(word, 'pastp') ]
                        : shift, xlink(TP.Spec, VP.Comp);
                        [...]

      Cmpl -
```

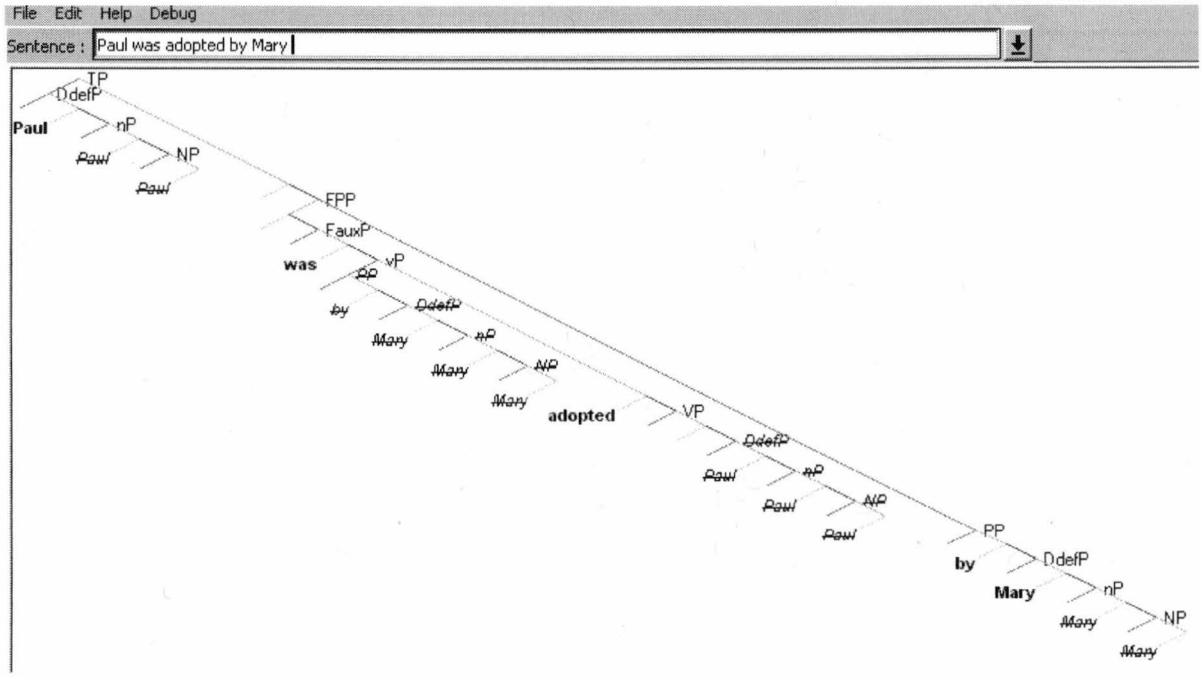


Figure 11 - Analyse de la construction passive *John was adopted by Mary.*

Une proposition interrogative est toujours analysé comme un CP (dans les questions oui/non, la position [Spec, CP] est vide, mais la tête contient le verbe

auxiliaire. Le mouvement de ce dernier de la position [Head, AuxP] vers [Head, CP] est dérivé par *Link*.) La proposition interrogative dans la Figure 12 inclut dans le constituent CP un mot *wh-* dans *spec.CP* et avec des liens vers *spec.TP* et *spec.vP*. En outre, l'analyseur récupère correctement les différentes structures pour les verbes transitifs comme *read* dans la Figure 10 et les inaccusatifs comme *arrive* dans la Figure 12.

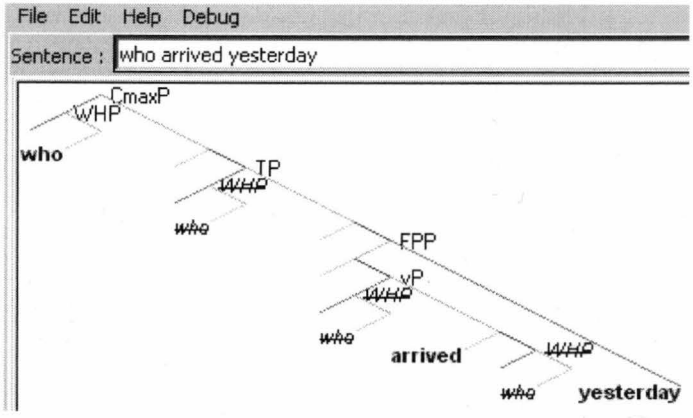


Figure 12 - Analyse de la proposition interrogative *Who arrived yesterday?*

Une des caractéristiques de ce parseur est le fait que toute opération générique est implémentée de manière à permettre le changement des paramètres de variation linguistique, de sorte que le parseur peut facilement analyser d'autres langues que l'anglais. Ainsi, *Shift*, *Link*, *Agree* et *Flip* ont une forme générique applicable à toute langue et les différences entre les langues sont le résultat du changement des paramètres de variation.

Les avantages généraux d'un parseur basé sur la théorie de l'asymétrie (Di Sciullo, 2005) sont multiples. Celui-ci permet un traitement unifié des langues: les principes de base restent identiques, mais les paramètres changent, pour accommoder les particularités de chaque langue (par exemple : la position du verbe monte à une position supérieure (TP) en français, mais pas en anglais; l'adjectif

s'accorde avec le nom dans les langues romanes, mais pas en anglais). Il permet également un traitement des expressions linguistiques basées sur les traits formels et sémantiques. De plus, il permet un traitement basé sur des traits structurés (la reconnaissance des traits sémantiques permet la prédiction des structures de manière à optimiser la vitesse de traitement). Ensuite, il permet un traitement des relations anaphoriques, étant donné l'implémentation de l'opération *Agree*, ce qui permet de cibler les antécédents potentiels d'une anaphore (les relations sémantiques permettront par la suite la détermination de l'antécédent optimal). En outre, il permet l'intégration d'autres modules avec des tâches spécifiques : QA (systèmes de questions-réponses), création des ontologies, etc. Mais, plus important encore pour l'objectif que nous nous avons fixé, l'analyseur LAD est une implémentation d'une théorie de la grammaire qui offre une modèle adéquat du traitement du langage tel que fait par l'humain et qui fait des prédictions validées par des tests neuro et psycholinguistiques sur le fonctionnement de la faculté du langage, tel que vu au Chapitre 2. L'information qui fait partie du lexique du LAD soutient davantage l'implémentation d'une solution optimisée de résolution des anaphores pronominales.

Ce chapitre a porté sur le traitement computationnel des relations anaphoriques, et, en particulier, des relations anaphoriques pronominales. Nous avons comparé les deux classes d'applications informatiques, pauvres et riches du point de vie linguistique. Dans la classe des applications informatiques riches, nous avons identifiés les propriétés générales de l'analyseur LAD, basé sur la théorie de l'asymétrie. Nous présenterons l'aLAD, le module de résolution des anaphores pronominales, associé au LAD, dans le chapitre suivant.

CHAPITRE IV

CONTRIBUTION DE LA THÈSE

Dans ce chapitre nous exposons la solution pour la résolution anaphorique pronominale que nous avons développée : aLAD, un module de résolution des anaphores pronominales qui utilise des connaissances linguistiques riches fournies par le parseur LAD.

Pour le but de notre recherche nous nous sommes basés sur les fondements théoriques exposés dans le deuxième chapitre et nous avons étendu la couverture du parseur LAD pour créer les prémisses pour le développement de aLAD.

Le parseur LAD (Di Sciullo et al. 2006) a été conçu pour traiter les relations asymétriques qui font partie de la structure hiérarchique des expressions linguistiques. Cependant, au moment où nous avons commencé notre contribution au développement du parseur, certaines fonctions n'étaient pas encore implémentées. Pour rendre possible le traitement des anaphores, nous avons implémenté, entre autres, un lexique augmenté des nouveaux traits morphologiques et sémantiques, nous avons étendu l'analyse au domaine du discours, nous avons implémenté des fonctions d'analyse syntaxique basée sur les relations asymétriques (l'accord et de la c-commande asymétriques) et le traitement des catégories vides (selon Di Sciullo 2006).

4.1 Le module de résolution anaphorique pronominale aLAD

aLAD est un module de résolution d'anaphores pronominales qui étend la couverture des applications computationnelles de la Théorie de l'asymétrie en ajoutant à l'analyseur LAD la fonction de résolution des anaphores pronominales.

Le module de résolution de l'anaphore pronominale intra- et interpropositionelle intègre un mécanisme de filtrage morphosyntaxique incluant une implémentation de l'accord asymétrique (Di Sciullo 2003, 2005), des spécifications sur le domaine de liage - l'information syntaxique utilisée provenant du parseur LAD (Di Sciullo et al. 2006), parseur qui fournit une analyse des relations asymétriques sous-jacentes aux expressions linguistiques de l'anglais.

Les pronoms sont des expressions linguistiques sans référence indépendante. Leur possibilité d'être liés à l'intérieur d'une proposition est gérée par les principes de la théorie du liage de Chomsky (1981), reprises en (62) pour faciliter la lecture. Le liage intrapropositionnel est soumis à la c-commande asymétrique ; la définition de la c-commande reprise en (63), et la définition de la c-commande asymétrique en (64):

(62) A. Une anaphore doit être liée dans son domaine de liage

B. Un pronom ne peut pas être lié dans son domaine de liage

C. Une expression référentielle doit être libre

(63) α c-commande β, ssi:

tout γ qui domine α domine β et

α ne domines pas β

(64) α c-commande asymétriquement β ssi :

α c-commande β et

β ne c-commande pas α

L'analyseur LAD peut offrir les ressources formelles et sémantiques nécessaires pour identifier pour chaque pronom son domaine de liage et son antécédent.

Conformément à (62B), les pronoms sont libres dans leur domaine de liage propositionnel. Cependant, les pronoms n'ont pas de référence indépendante, et doivent chercher un antécédent au-delà des limites du domaine du liage propositionnel, soit dans le discours. Di Sciullo (2006) postule (65) la condition d'interface pour l'anaphore pronominale, DD-Linking, et la relation d'accord, définie en (66), qui s'applique dans le domaine discursif:

(65) DD-Linking (Discourse Domain-Linking) - A pronominal must be linked in its DD.

(66) Agree (φ1, φ2) - Given two sets of features φ1 and φ2, Agree (φ1, φ2) applies iff φ1 properly includes φ2.

Étant donné que l'opération utilisée, *Link* (présentée en (55), est sujette à la vérification de l'accord asymétrique, la résolution de l'anaphore pronominale revient à trouver l'antécédent le plus proche du pronom tel que les traits de l'antécédent et ceux du pronom sont en relation de sous-ensemble propre[24].

Par exemple en (67), *John* [D, 3pers, masc, sing, +ani][25] est un antécédent possible pour le pronom *him* [D, 3pers, masc, sing, +ani, -w], car les traits de ces deux constituants sont en relation de sous-ensemble propre. L'ensemble de traits du pronom *him* inclut le trait [-w], qui est spécifique aux pronoms personnels, tel que proposé par Di Sciullo (2005c), ce qui n'est pas le cas des expressions référentielles telles que *John*.

---

[24] Sous-ensemble : pour les deux ensembles A et B, si tous les membres de A appartient aussi à B, alors A est un sous-ensemble de B. A est un sous-ensemble propre de B si et seulement si A est un sous-ensemble de B et A n'est pas égal à B.

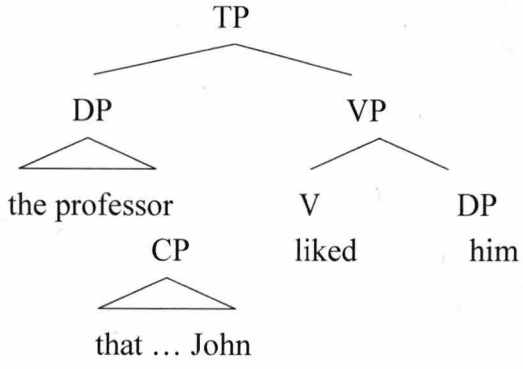
[25] Les traits sémantiques utilisés sont : référence indépendante [±ir], animé [±ani], partie-tout [±w].

Par ailleurs, *the professor* n'est pas un antécédent possible pour *him* puisque *the professor* c-commande asymétriquement *him*, voir (68), et selon la condition B de la Theorie du Liage, un pronom ne peut être lié dans son domaine de liage.

(67) The professor that interviewed John$_i$ liked him$_i$.
'Le professeur qui a interviewé John l'aime.'

[D]	[D]	[D]
[3pers]	[3pers]	[3pers]
[masc]	[masc]	[masc]
[sing]	[sing]	[sing]
[+ani]	[+ani]	[+ani]
		[-w]

(68)



4.2 Description de l'algorithme de résolution de l'anaphore pronominale

Notre algorithme de résolution des anaphores pronominales est implémenté en Python et a été conçu pour s'intégrer fonctionnellement dans l'interface de l'analyseur *LAD*. Le système ainsi obtenu, *aLAD*, a été conçu pour être autonome et

peut parcourir toutes les phases de l'analyse sans que l'intervention d'un tiers (humain ou logiciel) soit nécessaire. *aLAD* traite les anaphores pronominales autant dans le domaine de la proposition que dans le domaine interphrastique.

