L'ambiguïté anaphorique et la résolution automatique de l'anaphore pronominale

Afef Selmi

THESE DE DOCTORAT DE L'UNIVERSITE BOURGOGNE FRANCHE-COMTE

PREPAREE AU CENTRE INTERLANGUES - TEXTE, IMAGE, LANGAGE (TIL EA 4182)

Ecole doctorale n°592

LECLA (Lettres, Communication, Langues, Arts)

Doctorat de Sciences du langage

Par

Afef SELMI

L'ambiguïté anaphorique et la résolution automatique de l'anaphore pronominale

Propositions méthodologiques à partir d'un corpus fermé de résumés d'œuvres littéraires en français (RESUMAN)

Thèse présentée et soutenue à Dijon, le 11/12/2019

Composition du Jury :

M. Éric WEHRLI	Professeur honoraire, Université de Genève	Président
M. Lotfi ABOUDA	Maitre de conférences HDR, Université d'Orléans	Rapporteur
Mme Estelle DUPUY	Maitre de conférences, Université de Poitiers	Examinatrice
M. Laurent GAUTIER	Professeur, Université de Bourgogne	Directeur de thèse

ED 592
LECLA

Titre : L'ambiguïté anaphorique et la résolution automatique de l'anaphore pronominale « Propositions méthodologiques à partir d'un corpus fermé de résumés d'œuvres littéraires en français (RESUMAN)»

Mots clés : Anaphore, pronom, ambiguïté, corpus, TAL, résolution.

Résumé : Les anaphores pronominales sont sources d'ambiguïtés potentielles lors de l'exécution de requêtes cherchant bon antécédent. Dans un contexte d'automatisation du traitement linguistique et textuel, la recherche en traitement automatique des langues (TAL) a déjà mis en place diverses approches pour résoudre le problème. Nous nous proposons dans cette thèse d'adapter ces approches à la résolution automatique des anaphores pronominales dans un corpus fermé de résumés d'œuvres littéraires en français (RESUMAN).
Dans un premier temps sont présentés les modèles linguistiques, textuels et cognitifs, qui rendent compte du fonctionnement et de l'interprétation des pronoms personnels dans des textes en français. Ensuite sont exposées les procédures d'interprétation des anaphores pronominales : la procédure de distance minimale, la procédure des fonctions parallèles, la procédure du sujet ou la procédure thématique, ainsi que les procédures d'analyse morphologique, sémantique et pragmatique. Le corps de la thèse présente, enfin, une nouvelle version d'algorithme, se basant sur une approche statistique. L'outil RESUMAN permet d'améliorer les performances d'un TAL en cas de textes denses en anaphores pronominales comme ceux du corpus-échantillon étudié. Les performances de cet outil sont évaluées et ses limites sont commentées.

Title : Anaphoric ambiguity and automatic resolution of the pronominal anaphora "Methodological Proposals from a closed body of abstracts of literary works in french (RESUMAN)"

Keywords : Anaphora, pronoun, ambiguity, corpus, TAL, resolution.

Abstract : Pronominal anaphores can create ambiguities when requesting the correct antecedent. In a context of automation of linguistic and textual processing, several researchers in the field of automatic language processing (TAL) have put in place various approaches to solve the problem. We propose to adapt these approaches to the automatic resolution of pronominal anaphors in a closed body of abstracts of French literary works (RESUMAN).
Firstly, linguistic, textual and cognitive models are presented, which report on the functioning and interpretation of personal pronouns in French texts. Then, the interpretation procedures of the pronominal anaphors are exposed : the procedure of minimal distance, the procedure of the parallel functions, the procedure of the subject or the thematic procedure, as well as the procedures of morphological, semantic and pragmatic analysis. Finally, a new version of algorithm, based on a statistical approach, is presented. The RESUMAN tool makes it possible to improve the performance of a TAL in the case of texts that are pronounced pronominal anaphors like those of the corpus-sample studied. The performance of this tool is evaluated and its limits are commented.