Les opérations de l'algorithme de résolution d'*aLAD* sont les suivantes :

```
soit T le texte constitué des propositions Prop et des mots M
soit EAPR l'ensemble d'anaphores pronominales à résoudre
soit F la fenêtre d'attention
soit EAPOS l'ensemble d'antécédents possibles
soit d le domaine de recherche de l'antécédent
pour chaque Prop en T
    pour chaque mot M en Prop
        assigner indices M_prop,m
        si (M est un pronom personnel, possessif ou réfléchi de la
            troisième personne) et
           (M n'est pas non référentiel)
           alors
            ajouter M_prop,m à EAPR
        finsi
    finpour
finpour
pour chaque Pron_prop,m en EAPR
    si (Pron est une anaphore/pronom réfléchi) alors
        définir d_Pron comme étant T_prop
        pour chaque phrase nominale DP en d_Pron
            ajouter DP à EAPOS_Pron
        finpour
        pour chaque antecedent possible DP en EAPOS_Pron
            vérifier les conditions d'anaphoricité DP - Pron
        finpour
    sinon
        définir d_Pron comme étant le fragment de texte entre T(0) et
          T(prop)
        définir F_Pron comme étant le fragment de texte entre d_Pron(prop-9)
          et d_Pron(prop)
        pour chaque phrase nominale DP en F_Pron
            si (DP en F_Pron(prop)) alors
                si (DP ne c-commande pas asymétriquement Pron) alors
                    ajouter DP à EAPOS_Pron
                finsi
            sinon
                ajouter DP à EAPOS_Pron
            finsi
        finpour
```

```
        pour chaque antecedent possible DP en EAPOS_Pron
            vérifier les autres conditions d'anaphoricité DP - Pron
        finpour
    finsi
finpour
```

4.2.1 Opérations pré-résolution

Avant l'étape de résolution, le texte brut doit être traité par l'analyseur LAD. À cette fin, le texte peut être saisi dans le champ d'entrée d'une interface évoluée de LAD, Figure 13, qui intègre la fonction de résolution, l'*aLAD*. Le texte en entrée peut également être chargé à partir d'un fichier pré-formaté, une proposition par ligne.

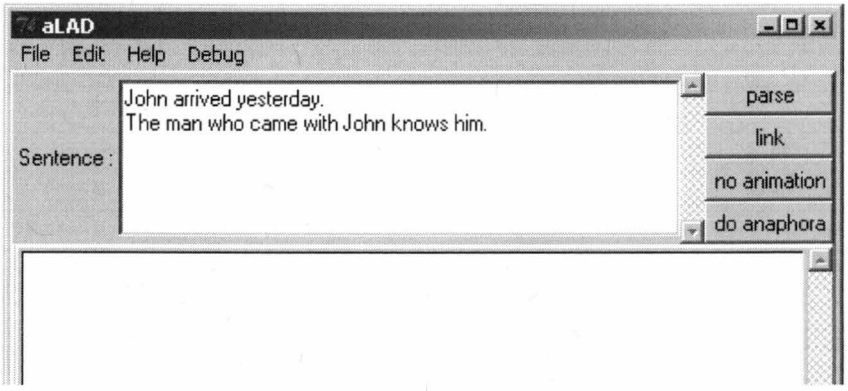


Figure 13 - L'interface graphique en aLAD

a. L'analyse syntaxique

Le parseur syntaxique *LAD* analyse le texte en entrée et la représentation obtenue à la sortie fournit l'information linguistique riche nécessaire à l'étape de résolution : les traits φ et les traits sémantiques de chaque entrée lexicale analysée, l'arbre syntaxique contenant de l'information détaillé sur les relations de c-commande, ainsi que d'accord entre constituants sont passés au module de résolution. Voici, en Figure 14, un exemple d'analyse syntaxique en *aLAD*.

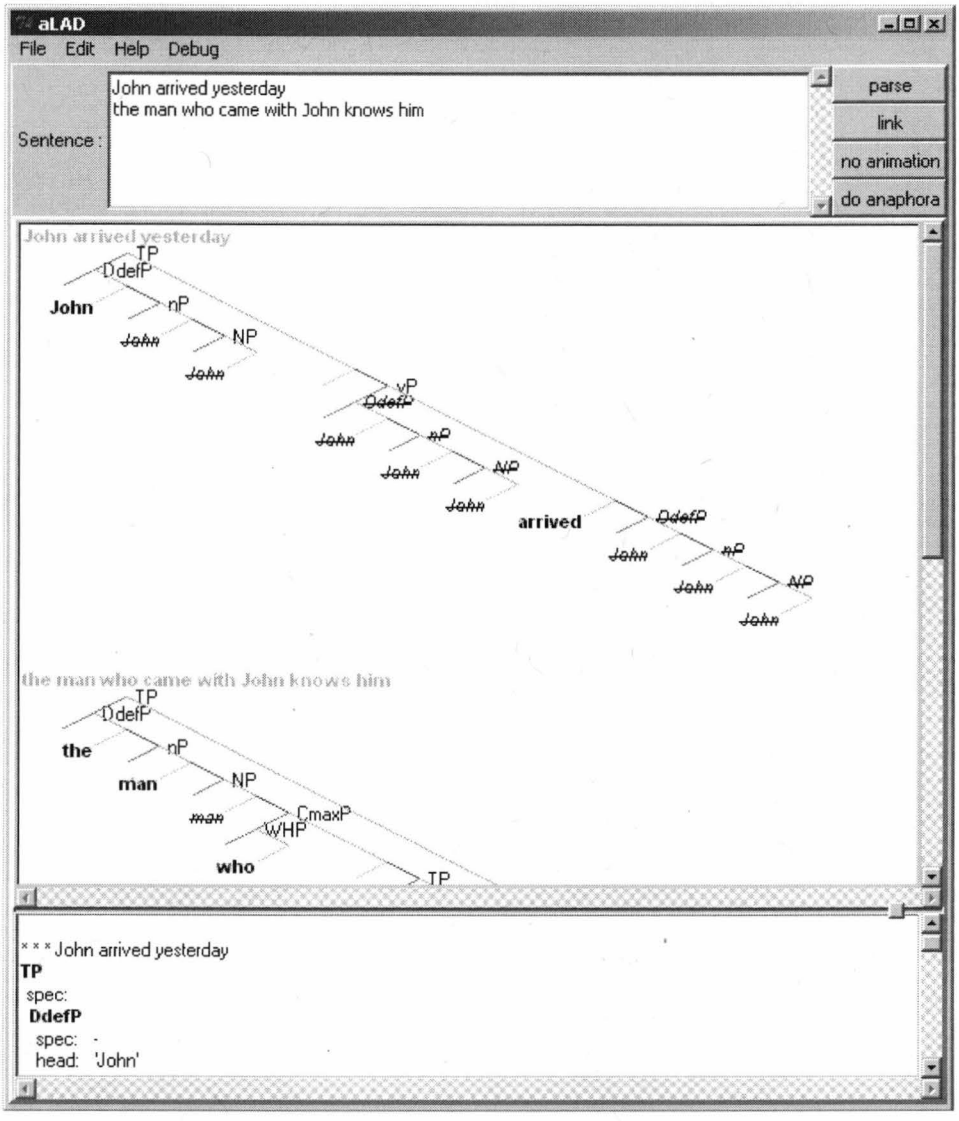


Figure 14 - Exemple d'analyse syntaxique en *aLAD*

b. Le modèle du discours

Le modèle du discours utilisé est réduit au minimum nécessaire à la démonstration. L'étendue du discours est établie par le texte à l'entrée du parseur *LAD* qui, dans l'état présent de l'application, a été testé sur des textes allant d'une seule à plus de 1000 propositions.

Il existe une relation directe entre la distance qui sépare l'anaphore de son antécédent et le degré de difficulté de la tâche de résolution : plus la distance est grande, plus le calcul est coûteux. Plusieurs recherches faites sur des corpus contenant des anaphores pronominales montrent que, généralement, l'antécédent d'une anaphore pronominale se trouve à une distance de maximum cinq propositions.[26] Nous avons implémenté une modélisation minimale des contraintes d'attention et de mémoire sous la forme d'une fenêtre F limitant l'espace de recherche à un maximum de 10 propositions. Ainsi, la fenêtre d'attention au moment du traitement du pronom *Pron* est F_{Pron} et correspond au fragment du texte T qui commence 10 propositions avant la proposition qui contient *Pron* et finit par la proposition courante T_{prop} : $F_{Pron} = T_{prop-9} -> T_{prop}$

Chaque fois qu'une anaphore est analysée dans une nouvelle proposition, un rang d'attention de 10 est assigné à chaque antécédent possible figurant dans la même proposition. Les antécédents possibles figurant dans la proposition qui précède la proposition courante seront assignés un rang d'attention de 9, et ainsi de suite jusqu'à la dernière proposition de F_{Pron} qui aura des antécédents possibles avec un rang d'attention de 1. Une décision est prise dans le processus de résolution de l'anaphore courante après que les 10 propositions de F_{Pron} sont parcourues, auquel moment F_{Pron} est déplacée vers l'avant dans le texte jusqu'à la proposition contenant l'anaphore suivante. Suite à ce repositionnement, le système reprend le processus d'assignation de rangs d'attention aux antécédents possibles présents dans la nouvelle F_{Pron}.

Un mécanisme d'indexation est nécessaire pour permettre l'identification des différents fragments du texte. Un algorithme de parcours en largeur (breadth-first) est utilisé pour produire un index hiérarchisé à deux niveaux : propositions et mots. Chaque mot M est reçoit des indices *prop* et *m* ($M_{prop,m}$).

---

[26] Charniak (1972), Klappholz & Lockman (1975), Pérez (1994) et autres trouvent la majorité des antécédents dans les cinq propositions précédant l'anaphore pronominale. Une distance record de 13 propositions est mentionnée par Hobbs (1978).

Nous excluons le traitement de la référence pronominale intertextuelle dans cette étude.

4.2.2 Création de l'ensemble d'anaphores pronominales à résoudre

L'algorithme parcourt le texte de gauche à droite et extrait les pronoms personnels de la troisième personne du singulier et du pluriel. De l'ensemble ainsi constitué, *EAPR*, sont éliminés les pronoms non référentiels.

Le pronom personnel de la troisième personne du singulier, *it*, est le pronom personnel le plus souvent utilisé en anglais.[27] Cependant, son usage n'est pas toujours référentiel, mais souvent explétif, comme en (69). En fait, Lappin et Leass (1994) constatent que 8% des pronoms traités par leur système sont non référentiels. Il est important d'identifier et d'éliminer d'*EAPR* les occurrences non-référentielles (ou impersonnelles) de *it* avant de déclencher la procédure de résolution pour des raisons de performance et d'économie du système.

(69) *It* was suggested that a report be created.

Des heuristiques pour l'élimination des pronoms non référentiels ont été mises en place. Ces heuristiques utilisent des patrons syntaxiques et des restrictions lexicales définies par Paice et Husk (1987) et Lappin et Leass (1994).

Nous avons utilisé deux classes lexicales fermés. STATUS – formée des adjectives du type : *advisable, certain, convenient, desirable*, etc., et VERB formée de verbes du type: *anticipated, assumed, believed*, etc. Une troisième classe ouverte de verbes s'ajoute pour créer des patrons syntaxiques en vue d'identifier et d'éliminer les pronoms non référentiels. Un des patrons utilisés est présentée en (70) avec un

---

[27] Nous avons fait une recherche dans le British National Corpus qui montre que *it* est le pronom avec l'utilisation la plus fréquente (10875 occurrences), le septième mot le plus souvent utilisé dans le BNC écrit-parlé.

exemple de proposition contenant un pronom non référentiel en (71). Une liste des classes utilisées est présentée dans l'Annexe 5.

(70) it VERB STATUS to TASK

(71) It is critical to recognize stroke symptoms and act quickly.

4.2.3 Identification du domaine de recherche de l'antécédent

Les anaphores et les pronoms ne sont pas autonomes du point de vue de la référence, mais ils acquièrent leur référence via des éléments « référentiellement » indépendants auxquels ils sont liés. La notion de liage est définie comme la relation entre deux nœuds *a* et *b* dans laquelle *a* lie *b* si et seulement si *a* et *b* sont co-indexés et *a* c-commande asymétriquement *b*. Le domaine de liage d'un pronom *Pron* est son domaine local propositionnel TP_{Pron}, qui est la proposition courante T_{prop}.

La Théorie du liage (62) spécifie les conditions qui régissent le liage pour les anaphores, les pronoms et les expressions référentielles. Une conséquence qui découle du principe B de la théorie du liage est qu'un pronom *Pron* peut néanmoins être lié dans son domaine local propositionnel TP_{Pron}, mais seulement par un élément qui ne le c-commande pas. Appelons $A(TP_{Pron})$ le sous-ensemble d'éléments TP_{Pron} qui ne c-commande pas *Pron*. Le sous-ensemble $A(TP_{Pron})$ sera ensuite ajouté à l'ensemble d'antécédents possibles extraits du domaine d_{Pron} pour créer l'ensemble complet des antécédents possible de *Pron*.)

Pour définir l'étendue du domaine *d* (72) dans lequel le pronom *Pron* peut trouver un antécédent, nous allons utiliser la modélisation du processus d'attention décrite ci-haut. Conséquemment, le pronom *Pron* devra chercher sa référence dans le domaine obtenu par l'élimination de la fenêtre d'attention du domaine dans lequel *Pron* doit être libre:

(72) $d_{Pron} = F_{Pron} - TP_{Pron}$

4.2.4 Création de l'ensemble d'antécédents possibles pour chaque élément

Avant de procéder à l'étape finale de la résolution de l'anaphore pronominale, il est nécessaire de créer l'ensemble complet des antécédents possibles d'un pronom, *EAPOS*, pour chaque élément *Pron* d'*EAPR*, selon son domaine de recherche. L'ensemble d'antécédents possibles de *Pron*, $EAPOS_{Pron}$ (73), se définit comme la totalité de phrases nominales appartenant au domaine d_{Pron} auxquelles s'ajoutent les éléments qui ne c-commande pas *Pron* dans son domaine local:

(73) $EAPOS_{Pron} = DP(d_{Pron}) + A(TP_p)$

4.2.5 Vérification des conditions d'anaphoricité

Pour toutes les paires formées de *Pron* et chaque élément de $EAPOS_{Pron}$ les conditions d'anaphoricité sont vérifiées en commençant avec les antécédents les plus proches.

a. La vérification de l'accord morphologique

Une comparaison des traits Φ des deux éléments de chaque paire permet d'éliminer très tôt un antécédent dont les traits ne sont pas en relation de sous-ensemble propre avec les traits de l'anaphore. En (74) *Paul* et *the waiter* sont éliminés à cause de l'inaccord causé par les traits [masc] et [+ani].

(74) Paul gave the tip to the waiter who deserved it.
'Paul a donné le pourboire au serveur qui le mérité.'

[D]	[D]	[D]	[D]
[3pers]	[3pers]	[3pers]	[3pers]
[masc]		[masc]	
[sing]	[sing]	[sing]	[sing]
[+ani]	[-ani]	[+ani]	[-ani]
			[-w]

b. La résolution dans le domaine de liage

Selon les conditions nécessaires à la bonne formation des relations anaphoriques dans le domaine de liage, Tableau 5, une anaphore doit trouver son antécédent dans la même proposition et la relation sera bien formée seulement si les conditions de c-commande asymétrique et d'accord asymétrique sont remplies simultanément.

	c-commande asymétrique	accord asymétrique
anaphore (réfléchi)	obligatoire	obligatoire
anaphore pronominale	anti c-commande	obligatoire

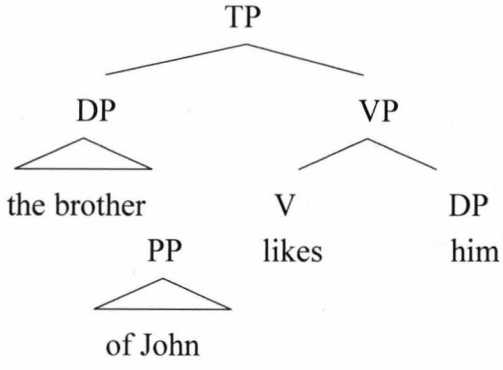
Tableau 5 - Conditions de bonne formation des relations anaphoriques dans le domaine propositionnel

Une anaphore pronominale peut trouver son antécédent dans la même phrase seulement si les conditions de anti c-commande (ou absence de c-commande) et d'accord asymétrique sont remplies simultanément. Par exemple, en (75) *a* et *b*, la condition de l'accord asymétrique est remplie pour les deux antécédents possibles, mais la condition d'anti c-commande ne l'est pas, *brother* c-commandant asymétriquement *him*. En (75c) aucune des deux conditions de bonne formation des relations anaphoriques dans le domaine de liage n'est remplie dans le cas de *Mary* : il n'y a pas de relation de sous-ensemble propre entre les traits de l'anaphore et les traits de *John* et *Mary* c-commande asymétriquement *him*. En (75d) *the man* c-commande asymétriquement *him*. En (75e) le sujet de phrase enchâssée, *he*, c-commande asymétriquement l'objet de la phrase enchâssée, *a genius* (mais il coréfère avec le sujet de la phrase matrice, la condition de l'anti c-commande s'appliquant seulement à l'intérieur de la proposition.

(75) a. The brother of John$_i$ likes him$_i$.

'Le frère de John l'aime.'

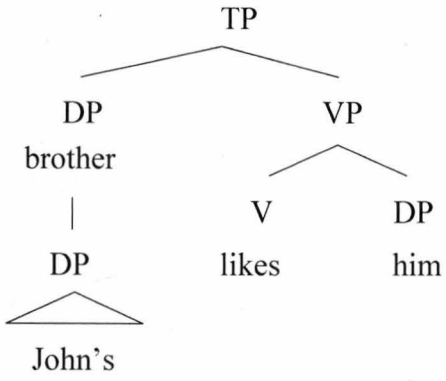
[D]	[D]	[D]
[3pers]	[3pers]	[3pers]
[masc]	[masc]	[masc]
[sing]	[sing]	[sing]
[+ani]	[+ani]	[+ani]
		[-w]



b. John$_i$'s brother likes him$_i$.

'Le frère de John l'aime.'

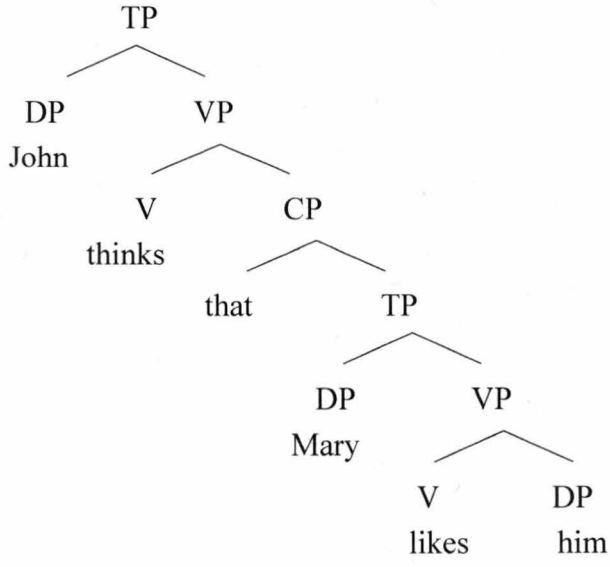
[D]	[D]	[D]
[3pers]	[3pers]	[3pers]
[masc]	[masc]	[masc]
[sing]	[sing]	[sing]
[+ani]	[+ani]	[+ani]
		[-w]



c. $John_i$ thinks that Mary likes him_i.

'John pense que Mary l'aime.'

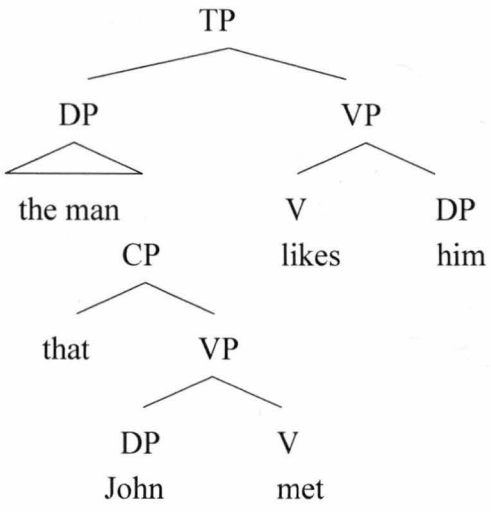
[D]	[D]	[D]
[3pers]	[3pers]	[3pers]
[masc]	[fem]	[masc]
[sing]	[sing]	[sing]
[+ani]	[+ani]	[+ani]
		[-w]



d. The man that $John_i$ met likes him_i.

'L'homme que John a rencontré l'aime.'

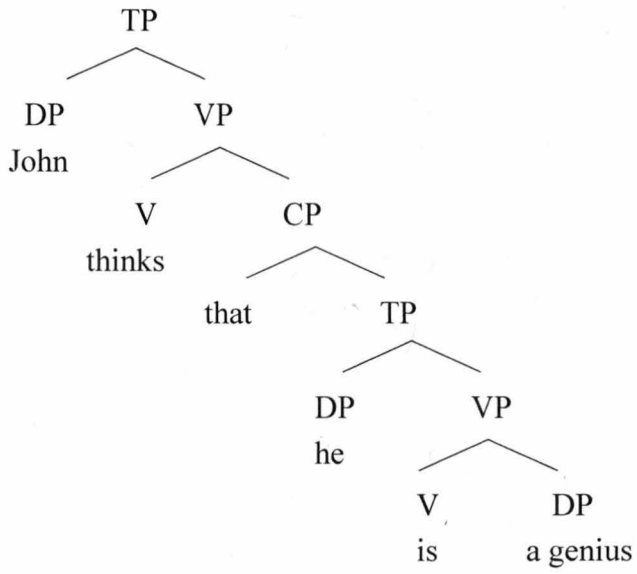
[D]	[D]	[D]
[3pers]	[3pers]	[3pers]
[masc]	[masc]	[masc]
[sing]	[sing]	[sing]
[+ani]	[+ani]	[+ani]
		[-w]



e. $John_i$ thinks that he_i is a genius.

'John pense qu'il est un génie.'

[D]	[D]	[D]
[3pers]	[3pers]	[3pers]
[masc]	[masc]	
[sing]	[sing]	[sing]
[+ani]	[+ani]	[+ani]
	[-w]	



Une relation anaphorique entre pronom et antécédent est exclue si les conditions de c-commande asymétrique et d'accord asymétrique ne sont pas satisfaites. Dans le domaine propositionnel (76) un pronom ne peut prendre comme antécédent un élément qui le c-commande asymétriquement. En (77) la relation d'accord asymétrique n'est pas satisfaite.

(76) a. #John$_i$ likes him$_i$.

'John l'aime.'

b. #He$_i$ likes John$_i$.

'Il aime John.'

(77) The man that Mary$_i$ met likes him$_j$.

'L'homme que Mary a rencontré l'aime.'

Si l'une des deux conditions dans le Tableau 5 n'est pas remplie, soit la c-commande asymétrique et l'accord asymétrique entre antécédent et anaphore, l'anaphore doit trouver son antécédent à l'extérieur de son domaine local.

c. La résolution dans le domaine du discours

Dans le domaine du discours, Tableau 6, une anaphore pronominale doit trouver son antécédent à l'extérieur de son domaine local, soit la proposition, et une relation anaphorique sera possible si la condition de d'accord asymétrique est satisfaite. La condition de c-commande asymétrique ne s'applique pas dans le domaine du discours, soit entre les propositions.

	c-commande asymétrique	accord asymétrique
anaphore (réfléchi)	-	-
anaphore pronominale	-	obligatoire

Tableau 6 - Conditions de bonne formation des relations anaphoriques dans le domaine du discours

Les relations de c-commande asymétrique et par conséquent la condition d'anti c-commande ne sont pas propres au domaine du discours et ne s'appliquent pas dans la formation des relations anaphoriques interpropositionnelles. Dans les exemples en (78) *a*, *b* et *c* la réalisation de l'accord asymétrique autorise les relations anaphoriques indiquées. Si une relation d'accord asymétrique ne se réalise pas, la recherche de l'antécédent doit poursuivre plus loin dans le discours (79).

(78) a. John$_i$ met Mary$_j$. She$_j$ really likes him$_i$.

'John a rencontré Mary. Elle l'aime beaucoup.'

[D]	[D]	[D]	[D]
[3pers]	[3pers]	[3pers]	[3pers]
[masc]	[fem]	[fem]	[masc]
[sing]	[sing]	[sing]	[sing]
[+ani]	[+ani]	[+ani]	[+ani]
		[-w]	[-w]

b. John$_i$ met the Simpsons$_j$. He$_i$ really liked them$_j$.

'John a rencontré les Simpsons. Il les a beaucoup aimé.'

[D]	[D]	[D]	[D]
[3pers]	[3pers]	[3pers]	[3pers]
[masc]		[masc]	
[sing]	[pl]	[sing]	[pl]
[+ani]	[+ani]	[+ani]	[+ani]
		[-w]	[-w]

c. French$_i$ is a Romance language. Mary knows it$_i$ well.

'Le francais est une langue romane. Mary la connait bien.'

[D]	[D]	[D]	[D]
[3pers]	[3pers]	[3pers]	[3pers]
		[fem]	
[sing]	[sing]	[sing]	[sing]
[-ani]	[-ani]	[+ani]	[-ani]
			[-w]

(79) John met Paul. She couldn't come.

'John a rencontré Paul. Elle n'a pas pu venir.'

[D]	[D]	[D]
[3pers]	[3pers]	[3pers]
[masc]	[masc]	[fem]
[sing]	[sing]	[sing]
[+ani]	[+ani]	[+ani]

d. aLAD - Exemples d'analyse

Nous reprenons ici un exemple analysé par LAD à la Figure 14 (*The man who came with John knows him*) afin de détailler la résolution du pronom anaphorique *him*. Le texte en entrée est composé d'une proposition qui contient le pronom personnel *him,* dont l'antécédent est contenu dans une proposition enchâssée.

Le traitement débute avec la phase pré-résolution - l'analyse syntaxique faite par l'analyseur LAD. Ensuite, l'aLAD extrait les pronoms personnels de la troisième personne. *EAPR* est constitué dans ce cas d'un seul pronom, *him*. L'aLAD identifie les traits du pronom et cherche un antécédent possible dans le domaine de la proposition qui le contient. Étant donné qu'il s'agit d'un pronom, et non d'un réfléchi, l'aLAD cherchera un antécédent possible qui ne le c-commande pas asymétriquement. De plus, l'aLAD cherche à identifier un antécédent dont les traits sont en relation de sous-ensemble avec les traits du pronom, tel que requis par la condition d'accord asymétrique.