UBFC UNIVERSITÉ BOURGOGNE FRANCHE-COMTE
Université Bourgogne Franche-Comté
32, avenue de l'Observatoire
25000 Besançon

À mon phare qui me dirige toujours vers mon havre de paix, à PAPA
À ma muse pleine de bonté et de grâce, pleine de talents, à ma chère MAMA
À mon petit soleil d'amour, à ma fille Amira

Remerciements

Un de mes anciens enseignants disait que faire un travail de recherche c’est emprunter un chemin qui se construit au fur et à mesure que l’on avance. Il n’est déjà pas toujours facile d’arriver au bout d’un chemin déjà tracé, que dire alors d’un chemin non tracé ?

Grâce à un certain nombre de personnes à qui je suis redevable, mon chemin a été tracé et une étape franchie.

Tout d’abord, je tiens à exprimer toute ma gratitude à Laurent GAUTIER de m’avoir accordé sa confiance en acceptant de diriger ce travail : je me souviendrai toujours de ses précieux conseils, ses échanges fructueux, son soutien sans faille et ses encouragements constants durant ces années. Qu’il veuille trouver ici l’expression de mes sincères remerciements pour ses relectures méticuleuses de chacun des chapitres de cette thèse, pour sa rigueur et sa bienveillance.

Je tiens également à remercier les membres de jury pour l’intérêt qu’ils ont manifesté pour ce travail. Je tiens à exprimer toute ma reconnaissance à Eric WEHRLI, Lotfi ABOUDA et Estelle DUPUY d’avoir accepté de donner de leur temps pour évaluer ce travail.

Je remercie vivement les membres du laboratoire TIL et l’école doctorale LECLA qui m’ont soutenue dans les moments difficiles ainsi que pour leurs encouragements durant la dernière ligne droite de cette thèse.

Je ne saurai omettre d'exprimer toute ma gratitude à Frédéric SABIO qui, malgré la distance et ses charges professionnelles à l’Université d’Aix-Marseille, n’a cessé de s'intéresser à mon travail en me prodiguant des encouragements et des conseils judicieux. J’en profite pour dire Grand merci aussi à tous mes anciens collègues et étudiants pour la bonne humeur durant les années passées ensemble.

Évidemment je ne remercierai jamais assez toute ma famille (mes parents, mes sœurs et frères) pour le soutien sans faille qu’elle m'a assurée malgré la distance qui nous sépare. Notamment mon père, pour ses conseils sages et éclairés et ses encouragements,

ainsi que ma mère qui ne manque jamais à chaque occasion qui se présente, de s’assurer si je vais bien « avec la thèse ». Mes sœurs et frères qui, avec cette question récurrente, « quand est-ce que tu la soutiens cette thèse ? », bien qu’angoissante en période fréquente de doutes, m’ont permis de ne jamais dévier de mon objectif final.

Je remercie affectueusement ma belle princesse Amira de faire de moi la maman la plus heureuse et de m’aider à achever cette thèse malgré son bas-âge (elle saura comment…).

Enfin à tous ceux que j'ai connus et qui m'ont soutenu durant ces difficiles longues années, je dis merci. Je n'oublie personne, de l'inconnu d’un coin de la rue au hasard d’une brève rencontre qui m'adresse un sourire ou des mots réconfortants, aux plus proches amis.

Qu'ils veuillent tous trouver ici, l'expression de mes sincères remerciements.

Table des matières

Liste des tableaux

Liste des figures

Introduction

« Fondamentalement, l'ordinateur et l'homme sont les deux opposés les plus intégraux qui existent. » Gérard Berry (2015)

Motivation

La révolution créée par l'apparition du réseau mondial Internet a engendré une augmentation inexorable du nombre de documents en ligne accessibles et disponibles pour les utilisateurs. Paradoxalement, il est devenu de plus en plus difficile de rechercher l'information désirée dans ce foisonnement de documents. Cette profusion des ressources textuelles a, entre autres, induit le développement du Traitement Automatique des Langues Naturelles (TALN) qui a permis le perfectionnement des systèmes de diffusion de l'information nécessaires aux différentes organisations, tels que les laboratoires de recherche et les administrations. Ainsi, une impressionnante banque de données textuelles sous format électronique a vu le jour. Cette évolution technologique pose plusieurs défis aux entités organisationnelles qui doivent disposer d'outils capables de retrouver les informations appropriées. Les données textuelles de ces banques sont en constante évolution puisque la communauté de recherche ne cesse de les alimenter. En revanche, ces données sont la plupart du temps « brutes ». Ainsi, se pose la problématique suivante : comment récupérer, automatiquement et de manière rapide, les informations recherchées dans ces banques textuelles qui sont accessibles en ligne ?