L'aLAD trouve effectivement que le DP *John* se trouve dans le domaine requis et que les traits de du DP *John* {[D], [3pers], [masc], [sing], [+ani]} et du pronom *him*

him {[D], [3pers], [masc], [sing], [+ani], [-w]} sont en relation de sous-ensemble propre. L'analyse se conclue par l'assignation du rôle d'antécédent au DP *John*. Les mêmes structures et le même processus décisionnel sont en œuvre dans l'analyse de l'exemple en Annexe 2.

Considérons maintenant le traitement que l'aLAD effectue dans le domaine du discours, dans l'exemple : *Dan Morgan told himself he would forget Ann Turner. He was well rid of her.* (Voir la trace en Annexe 2)

Après l'analyse syntaxique, une fois l'ensemble de pronoms identifiés, l'aLAD commence avec le premier pronom dans cet ensemble et procède de la même manière pour trouver son antécédent.

La première phrase contient une proposition principale (matrice) et une proposition enchâssée. Le pronom réfléchi *himself* se trouve dans le phrase matrice et doit trouver un antécédent dans cette proposition, conformément au principe A de la Théorie du liage (62). En effet, il y existe dans cette proposition un DP, *Dan Morgan*, avec lequel il est en relation d'accord asymétrique. En conséquence, le rôle d'antécédent est assigné au DP *Dan Morgan*.

L'analyse se poursuit avec le pronom *he* qui est en position sujet de la phrase enchâssée, et donc libre dans son domaine de liage, conformément au principe B de la théorie du liage (62). L'aLAD cherche dans la proposition matrice un antécédent possible pour ce pronom, et trouve le DP *Dan Morgan* qui satisfait la condition d'accord asymétrique pour l'anaphore, mais non la condition de c-commande asymétrique. En conséquence, le rôle d'antécédent est assigné au DP *Dan Morgan*.

Ensuite l'aLAD passe à l'analyse de la seconde phrase du discours en ne trouve aucun antécédent pour les pronoms *He* et *her* dans ce domaine local, conformément au principe B de la théorie du liage (62). L'analyse est étendue à la phrase précédente dans la fenêtre d'attention.

L'analyse continue avec la recherche d'un antécédent pour le pronom *He*. Dans la phrase initiale aLAD trouve un seul DP, avec lequel pronom *He* {[D], [3pers], [masc], [sing], [+ani], [-w]} est en relation d'accord asymétrique, *Dan Morgan* {[D], [3pers], [masc], [sing], [+ani]}. En conséquence, le rôle d'antécédent du pronom *He* est assigné au DP *Dan Morgan*.

Le processus de résolution continue pour le pronom *her*. Dans la phrase initiale aLAD trouve un seul DP, avec lequel pronom *her* {[D], [3pers], [fem], [sing], [+ani], [-w]} est en relation d'accord asymétrique, soit le DP *Ann Morgan* {[D], [3pers], [fem], [sing], [+ani]}. En conséquence, le rôle d'antécédent du pronom *her* est assigné au DP *Ann Turner*.

L'aLAD reconnait les relations d'accord asymétrique entre antécédent et anaphore et identifiera *Dan Morgan* comme un antécédent possible du pronom *He* mais non pas du pronom *her*. La raison étant que *Dan Morgan* {[D], [3pers], [masc], [sing], [+ani]} et *her* {[D], [3pers], [fem], [sing], [+ani], [-w]} ne sont pas en relation de sous-ensemble propre : ils ne partagent pas l'ensemble de leurs traits *phi*.

Nous verrons au Chapitre 5 qu'un système de résolution anaphorique qui n'est pas basé sur des relations asymétriques, comme c'est le cas du prototype aLAD, donne lieu à un traitement insatisfaisant des anaphores pronominales discursives et phrastiques.

4.2.4 Types de stratégies de résolution

Le processus de résolution d'une anaphore est clos au moment où la décision a été prise quant à l'identité de son antécédent. Le moment de prendre cette décision survient tard dans l'analyse, soit après que toute l'information requise ait été obtenue. Cependant, l'humain procède de manière incrémentielle à la construction de la structure référentielle d'un texte, sans qu'une analyse morphosyntaxique de ce texte soit effectuée au préalable. Il résulte des études en suivi des mouvements oculaires

(eye tracking) que l'identification de l'antécédent d'un pronom qui se trouve dans la même proposition que ce dernier se fait avant que le lecteur ne passe à la phrase suivante (Carminati 2002, Costa et al. 2011). Toutefois, nous n'avons pas trouvé dans la littérature de système qui propose une stratégie de résolution adaptée à cette réalité cognitive. Le module de résolution des anaphores pronominales que nous avons construit ajoute à la résolution classique, ou tardive, une stratégie de résolution *à la volée* (on the fly).

Lors de l'utilisation de la stratégie de résolution tardive il est nécessaire qu'une analyse complète du texte entier soit faite avant de procéder à la recherche des antécédents des anaphores pronominales. Avec la stratégie à la volée, la résolution se fait graduellement, en deux étapes. Le module *aLAD* interagit dynamiquement avec l'analyseur *LAD* et, dans une première phase, la résolution est effectuée au niveau de la phrase pour ensuite reléguer seulement les anaphores non résolues à un moment ultérieur où l'analyse syntaxique de l'ensemble du texte est complète.

À la fin de l'étape d'analyse syntaxique la stratégie de résolution tardive s'applique à l'ensemble d'anaphores non résolues.

Dans ce chapitre nous avons décrit *aLAD*, le module de résolution d'anaphores pronominales basé sur la théorie de l'asymétrie et qui utilise des connaissances linguistiques riches. Le chapitre suivant est consacré à une analyse comparative de la performance de l'*aLAD* et de la performance de deux autres systèmes de résolution de l'anaphore pronominale.

CHAPITRE V

TESTS ET ÉVALUATION DE LA PERFORMANCE

Dans ce chapitre nous présenterons les tests que nous avons effectués sur le corpus choisis et les résultats que nous avons obtenus. Nous montrerons aussi quels sont les effets de l'implémentation de la stratégie de résolution à la volée.

5.1 Introduction

Il est bien connu que l'évaluation des résultats dans la résolution de l'anaphore est très difficile. Une évaluation comparative crédible de la performance des systèmes de résolution requiert une base commune sur laquelle les tests d'évaluation peuvent être appliqués. Or les applications dans ce domaine se ressemblent rarement.

La première différence est due à la nature du texte en entrée. Le besoin de traiter des textes spécialisés, comme les textes de biomédecine, a mené à l'implémentation des systèmes adaptés qui répondent bien aux exigences du domaine (Chen et al. 2008, Huang 2010, etc.). Ces sont des systèmes qui se différencient des systèmes voués au traitement des textes aux sujets communs, non spécifiques.

Par la suite il y a le choix du type et du volume des connaissances avec lesquelles les systèmes opèrent. Il existe systèmes qui fonctionnent sans analyseur morphosyntaxique (Kennedy et Boguraev 1996) ainsi que des systèmes qui utilisent des connaissances très riches de nature linguistique et du monde réel (Klappholz et Lockman 1975). Nous soulignions aussi la différence qui existe entre, d'une part, les systèmes de résolution automatiques qui sont autosuffisants, pouvant assurer de manière autonome toutes les phases du traitement, et d'une autre part, celles qui prennent en entrée le texte pré-analysé.

Ce problème est connu et il a été abordé par plusieurs chercheurs (Vilain 1995, Mitkov et Barbu 2001, Kabadjov 2007, etc.) Des propositions ont été faites pour l'adoption d'une méthodologie unifiée dans le but de rendre le processus d'évaluation plus fiable. Entre autres, des nouvelles mesures de performance ont été ajoutées aux mesures classiques (précision et rappel). Ces nouvelles, les classes d'équivalence et le taux de succès, permettent l'évaluation comparative entre systèmes automatiques de résolution et les algorithmes dépendants de l'apport d'un tiers logiciel.

Néanmoins, il reste que pour faire une évaluation comparative de deux ou plusieurs systèmes de résolution anaphorique de grands efforts doivent être déployés, efforts qui vont des fois jusqu'à la reprogrammation entière des systèmes.

Nous nous sommes donné comme tâche d'identifier des propriétés fondamentales de la faculté du langage, telles que la c-commande asymétrique dans le cas du liage et l'accord asymétrique dans le cas de la coréférence, que les humains utilisent lors du processus de résolution des anaphores pronominales et de montrer que l'intégration de ces propriétés dans les systèmes automatiques de résolution peut augmenter la performance de ces systèmes.

5.2 Outils employés pour la comparaison

Nous nous proposons de comparer aLAD à des systèmes qui ne lui sont pas trop éloignés du point de vue de la méthode de traitement. Pour effectuer une évaluation comparative des résultats de notre travail nous choisissons de systèmes qui correspondent aux critères suivants :

- systèmes qui utilisent des connaissances syntaxiques, même si moins riches, et non pas basés sur des étiquettes morpho-syntaxiques et des heuristiques (p. ex. Kennedy et Boguraev 1996)
- systèmes automatiques (qui font toutes les étapes du traitement sans qu'une intervention humaine soit nécessaire) ; il en découle que :
- le texte en entrée ne doit pas être traité manuellement ;
- enfin, une implémentation informatique doit être accessible pour en faire l'évaluation.

Nous avons choisi parmi les systèmes présentés dans le Tableau 4 deux systèmes automatiques qui se démarquent par leurs performances du reste du peloton des systèmes à connaissances pauvres : le MARS (Mitkov 2002) et le RAP (Lappin & Leass 1994).

RAP - Resolution of Anaphora Procedure

RAP affichait une précision de 57,9% lors de la MUC-6 en 1995, qui montait, après les réglages fins, à 85% pour des textes provenant des manuels informatiques. (Lappin & Leass 1994)

RAP est un algorithme conçu pour la résolution automatique des anaphores intra- et interpropositionnelles. Dans son implantation originelle il utilise le parseur de McCord qui est basé sur une grammaire de dépendances. RAP est constitué de plusieurs composantes qui s'appliquent à la sortie du parseur :

- un filtre syntaxique qui a le rôle d'éliminer la coréférence entre un pronom P et une phrase nominale N si une des six conditions spécifiées ci-dessous est remplie (voir Figure 15) ;
- un filtre morphologique qui a le rôle d'éliminer la coréférence entre un pronom et une phrase nominale dans l'absence de l'accord en traits Φ ;
- une procédure pour éliminer les pronoms explétifs ;
- une procédure pour identification des antécédents possibles des pronoms réfléchis à l'intérieur de la même proposition ;
- une procédure qui assigne une valeur de proéminence aux antécédents possibles selon le rôle grammatical, la fréquence de mention, la proximité de l'anaphore, etc. ;
- une procédure pour l'identification de chaines référentielles.

1. P and N have incompatible agreement features.
2. P is in the argument domain of N.
3. P is in the adjunct domain of N.
4. P is an argument of a head H, N is not a pronoun, and N is contained in H.
5. P is in the NP domain of N.
6. P is a determiner of a noun Q, and N is contained in Q.

Figure 15 – Conditions du filtre syntaxique en RAP (Lappin et Leass 1994, p. 537)

L'algorithme crée d'abord une liste de phrases nominales présentes dans la proposition courante. Il procède ensuite à la résolution de pronoms présents dans cette proposition et continue avec la résolution des pronoms suivant dans l'ordre d'apparition dans le texte.

Quelques mentions que nous devons faire sur RAP :

- même si les grammaires basées sur les dépendances produisent une analyse en termes de relations tête-argument et tête-adjoint, elles ne récupèrent pas les relations d'asymétrie qui s'établissent entre, par exemple, le prédicat verbal et

son argument interne, d'une part, et, d'une autre, avec son argument externe, tel que présenté en Figure 1. Ce type d'analyse conduit à des erreurs de résolution en contexte d'anti-c-commande comme celui en (45a), dont la structure est présentée en (46) et pour lequel RAP ne trouve pas d'antécédent (Annexe 1, ex. 6).

- la configuration courante de RAP a été optimisée sur un corpus spécialisé formé de manuels pour ordinateurs. Différentes configurations des valeurs de proéminence ont été itérés afin d'optimiser les performances de RAP. Les bons résultats obtenus (précision de 85%) sont en partie une conséquence du fait que l'évaluation a été faite sur le même type de corpus que celui sur lequel l'algorithme a été entrainé. La précision de la résolution sur un corpus non spécialisé tombe en fait à 58% (Qiu 2004).
- dans l'évaluation comparative nous avons utilisé RAP dans son implémentation Java[28] (Qiu 2004) qui utilise un parseur statistique (Charniak 2000). L'information fournie à RAP par ce parseur est complétée par des heuristiques appliquées à l'arbre syntaxique pour inférer les relations tête-argument et tête-adjoint afin d'offrir à RAP une information équivalente à celle fournie par le parseur McCord.

MARS - Mitkov's Anaphora Résolution System

MARS, utilisant l'analyseur FDG (Tapanainen and Jarvinen 1997), avait atteint un seuil de précision de 62,4% en mode automatique, mais pouvait atteindre 89,7% dans des contextes spécialisés. (Mitkov 2002)

---

[28] Nous utilisons l'implémentation Java de l'algorithme originel de Lappin et Leass disponible grâce à Long Qiu et accessible à l'adresse suivante: http://www-appn.comp.nus.edu.sg/~rpnlpir/cgi-bin/JavaRAP/JavaRAPdemo.cgi

L'approche de Mitkov est de concevoir un système automatique de résolution des anaphores qui éviterait les obstacles éventuels résultant de l'utilisation des représentations syntaxiques et sémantiques riches. MARS est une optimisation d'une variante antérieure de l'algorithme de résolution des anaphores, appelé robuste, pour lequel Mitkov n'avait pas utilisé de parseur syntaxique, mais un étiqueteur morpho-syntaxique (*POS tagger*), un extracteur de phrases nominales et une série d'indicateurs d'antécédence (Mitkov 1998). Les indicateurs suivants sont utilisés pour assigner des scores (de -1 à 2) aux phrases nominales qui se trouvent dans les deux propositions précédant l'anaphore : *first noun phrase, indicating verbs, lexical reiteration, section heading preference, collocation match, immediate reference, sequential instructions, term preference, indefiniteness, prepositional noun phrase*. L'algorithme de résolution procède de la manière présentée à la Figure 16.

1. Examine the current sentence and the two preceding sentences (if available). Look for noun phrases only to the left of the anaphor.
2. Select from the identified noun phrases only those which agree in gender and number with the pronominal anaphor and group them as a set of potential candidates.
3. Apply the antecedent indicators to each potential candidate and assign scores; propose the candidate with the highest aggregate.

Figure 16 - Description de l'algorithme robuste de Mitkov (2002, p. 149)

Pour implémenter MARS, Mitkov ajoute trois nouveaux indicateurs d'antécédence (*boost pronoun, syntactic parallelism* et *frequent candidates*), une procédure qui identifie les pronoms explétifs, mais l'addition la plus importante est le parseur FDG qui en plus d'identifier les relations de dépendance dans le texte, permet le passage à une exécution entièrement automatique de la résolution.

La résolution en MARS est exécutée en cinq phases :

- analyse syntaxique en FDG

- identification des pronoms anaphoriques et non anaphoriques (*it* explétif inclus)
- pour chaque pronom anaphorique identifié, recherche des phrases nominales qui se trouvent dans la même proposition et dans les deux propositions antérieures. De cette liste sont éliminés les antécédents possibles qui ne remplissent pas les conditions d'accord morphologique et des filtres syntaxiques.
- les scores spécifiés par les indicateurs d'antécédence sont ensuite assignés aux candidats qui restent dans la liste
- le candidat avec le score le plus élevé est choisi comme antécédent

Les scores assignés par les indicateurs d'antécédence originaux ont été dérivés de manière empirique. Avec MARS ces scores sont optimisés en utilisant un algorithme génétique qui permet de trouver l'ensemble d'indicateurs pour lesquels la performance du système est maximale. Cependant, l'optimisation est dépendante du type de texte utilisé en entrée dans le processus d'optimisation. Ainsi, au moment où le texte en entrée n'est plus issu des manuels pour logiciels, la performance de MARS diminue.

5.3 Analyse sur corpus diagnostique

Nous avons créé un corpus (Annexe 1) contenant des cas simples d'analyse, mais aussi plusieurs types de structures qui s'avèrent problématiques pour les systèmes à base de connaissances linguistiques pauvres. Le corpus, annoté manuellement, a été utilisé dans l'évaluation comparative de la précision, du rappel et de la F-mesure pour les trois systèmes. Ce corpus n'est pas extrait d'un texte unitaire, mais est plutôt une collection de propositions indépendantes contenant pour la plupart des anaphores

intrapropositionnelles. Il est composé de 20 propositions contenant 26 pronoms, dont une proposition incluant un pronom impersonnel, pour un total de 22 relations anaphoriques, dont 4 utilisant des antécédents non-manifestes (vides) et une cataphore. Nous présentons des détails et des traces d'analyse des trois systèmes dans les Annexes 1 et 2.

Nous effectuerons trois tests. D'abord un test pour vérifier quel est l'effet de l'utilisation des relations asymétriques dans la résolution des anaphores pronominales sur un corpus (Annexe 1) conçu de manière à inclure des structures problématiques pour les systèmes à connaissances pauvres. Ensuite nous allons effectuer un deuxième test en éliminant les cas d'antécédents non-manifestes. Finalement, le troisième test vise à comparer les performances de notre système pour les deux types de résolution implémentées : la résolution tardive et la résolution à la volée.

Nous utiliserons les mesures d'évaluation les plus connues, soit la *précision*, le *rappel*, la *F-mesure*, dérivée des deux premières (80).

(80)
$$\text{Précision} = \frac{\text{nombre d'anaphores correctement résolues}}{\text{nombre de résolutions essayées}}$$

$$\text{Rappel} = \frac{\text{nombre d'anaphores correctement résolues}}{\text{nombre d'anaphores identifiées par le système}}$$

$$\text{F-mesure} = 2 \times \frac{\text{précision x rappel}}{\text{précision + rappel}}$$

5.3.1. Analyse des résultats

Les résultats où l'antécédent existe dans le texte, mais il n'a pas été identifié (ou il a été identifié partiellement) ont été considérés comme échec. Les résultats *null* ou *nothing* ont été considérés comme succès là où un antécédent n'existe pas dans le

texte. L'identification du pronom explétif a été considérée comme succès. MARS et RAP échouent dans l'identification du pronom explétif it en (20), fait qui porte le nombre de résolutions essayées à 26.

Après le test initial, les trois systèmes ont très bien performé sur les phrases avec des structures simples pour l'analyse desquelles l'information de surface est suffisante. Des difficultés se sont manifestées dans les cas plus complexes, qui nécessitaient une analyse complète. Nous présentons dans le Tableau 7 les résultats du premier test.

La précision, le rappel et la mesure F ont été calculés selon les formules ci-dessus (présentés en pourcentage) en utilisant le nombre total de relations anaphoriques et, pour chaque système, le nombre de relations anaphoriques résolues correctement et le nombre d'essais.

	Précision	**Rappel**	**Mesure F**
MARS	46	55	50
RAP	50	59	54
aLAD	96	96	96

Tableau 7 - Test sur le corpus

Après l'analyse des résultats, nous remarquons un nombre de facteurs qui ont causé les erreurs. Il y a des facteurs ponctuels qui provoquent des erreurs localisées, comme c'est l'autoréférence produite par RAP dans le cas du deuxième *it* dans la phrase analysée dans la Figure 17.

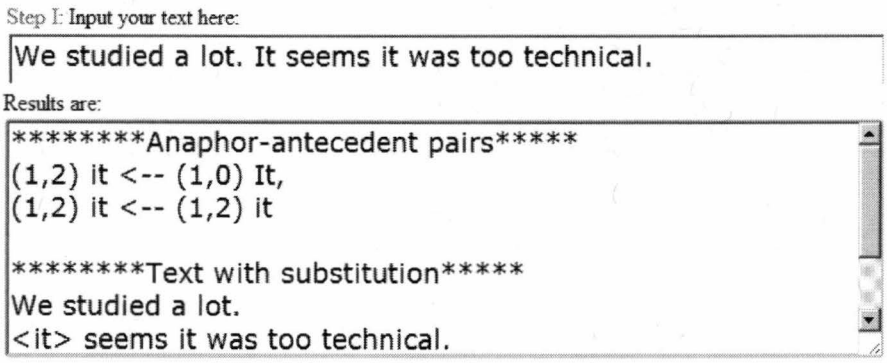


Figure 17 - Autoréférence en RAP

Ensuite il y des facteurs systémiques qui risquent de produire un nombre plus grand d'erreurs. RAP et MARS utilisent dans leurs analyses des connaissances linguistiques de surface qui ne leur permettent pas de traiter la structure sous-jacente du texte en entrée. Par exemple, dans l'analyse de la phrase : "Paul who knows the man walking on the street respects him" il y a deux constituants qui pourraient être considérés des candidats antécédents de l'anaphore pronominale *him* : *Paul* et *the man walking on the street*. En l'absence des connaissances linguistiques profondes, MARS ne reconnaît pas le cas d'anti-c-commande dans lequel se trouve le deuxième constituant, et, par conséquent, échoue dans cette analyse en choisissant *the street* comme antécédent. Dans ce même cas RAP ne précise pas avec exactitude les limites du constituant complet et choisit *the man* comme antécédent.

Le nombre d'analyses qui ont échoué à cause de l'absence des connaissances linguistiques profondes est de 8 pour MARS et de 9 pour RAP, ce qui cause une diminution de près de moitié de la précision des deux systèmes. Mentionnons en outre que MARS ne traite pas les pronoms réfléchis, ainsi *himself*, n'est pas reconnu comme une anaphore.

aLAD obtient de très bonnes performances grâce à l'information linguistique basée sur des relations asymétriques, incluant la c-commande et l'accord asymétrique.

Un pourcentage non-négligeable d'antécédents, (10%) ont été résolus erronément par MARS à cause du fait que le trait ±*animé* n'est pas pris en considération. Ainsi, MARS identifie en (11) *the street* comme l'antécédent de *him*, et en (12) *the other day* comme l'antécédent de *him.* RAP montre une plus grande mobilité quand il s'agit de chercher l'antécédent d'une anaphore interpropositionnelle, mais l'absence de traits lexicaux conduit à une mauvaise analyse, comme c'est le cas dans la proposition 15, exemple que MARS échoue aussi. Ce type d'erreur est généré de manière consistante par MARS qui n'utilise pas des représentations lexicales riches. Nous donnons en exemple en (83) la résolution échouée du pronom *him* :

(83) *The easiest thing would be to sell out to Al Budd and leave the country, but there was a stubborn streak in him that would_ n't allow it .*

him appears in paragraph 8, sentence 1, from position 24 to position 24. It is singular. The antecedent is indicated to be **the country** in paragraph 8, sentence 1, from position 14 to position 15.

Étant donné que MARS et RAP ne possèdent pas les connaissances linguistiques nécessaires pour traiter les cas de d'antécédents non-manifestes (non prononcés), nous avons considéré un deuxième scénario d'analyse où nous excluons les relations anaphoriques utilisant des antécédents non-manifestes (au nombre de quatre (4)), et considérons seulement les relations anaphoriques réalisées avec un antécédent manifeste. Dans ce deuxième cas les mesures de performances sont les suivantes (en pourcent).

	Précision	**Rappel**	**Mesure F**
MARS	55	57	56
RAP	59	62	60
aLAD	95	95	95

Tableau 8 - Test sur corpus sans antécédents vides

5.4 Test comparatif pour les deux types de résolution

En tant que locuteurs natifs d'une langue, nous avons une connaissance implicite du fonctionnement des relations qui sont à la base de la faculté du langage. Les interprétations de la phrase que nous sommes en train d'écouter commencent à se former à partir des premiers mots articulés. Il existe des études (Carminati 2002, Costa 2011) qui démontrent l'existence d'une phase locale de traitement où certains pronoms sont traités avant que le traitement passe à la phase suivante.

Pour que notre système automatique de résolution des anaphores puisse s'approcher plus du modèle humain de traitement des anaphores, nous avons conçu la stratégie de résolution *à la volée* qui interagit avec l'analyseur LAD pendant le traitement d'une proposition et résout les anaphores qui peuvent être résolues à ce niveau avant de passer à la phrase suivante. Une telle approche peut être très intéressante pour les applications avec des tâches de traitement de haut volume ainsi que pour les systèmes de réponse automatique à des questions. La rapidité de traitement de ces systèmes peut être augmentée puisque ces systèmes seraient en mesure de fournir des réponses sans faire l'analyse complète d'un texte.

Nous avons analysé le corpus utilisant la résolution tardive au préalable et la stratégie de résolution *à la volée* par la suite. Les tests ont été exécutés sur une machine Windows XP P4 avec un processeur de 2,99GHz et 1Go de RAM. Nous en présentons les résultats dans le Tableau 9.

	Précision	Temps CPU
Résolution tardive	inchangée	≈750ms./proposition
Résolution *à la volée*	inchangée	>1500ms./proposition

Tableau 9 - Comparaison des résultats obtenus avec les deux types de stratégies

Les résultats montrent que du point de vue de la précision les résultats sont identiques dans les deux configurations. Cependant, il y a un changement drastique dans le temps d'exécution qui passe de ≈750ms./proposition à plus de 1500ms./proposition.

Cet effet était attendu et il s'explique par le fait que, dans la résolution à la volée, toutes les anaphores qui ont été traitées, mais n'ont pas été résolues dans la phase propositionnelle doivent être traitées une deuxième fois au moment où l'analyse morphosyntaxique est terminée pour l'ensemble du texte. A ce moment la stratégie de résolution tardive s'applique sur l'ensemble d'anaphores non résolues.

Quoique le délai obtenu en utilisant stratégie de résolution à la volée soit non-négligeable, l'intérêt de cette stratégie demeure pour les applications qui requièrent une réponse avant la fin de l'analyse complète. En fait le nombre des résultats partiels possibles est directement proportionnel avec le nombre d'anaphores intrapropositionnelles.

5.5 Conclusion

Nous avons créé un corpus pour évaluer l'importance des données linguistiques riches dans la résolution des anaphores pronominales. Nous avons comparé les performances de notre système, aLAD, avec les performances des deux systèmes utilisant des connaissances linguistiques de surface. Les résultats obtenus valident l'hypothèse de départ que l'utilisation des relations d'asymétrie est profitable pour la résolution des anaphores pronominales et pour le TALN plus généralement. Nous avons montré l'importance d'accompagner l'utilisation de l'information linguistique riche par un mécanisme de suivi pondéré de la proéminence des entités participant à la dynamique du discours.

Finalement, nous avons testé la nouvelle stratégie de résolution à la volée et nous avons montré l'intérêt que pourraient avoir pour cette approche d'autres applications du TALN.

CONCLUSION ET TRAVAUX FUTURS

Le fait que les ordinateurs fassent de plus en plus partie de notre quotidien fait ressentir un besoin aigu d'amélioration de la qualité des transferts d'informations entre l'homme et l'ordinateur.

Le TALN est continuellement à la recherche des moyens pour améliorer les performances des systèmes automatiques de traitement du langage humain. Une des voies explorées par les chercheurs en TALN favorise un traitement utilisant des informations linguistiques de surface pour renforcer la stabilité des systèmes, ce qui va au détriment de la précision d'analyse.

Nous avons analysé la possibilité d'améliorer les performances des systèmes de résolution de l'anaphore pronominale par l'intégration dans ces systèmes des propriétés fondamentales de la faculté du langage. Nous avons précisé pourquoi les relations d'asymétrie sont privilégiées dans le traitement cognitif par opposition aux relations impliquant des points de symétrie. Notre hypothèse de départ est qu'un système de résolution automatique de l'anaphore pronominale aurait des meilleures performances s'il utilisait un modèle fidèle au traitement que les humains font des anaphores.

Pour la mise en œuvre de notre système, nous avions besoin d'un analyseur syntaxique qui reconnaisse les relations d'asymétrie et permet ensuite leur traitement dans un algorithme de résolution d'anaphores. Nous avons utilisé l'analyseur LAD, un analyseur basé sur la reconnaissance des asymétries d'interface. Nous avons en premier lieu étendu la couverture de LAD pour intégrer la vérification de l'accord et la c-commande asymétriques et nous avons implémenté des changements pour permettre au LAD d'analyser le discours.

Nous avons implémenté un algorithme de résolution des anaphores pronominales qui attribue aux relations asymétriques la place privilégiée qu'elles ont dans la cognition humaine en général et dans la faculté du langage en particulier.

Les résultats obtenus nous permettent de valider l'hypothèse de départ de notre recherche. Les tests que nous avons réalisés montrent que le fait d'utiliser des connaissances linguistiques riches, accompagné d'un modèle adéquat du traitement discursif, mène un système automatique de résolution des anaphores pronominales à des performances supérieures aux systèmes qui utilisent des connaissances linguistiques de surface.

Limites et travaux futurs

Bien que aLAD fonctionne en mode automatique quelques ajouts seraient nécessaires pour étendre plus la couverture du système. Le prototype LAD a encore un lexique réduit et il bénéficierait de l'ajout d'un lexique plus grand. Une connexion avec une base de données lexicales avait été considérée en cour d'implémentation et le travail préliminaire sur la correspondance des représentations lexicale dans les deux plateformes peut être récupéré pour accélérer l'implémentation. Après l'optimisation du lexique d'autres avenues sont possibles : grâce à l'analyse linguistique riche de LAD, le module de résolution anaphorique peut contribuer à l'extraction de connaissances des contextes jusque là opaques et à la création automatique d'ontologies.

De plus, le modèle de discours utilisé actuellement par notre système peut être amélioré avec l'inclusion d'une mémoire à long terme. Cette mémoire permettra de mettre en place un mécanisme de sécurité servant à réviser une décision si de l'information contraire est obtenue plus tard dans l'analyse. Le même ajout permettrait un suivi plus étroit de la dynamique du discours qui serait profitable pour l'utilisation du topique dans le processus de résolution.

aLAD ne traite pas les cas de cataphores (backwards anaphora), ce cas spécial d'anaphore où le pronom (dans notre cas) apparait dans le texte avant l'expression autonome du point de vue référentiel qui assure son interprétation sémantique et référentielle. Notre recherche est un travail exploratoire qui peut être continué dans des développements futurs pour étendre la couverture computationnelle du système au cas des cataphores et ainsi que des anaphores verbales et adverbiales.

L'analyseur LAD a été conçu de manière à permettre le changement des paramètres de variation linguistique. Une grammaire du français pour LAD est maintenant en développement, ce qui pourra permettre d'étendre la couverture de aLAD aux anaphores du français.

BIBLIOGRAPHIE

Agre, Phillip E. 1997. *Toward a Critical Technical Practice : Lessons Learned in Trying to Reform AI.* In: Geof Bowker, Les Gasser, Leigh Star, and Bill Turner, eds, *Bridging the Great Divide: Social Science, Technical Systems, and Cooperative Work*, Erlbaum.

Aone, Chinatsu et Scott William Bennett. 1995. *Evaluating Automated and Manual Acquisition of Anaphora Resolution Strategies* In: *Proceedings of the 33rd Annual Meeting of the ACL*, Santa Cruz, New Mexico, 122-129.

Asher, Nicholas et Hajime Wada. 1988. *A Computational Account of Syntactic, Semantic and Discourse Principles for Anaphor Resolution.* Journal of Semantics 6, 309-344.

Bach, Emmon et Barbara Partee. 1980. *Anaphora and Semantic Structure*. In J. Kreiman and A. Ojeda, eds., Papers from the Parasession on Pronouns and Anaphora, Chicago Linguistic Society, Chicago (1980) 1-28.

Baldwin, Breck. 1997. *CogNIAC: High Precision Coreference with Limited Knowledge and Linguistic Resources.* In: *Proceedings of the ACL'97 / EACL'97 Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts*, Madrid, Spain, 38-45.

Barbu, Catalina et Ruslan Mitkov. 2001. *Evaluation Tool for Rule-based Anaphora Resolution Methods*. In *Proceedings of ACL'01*, Toulouse. France.

Bühler, Karl. 1982. *The Deictic Field of Language and Deictic Words.* Dans R. J. Jarvella et W. Klein (eds.). *Speech, Place, and Action.* New York : John Wiley

Byron, Donna K. 2001. *The Uncommon Denominator: A Proposal for Consistent Reporting of Pronoun Resolution Results.* In: *Computational Linguistics*, Vol. 27, Number 4, 569-578.

Byron, Donna K. et James F. Allen. 2002. *What's a Reference Resolution Module to do? Redefining the Role of Reference in Language Understanding Systems.* In: *Proceedings of the 4th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC2002)*, University of Lisbon, Portugal, 25-30.

Carbonell, Jaime G. et Ralf D. Brown. 1988. *Anaphora Resolution: A Multi-Strategy Approach.* Proceedings of the 12th International Conference on Computational Linguistics, Budapest, 96-101.

Carminati, Maria Nella. 2002. The Processing of Italian Subject Pronouns. Doctoral Dissertation. University of Massachusetts. Amherst.

Charniak, Eugene. 1972. *Toward a model of children's story comprehension.* AI Technical Report 266, Massachusetts Institute of Technology Artificial Intelligence Laboratory.

Chierchia, Gennaro. 1992. *Anaphora and dynamic binding, Linguistics and Philosophyy*, vol. 15, pp. 111-183.

Chieu, Hai et Hwee Ng. 2002. *A Maximum Entroy Approach to Information Extraction from Semistructured and Free Text.* In *Proceedings of the Eighteenth International Conference on Artificial Intelligence (AAAI-02).* Edmonton, Canada

Chomsky, Noam. 1970. *Remarks on nominalization.* In: *Reading in English Transformational Grammar*, Roderick A. Jacobs, Peter S. Rosenbaum, Eds., pp. 184-221. Waltham, Ginn.

Chomsky, Noam. 1981. *Lectures on Government and Binding.* Foris Publications, Dordrecht.

Chomsky, Noam. 1995. *The minimalist program*. Cambridge, Massachusetts, The MIT Press.

Chomsky, Noam. 2000. *Minimalist inquiries*, In: *Step by Step. Essays on Minimalist Syntax in Honor of Howard Lasnik*, R. Martin, D. Michaels and J. Uriagereka, Eds. Cambridge, Mass.: The MIT Press, pp. 89- 155.

Chomsky, Noam. 2008. *On Phases*, In: *Foundational issues in linguistic theory: essays in honor of Jean-Roger* Vergnaud, Robert Freidin, Carlos Peregrín Otero, Maria Luisa Zubizarreta eds., MIT.

Connolly, Dennis, John D. Burger et David S. Day. 1994. *A Machine-Learning Approach to Anaphoric Reference.* In: *Proceedings of the International Conference on New Methods in Language Processing (NEMLAP).*

Costa, A., G. Matos et P. Luegi. 2011. *Using eye-tracking to study anaphoric relations processing in European Portuguese.* In *Journal of Eyetracking, Visual Cognition and Emotion.* Vol. 1, No. 1. pp. 50-58.

Dagan, Ido et Alon Itai. 1991. *Automatic Processing of Large Corpora for the Resolution of Anaphoric References.* In: *Proceedings of the 13th International Conference on Computational Linguistics (COLING)*, Vol. 3, Helsinki. pp. 330-332.

Danlos, Laurence. 2001. *Event Coreference Between Two Sentences*. Dans H. Bunt, R. Muskens et E.Thijsse (éds.), *Computing Meaning.* Vol. 2. 271-288. Amsterdam: Kluwer Academic Publishers.

Danlos, Laurence et B. Gaiffe. *Event coference and discourse relations.* In L. Kulda (ed), *Language, Music and Cognition.* Amsterdam : Kluwer Academic Publishers.

Di Sciullo, Anna Maria. 1995. *X' Selection.* In *Phrase Structure and the Lexicon*, Johan Rooryck et Laurie Ann Zaring Eds, pp. 77-107. Dordrecht : Kluwer.

Di Sciullo, Anna Maria. 1996. Atomicity and Relatedness in Configurational Morphology. Chap. In Configurations: Essays on Structure and Interpretation, p. 17-41. New York : Cascadilla Press.

Di Sciullo, Anna Maria. 1997. *Selection and Derivational Affixes*. In *Progress in Morphology*. Wolfgang Dressler, Ed. 79-97. Berlin : Walter de Gruyter.

Di Sciullo, Anna Maria. 1997a. *Argument Structure Parsing*. In *Papers in Natural Language Processing*. Angela Ralli, Maria Grigoriadou, George Philokyprou et Dimitris Christodoulakis, Eds. p. 55-77. Athènes : Diavlos.

Di Sciullo, Anna Maria. 1999. *The Local Asymmetry Connection*, In: *MIT Working Papers in Linguistics,* Vol. 35, pp. 25-49.

Di Sciullo, Anna Maria, 1999a. *An Integrated Competence-Performance Model, A Prototype for Morpho-Conceptual Parsing and Consequences for Information Processing*. In *Proceedings of VEXTAL*. Università Ca'Foscari Venezia : San Servolo, V.I.U. pp. 369-377.

Di Sciullo, Anna Maria. 2000. *Parsing asymmetries. Natural Language Processing.* Springler Computer Science Press, pp. 24-39.

Di Sciullo, Anna Maria. 2003. *Morphological phases*, In: *Proceedings of the 4th GLOW in Asia 2003, Generative Grammar in a Broader Perspective. The Korean Generative Grammar Circle*, H.-J. Yoon, pp. 113-136.

Di Sciullo, Anna Maria. 2003a. *Morphological relations in asymmetry theory*, In: *Asymmetry in Grammar, volume 2: Morphology, Phonology and Acquisition* A.M. Di Sciullo, Ed. Amsterdam: John Benjamin, pp. 1-38.

Di Sciullo, Anna Maria. 2005. *Morpho-Syntax Parsing*. (avec Sandiway Fong) In *UG and External Systems*. Anna Maria Di Sciullo, Ed. Amsterdam : John Benjamins.

Di Sciullo, Anna Maria. 2005a. *Asymmetry in Morphology*. Cambridge, Mass.: The MIT Press.

Di Sciullo, Anna Maria. 2005b. *Asymmetry theory in internet infrastructure*, In: *International Journal of Electronic Business*, Vol 3: Special Issue on: Multidisciplinary, Interdisciplinary and Transdisciplinary Research in Electronic Business, pp. 228-238.

Di Sciullo, Anna Maria. 2005c. *Domains of Argument Structure Asymmetries*. In *Proceedings of the 9th World Multiconference on Systemics, Cybernetics and Informatics*, pp. 316-320. Orlando, Florida.

Di Sciullo, Anna Maria. 2006. *Pronominal anaphora processing*. Computer Science. Vol 11. pp. 67-74.

Di Sciullo, Anna Maria. 2009. *Why are compounds part of natural languages : a view from Asymmetry Theory*. In *Handbook of Compounds*, Rochelle Lieber et Pavol Štekauer, Eds. pp. 145-177. Oxford : Oxford University Press.

Di Sciullo, Anna Maria. 2010. *Information processing*, In: H. Fujita and D. Pisanelli (eds.), *New Trends in Software Methodologies Tools and Techniques. Frontiers in Artificial Intelligence and Applications*. Amsterdam: IOS Press.

Di Sciullo, Anna Maria. 2014. *Minimalism and I-Morphology*. In *Minimalism and Beyond: Radicalizing the Interfaces*. Peter Kosta, Steven Franks and Teodora Radeva-Bork, Eds. Amsterdam: John Benjamins.

Di Sciullo, Anna Maria, Ileana Paul et Stanca Somesfalean. 2003. *The Clause Structure of Extraction Asymmetries*. In *Asymmetry in Grammar*. Vol. 1: *Syntax and Semantics*. Anna Maria Di Sciullo, Ed. pp. 279-300. Amsterdam: John Benjamins.

Di Sciullo, Anna Maria, Philippe Gabrini, Calin Batori et Stanca Somesfalean. 2006. *Asymmetry, the grammar and the parser*. In: *Linguistica e modelli tecnologici de ricerca, XL Congresso SLI*. Vercelli, Italia.

Di Sciullo, Anna Maria. et S. Fong. 2002a. *Asymmetry, Zero Morphology and Tractability*. In Language, Information and Computation. PACLIC 15. Language Information Sciences Research Center, 61-72. University of Hong Kong.

Di Sciullo, Anna Maria. et S. Fong. 2002b. *Efficient Parsing for Word Structure*. In Natural Language Processing Pacific Rim Symposium, 741-748. Tokyo, Japon.

Di Sciullo, Anna Maria et S. Fong. 2005. *Morpho-syntax parsing*. In A. M. Di Sciullo, ed., *UG and External Systems*. Amsterdam: John Benjamins. Pp. 247-268.

Di Sciullo, Anna Maria, Stanca Somesfalean, Calin Batori et Philippe Gabrini. 2010. *Pronominal anaphora understanding*. Intelligent Systems Design and Applications (ISDA). Cairo, Egypt.

Di Sciullo, Anna Maria et Naoko Tomioka. 2008. *Priming with the argument/modifier asymmetry with Japanese compounds*. Ms UQAM.

Ducrot, Oswald. & T. Todorov. 1972. *Dictionnaire encyclopédique des Sciences du Langage*, Paris : Le Seuil.

Evans, Gareth. 1977. *Pronouns, Quantifiers and Relative Clauses.* Canadian Journal of Philosophy 7: 467–536.

Evans, Gareth. 1980. *Pronouns*, *Linguistic Inquiry* vol. 11, pp. 337-362.

Evans, Richard et C. Orasan. 2000. *Improving Anaphora Resolution by Identifying Animate Entities in Texts*. In Proceedings of the Discourse Anaphora and Reference Resolution Conference (DAARC2000), pp. 154 – 162 Lancaster, UK.

Fodor, Jerry. 1983. *The Modularity of Mind: An Essay on Faculty Psychology*, MIT Press.

Fodor, Jerry et Zenon Pylyshyn. 1988. *Connectionism and cognitive architecture: A critical analysis*, *Cognition*, 28, pp. 3-7.

Fradin, Bernard. 1984. *Anaphorisation et stéréotypes nominaux*. Lingua 64. 325-369.

Fradin, Bernard. 1986. *Pragmatique et constitution de la signification lexicale*. Cahiers de Linguistique Française. 7, 115-134.

Ge, Niye, John Hale et Eugene Charniak. 1998. *A statistical approach to anaphora resolution*. In *Proceedings of the Sixth Workshop on Very Large Corpora*. pp. 161-171.

Geach, Peter Thomas. 1962. *Reference and Generality*. Ithaca: Cornell University Press.

Grodzinsky, Yosef et Tanya Reinhart. 1993. *The Innateness of Binding and Coreference*. Linguistic Inquiry, Vol. 24, No. 1 (Winter, 1993), 69-101.

Grosz, Barbara J. 1977. *The Representation and Use of Focus in a System for Dialogue Understanding*. Technical Report 151, Artificial Intelligence Center, SRI International, Menlo Park, CA.

Grosz, Barbara J. et Candace L. Sidner. 1986. *Attention, Intentions, and the Structure of Discourse*. In: *Computational Linguistics*, Vol. 12, Number 3, July-September, 175-204.

Hale, Ken et Jay Keyser. 2002. *Prolegomena to a Theory of Argument Structure*. Cambridge, Mass.: The MIT Press.

Halliday Michael A K et R. Hasan. 1976. *Cohesion in English*. London, Longman.

Heim, Irene. 1982. *The Semantics of definite and indefinite NPs*. Ph.D. dissertation, University of Amherst.

Higginbotham, James. 1980. *Anaphora and GB: Some Preliminary Remarks*. In J. Jensen, ed., *Cahiers Linguistiques d'Ottawa: Proceedings of NELS 10*, University of Ottawa.

Hirst, Graeme. 1981. *Anaphora in Natural Language Understanding: A Survey*. Lecture Notes in Computer Science 119, Springer-Verlag.

Hobbs, Jerry R. 1978. *Resolving Pronoun References*. In: *Lingua*, Vol. 44, 311-338.

Hobbs, Jerry R. 1979. *Coherence and Coreference*. In: *Cognitive Science*, Vol. 3(1), 67-82.

Huang, Cuili. 2010. *Coreference resolution in biomedical full-text articles with domain dependent features. 2nd International Conference on Computer Technology and Development (ICCTD)*. Cairo. Egypt.

Ingria, Robert J. P. et David Stallard. 1989. *A Computational Mechanism for Pronominal Reference*. In: *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*. Vancouver.

Jackson, Peter et Isabelle Moulinier. 2002. Natural Language Processing for Online Applications: Text Retrieval, Extraction & Categorization. John Benjamins.

Kabadjov, Mijail. 2007. A Comprehensive Evaluation of Anaphora Resolution and Discourse-new Classification. Doctoral Dissertation. University of Essex.

Kayne, Richard. 2001. *Pronouns and their Antecedents.* In *Derivation and Explanation in the Minimalist Program.* Samuel Epstein and Daniel Seely, Eds. pp. 133-166. Malden, MA: Blackwell.

Klappholz, A. David et David Lockman. 1975. *Contextual reference resolution.* In: *American Journal of Computational Linguistics*, microfiche 36, 4-25.

Kennedy, Christopher et Branimir Boguraev. 1996. *Anaphora for Everyone: Pronominal Anaphora Resolution without a Parser.* In: *Proceedings of the 16th International Conference on Computational Linguistics (COLING),* Vol. I, Kopenhagen, 113-118.

Küppers, Günter et Johannes Lenhard. 2005. *Validation of Simulation: Patterns in the Social and Natural Sciences.* In: *Journal of Artificial Societies and Social Simulation* 8(4)3

Lappin, Shalom. 2003. *A Sequenced Model of Anaphora and Ellipsis Resolution.* In the 4th Discourse Anaphora and Anaphora Resolution Colloquium. Lisbon.

Lappin, Shalom et Herbert J. Leass. 1994. *An Algorithm for Pronominal Anaphora Resolution.* In: *Computational Linguistics,* 20 (4), 535-561.

Lappin, Shalom et Michael McCord. 1990. *A Syntactic Filter on Pronominal Anaphora for Slot Grammar.* In: *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*, 135-142.

Lappin, Shalom et Michael McCord. 1990. *Anaphora Resolution in Slot Grammar.* In: Computational Linguistics, 16 (4), 197-212.

Lin, Y., et T. Liang. 2004. *Pronominal and Sortal Anaphora Resolution for Biomedical Literature.* ROCLING XVI: Conference on Computational Linguistics and Speech. Taiwan.

Lyons, John. 1977. *Semantics.* Cambridge: Cambridge University Press.

Miller, G. A. (1977). *Practical and lexical knowledge.* In P. N. Johnson-Laird and P. C. Wason (eds.), *Thinking: Reading in Cognitive Science.* Cambridge: Cambridge University Press.

Mitkov, Ruslan. 2001. *Outstanding Issues in Anaphora Resolution.* In: Alexander Gelbukh (ed). *Proceedings of the Second International Conference on Computational Linguistics and Intelligent Text Processing (CICLing)*, Mexico City, 110-123.

Mitkov, Ruslan. 2002. Anaphora Resolution. London, Longman.

Moro, Adrea. 2000. *Dynamic Antisymmetry*, Cambridge, Mass.: The MIT Press.

Neale, Stephen. 2005. *Pragmatism and Binding*. In Szabo, ed. *Semantics versus Pragmatics*. Oxford, Oxford University Press.

Orasan Constantin et R. Evans. 2007. *NP Animacy Identification for Anaphora Resolution*. Journal of Artificial Intelligence Research, 29:79–103.

Paice, Chris et G.D. Husk. 1987. *Towards the automatic recognition of anaphoric features in English text: the impersonal pronoun "it"*. In *Computer Speech & Language*. Volume 2, Issue 2, June 1987, 109-132.

Partee, Barbara. 1978. Bound variables and other anaphors. In Theoretical Issues In Natural Language Processing 2 (TINLAP-2), ed. David L. Waltz, 79-85. Urbana, IL: University of Illinois.

Partee, Barbara et Emmon Bach. 1981. *Quantification, Pronouns and VP Anaphora*. In *Formal Methods in the Study of Language*, ed. J. Groenendijk, T. Janssen, M. Stokhof, Mathematisch Centrum, Amsterdam. 445-481.

Pérez, Celia Rico. 1994. *Estudio de la incidencia de differentes fuentes de la información en el establecimiento de relaciones anafóricas*. Bulletín de la Sociedad Española para el Procesamiento del Lenguaje Natural, No. 14, March.

Popescu-Belis, Andrei et Isabelle Robba. 1997. *Cooperation between Pronoun and Reference Resolution for Unrestricted Texts*, In: *Proceedings of the ACL'97/EACL'97 Workshop*, Madrid, Spain, pp.94-99

Preiss, Judita. 2002. *Choosing a Parser for Anaphora Resolution.* In: *Proceedings of the 4th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC2002)*, University of Lisbon, Portugal, 175-180.

Rappaport, Gilbert. 1986. *On Anaphor Binding in Russian*. Natural Language and Linguistic Theory 4, pp. 97–120.

Reboul, Anne. 1989. *Résolution de l'anaphore pronominale: sémantique et/ou pragmatique*. Cahiers de linguistique française. 10, 77-100.

Refoufi, Allaoua. 2014. *Pronominal Anaphora Resolution Using XML Tagged Documents*. Journal of Computer Engineering and Information Technology. 3:1.

Reinhart, Tanya. 1983a. *Anaphora and Semantic Representation*. Chicago: The University of Chicago Press.

Reinhart, Tanya. 1983b. *Coreference and Bound Anaphora: A Restatement of the Anaphora Questions*. Linguistics and Philosophy 6, 47–88.

Reinhart, Tanya. 1999. *Binding Theory*. In The *MIT Encyclopedia of the Cognitive Sciences*, eds. R. Wilson and Frank C. Keil, 86-88. Cambridge, Mass: MIT Press.

Reinhart, Tanya. 2006. *Interface Strategies: Optimal and Costly Computations*. Cambridge, Mass.: MIT Press.

Reinhart, Tanya et Eric Reuland. 1993. *Reflexivity*. In: *Linguistic Inquiry*. No. 24 pp. 657-720.

Reuland, Eric. 2011. *Anaphora and Language Design*. Cambridge, Mass: MIT Press.

Rich, Elaine et Susann LuperFoy. 1988. *An architecture for anaphora résolution.* ACL Proceedings, Second Conference on Applied Natural Language Processing, pp. 18-24.

Rizzi, Luigi. 1990. *Relativized Minimality*, Cambridge, Mass.: The MIT Press.

Roussarie, Laurent et P. Amsili. 2007. *Interpréter les pronoms phrastiques*, In Modèles Linguistiques, 56, 83-112.

Sag, Ivan A. 1976. *Deletion and Logical Form.* Doctoral Dissertation. Cambridge, Mass: MIT Press.

Sidner, Candace. 1978. A progress report on the discourse and reference components of PAL, In: Proceedings of the second national conference, Canadian Society for Computational Studies of Intelligence / Societé canadienne des études d'intelligence par ordinateur. Toronto, 206-213.

Sidner, Candace. 1979. *Towards a Computational Theory of Definite Anaphora Comprehension in English Discourse.* M.I.T. Artificial Intelligence Laboratory, TR-537.

Sidner, Candace. 1981. *Focusing for Interpretation of Pronouns*. American Journal of Computational Linguistics, Vol.7, 217-231.

Sidner, Candace. 1983. *Focusing in the Comprehension of Definite Anaphora.* In: Michael Brady, Robert C. Berwick (eds). *Computational Models of Discourse*. M.I.T. Press, Cambridge, MA.

Soames, Scott. 1994. *Attitudes and Anaphora.* In Philosophical Perspectives, Vol. 8, Issue Logic and Language. pp 251 – 272.

Sperber, Dan et D. Wilson. 1986. *Relevance: Communication and cognition*. Oxford: Blackwell.

Stuckardt, Roland. 2001. *Design and Enhanced Evaluation of a Robust Anaphor Resolution Algorithm.* In: *Computational Linguistics*, Vol. 27, Number 4, 479-506.

Stuckardt, Roland. 2004. *Three Algorithms for Competence-Oriented Anaphor Resolution.* In: *Proceedings of the 5th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC2004)*, Sao Miguel/Azores, 157-163.

Tapanainen, Pasi et T. Järvinen. 1997. *A non-projective dependency parser*, In: *Proceedings of the 5th Conference on Applied Natural Language Processing*, Washington, D.C, 1997, pp. 64-71.

Tetreault, Joel. 2001. *A Corpus-Based Evaluation of Centering and Pronoun Resolution.* In: *Computational Linguistics*, Vol. 27, Number 4, 507-520.

Tsapkini, Kyrana, Gonia Jarema et Anna Maria Di Sciullo. 2004. The role of configurational asymmetry in the lexical access of prefixed verbs: Evidence from French. Brain and Language 90: 143-150.

Vilain, Marc, John Burger, John Aberdeen, Dennis Connolly et Lynette Hirschman. 1995. *A Model-Theoretic Coreference Scoring Scheme.* In: *Proceedings of the*

Sixth Message Understanding Conference (MUC-6), Morgan Kaufmann Publishers.

Webber, Bonnie Lynn. 1988. *Discourse Deixis: Reference to Discourse Segments.* In: *Proceedings of the 26th Annual Meeting of the Association for Computational Linguistics (ACL)*, 113-121.

Soon, Wee Meng, Hwee Tou Ng et Daniel Chung Yong Lim. 2001. *A Machine Learning Approach to Coreference Resolution of Noun Phrases.* In: *Computational Linguistics*, Vol. 27, Number 4, 521-544.

Weinberg, Amy. 1999. A Minimalist theory of human sentence processing. In: Samuel D. Epstein and Norbert Hornstein (eds) Working Minimalism, Cambridge, Mass.: The MIT Press, pp. 283-315.

Williams, Edwin. 1977. *Discourse and Logical Form*. Linguistic Inquiry 8, 101-140.

		Pronom	Antécédent	MARS	RAP	aLAD
1.	John knows himself.	himself	John	none	John	John
2.	John likes him.	him	none	nothing	null	none
3.	Jane had left when he called her.	he; her	none; Jane	nothing; Jane	null; Jane	none; Jane
4.	John thinks that Mary likes him.	him	John	John	John	John
5.	John's brother likes him.	him	John	John's	null	John
6.	The brother of John likes him.	him	John	John	null	John
7.	The man who came with John likes him.	him	John	John	null	John
8.	The children that deserve them will receive the prizes they desire.	them (cataphora); they	the prizes; the children that deserve them	nothing; the children that deserve them	the children; the children that deserve them	null; the children that deserve them
9.	John knows the answer and Mary knows it too.	it	the answer	the answer	null	the answer
10.	Paul who knows John for one year respects him.	him	John	John for one year	null	John
11.	Paul who knows the man walking on the street respects him.	him	the man walking on the street	the street	the man	the man walking on the street
12.	The accountant that the manager brought Paul the other day betrayed him.	him	the manager/Paul	the other day	the manager	the manager
13.	John met Mary. She really likes him.	she; him	Mary; John	Mary; John	Mary; John	Mary; John
14.	They promised to leave. They did.	they; they	none; they	nothing; they	null; they	none; they
15.	Jane's sister arrived. Paul's sister saw him.	him	Paul	Jane's sister	Jane's sister	Paul
16.	An archeologist discovered a skeleton. He sent it to the museum.	he; it	an archeologist; a skeleton	a skeleton; an archeologist	an archeologist; an archeologist	an archeologist; a skeleton
17.	We ate *Pro* a lot. It was too spicy.	it	Pro	a lot	a lot	Pro
18.	John knows *Pro* and Mary knows it too.	it	Pro	nothing	null	Pro
19.	John suggested *Pro* to leave. They did.	they	Pro	nothing	null	Pro
20.	We studied *Pro* a lot. It seems it was too technical.	(it); it	none; Pro	a lot; a lot	it(cataphora); it(autoreference)	*expletif**; Pro
				12/25(+1)	13/25(+1)	24(+1)/25(+1)

ANNEXE 2 : TRACES D'ANALYSE

Texte en entrée : The man who came with John likes him.

Analyse Java RAP :

Step I: Input your text here:

The man who came with John likes him.

Or upload it:

Choose File No file chosen

Step II:

Select the operation:

Split Sentences Reminder: The **resolving** process might take a while ...

Step III:

Submit

Results are:

```
********Anaphor-antecedent pairs*****
NULL <-- (0,7) him

********Text with substitution*****
The man who came with John likes him.
```

Go back to the main page of JavaRAP.

You might want to take a look at the parsed text.

Syntactic analysis.

Listing collocation patterns.

Classifying *it*.

Resolving pronominal anaphors.

The man who came with John likes him

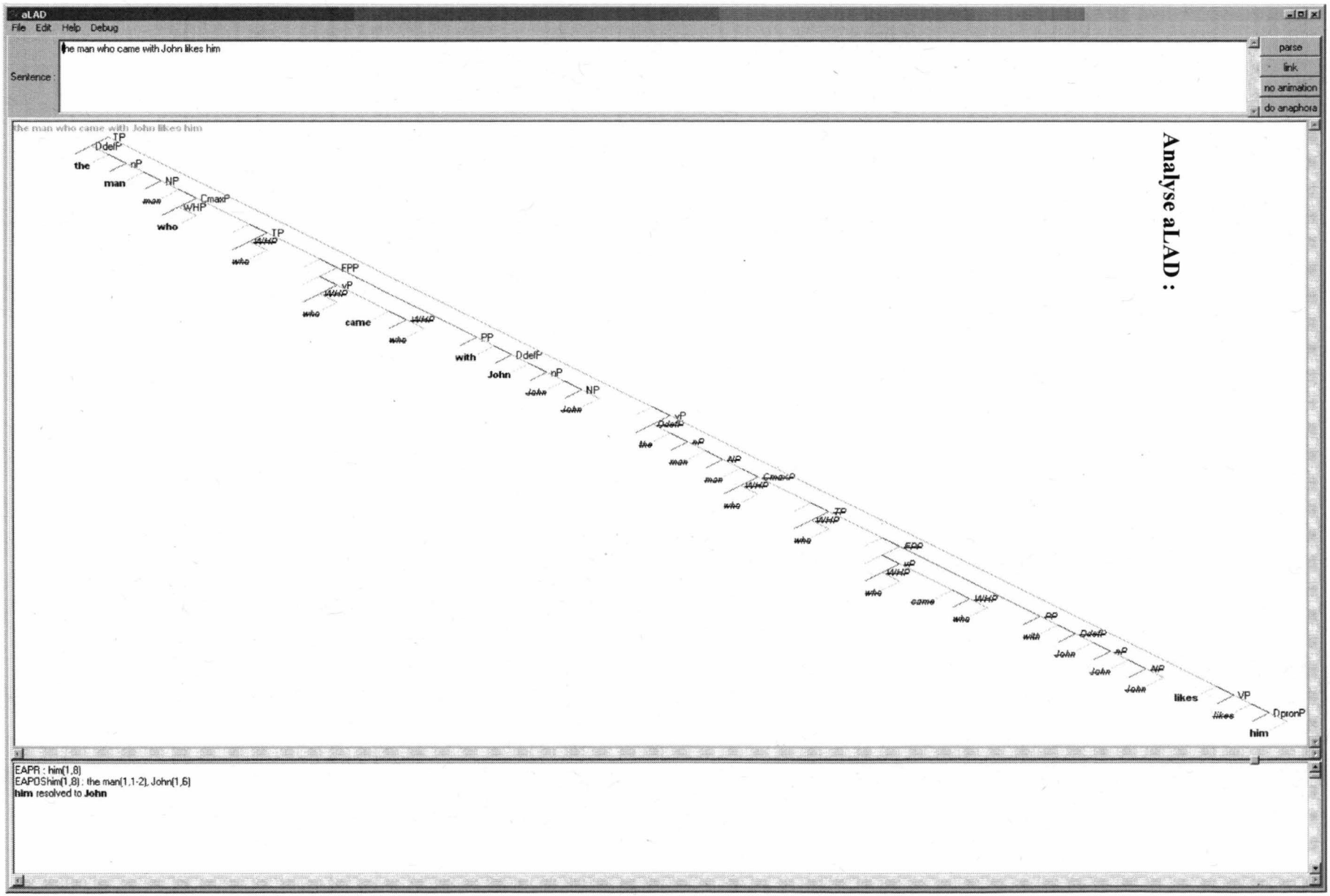
The man who came with John likes him

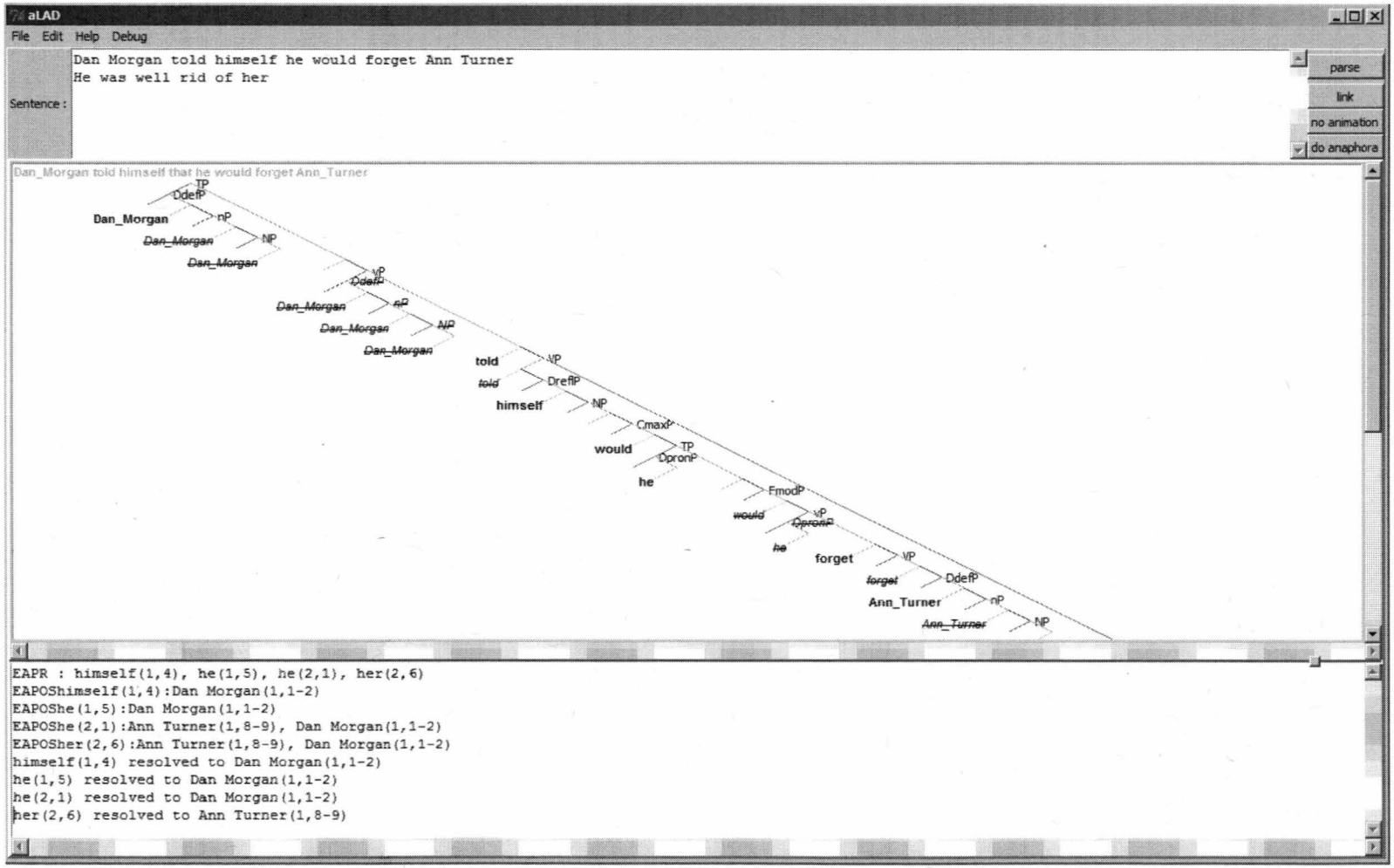
him appears in paragraph 1, sentence 1, from position 8 to position 8. It is singular. The antecedent is indicated to be **John** in paragraph 1, sentence 1, from position 6 to position 6.

Candidates were:

Candidates	Agreement	Sent	Pos	Def	Given	I-V	F-Cand	Rei	Sect-Hdg	N-PP-NP	Col-Pat	Im-Ref	Seq-Inst	Ref Dist	Term Pref	Syn_Par	B_P	TOTAL SCORE
John	singular, masc	1	6 - 6	Indefinite (-1)	Not subject, object, or indirect object (-1)	0	0	NP with same head repeated less than twice in paragraph (+0)	1	PP argument (-1)	0	0	0	Appears in same sentence as pronoun (+2)	0	0	0	0 (*indicators*)

Click here to return to the text input page. Alternatively, if you don't want to lose your last input, press 'back' from your browser.


aLAD
File Edit Help Debug
Sentence :
the man who came with John likes him
parse
link
no animation
do anaphora
Analyse aLAD :
EAPR : him(1,8)
EAPOShim(1,8) : the man(1,1-2), John(1,6)
him resolved to John


aLAD
File Edit Help Debug
Sentence :
Dan Morgan told himself he would forget Ann Turner
He was well rid of her
parse
link
no animation
do anaphora
Dan_Morgan told himself that he would forget Ann_Turner
EAPR : himself(1,4), he(1,5), he(2,1), her(2,6)
EAPOShimself(1,4):Dan Morgan(1,1-2)
EAPOShe(1,5):Dan Morgan(1,1-2)
EAPOShe(2,1):Ann Turner(1,8-9), Dan Morgan(1,1-2)
EAPOSher(2,6):Ann Turner(1,8-9), Dan Morgan(1,1-2)
himself(1,4) resolved to Dan Morgan(1,1-2)
he(1,5) resolved to Dan Morgan(1,1-2)
he(2,1) resolved to Dan Morgan(1,1-2)
her(2,6) resolved to Ann Turner(1,8-9)

ANNEXE 3 : HEURISTIQUES

Patrons:

VERB it STATUS to TASK

it VERB STATUS that

it VERB STATUS to TASK

it VERB STATUS that

it VERB that

VERB class:

anticipate

appear

assume

be

believe

find

follow

make

mean

seem

STATUS class:

advisable

believed

certain

convenient

correct

crucial

desirable

determinant

essential

fitting

handy

important

probable

sound

suggested

wise