Pour répondre à cette problématique, de nombreux travaux dans le domaine de la recherche d'information (RI) et de l'extraction d'information (EI) ont vu le jour ces dernières années et ont contribué à la mise en place d'outils facilitant l'accès à l'information textuelle. Néanmoins, malgré leur pertinence, l'utilisation de ces outils nécessite du temps, car elle implique évaluations et adaptations successives et continues. Dans les domaines de l'EI et la RI, ce sont les utilisateurs qui définissent leurs besoins en amont de la recherche spécifiée, mais les résultats ne garantissent pas un accès aux informations contenues dans les documents. Au vu du développement technologique actuel et de la place que prend l'informatique dans le domaine des sciences du langage, un enjeu adaptatif important s'avère être la communication entre l'Homme et la machine. Pour rendre les informations généralisées et accessibles à tous, il faut développer des outils permettant de numériser les données linguistiques, les modéliser et les traiter automatiquement. La structure profonde des documents doit de ce fait être analysée par ces outils afin d'en tirer les informations pertinentes.

Quelle que soit l'application du TALN, comme EI, RI, traduction, compréhension ou encore résumé automatique, le traitement et la résolution des anaphores, en d'autres termes l'identification de l'antécédent de chaque anaphore, est une tâche importante. L'utilisation

de l'outil informatique afin d'automatiser la résolution de l'anaphore, grâce à la désambiguïsation qu'elle apporte, apporte un avantage certain au vu de cette explosion de requêtes d'information. Il est intéressant de noter toutefois que, même si la production automatique de documents annotés en anaphores reste un enjeu d'actualité, cette branche n'a pas connu de réelles avancées les années passées, sur le plan théorique ou dans la création d'applications. C'est pour cette raison que nous avons choisi de travailler sur l'ambiguïté anaphorique et sa résolution automatique dans un corpus créé dans le cadre de cette thèse et permettant de traiter un certain nombre de difficultés propres aux textes à haute densité anaphorique.

Notions et concepts

Comme annoncé dans le titre de ce travail, notre recherche s'articule autour de l'ambiguïté anaphorique et sa résolution automatique. Nous éclairons dans cette section les grands traits définitoires des concepts de base, à commencer par la définition du mot clé d'anaphore. En rhétorique, l'anaphore indique un procédé cherchant de l'équilibre, d'insistance, ou autres, par « répétition d'un mot ou d'un groupe de mots au début de plusieurs énoncés ou syntagmes consécutifs » (Bonhomme 2005 : 64) et nous citons par exemple les emplois dans les discours politiques (Magri-Mourgues 2015). Syntaxiquement parlant, « un segment de discours est dit anaphorique lorsqu'il est nécessaire, pour lui donner une interprétation, de se reporter à un autre segment du même discours » (Ducrot/Todorov 1972 : 358).

Comment faut-il alors aborder la « matière anaphorique » ? Les approches classiques, nous le savons bien, proposent une réponse en termes de localisation dans le texte : pour trouver le référent d'une expression anaphorique, il « suffirait » de chercher son antécédent. Apparemment rien de plus simple que d'expliquer la référence du pronom *il* dans une séquence telle que :

[1] La semaine s'achève. Meursault a bien travaillé. C'est samedi, il retrouve Marie. (Résumé L'Étranger)

Il suffit de souligner qu'il s'agit d'un cas classique d'anaphore et que le pronom *il* ne fait que reprendre un référent déjà mentionné dans le contexte *(Meursault)*. Les travaux de ces dernières années ont toutefois clairement montré les limites et insuffisances d'une telle réponse, mais leurs contre-propositions, discordantes, ont conduit à un domaine en plein renouvellement théorique et méthodologique, avec des conceptions et des approches très diverses. Les

études, nombreuses, sur l'anaphore faites aussi bien en linguistique qu'en psycholinguistique, philosophie du langage, logique et intelligence artificielle, ont, de différentes manières, essayé d'aller plus loin dans la description et l'explication des stratégies et interprétations anaphoriques. Ce faisant, elles ont été amenées à bousculer des idées plus ou moins bien reçues sur le sujet et à poser de nouvelles questions à leur propos. Si l'accord semble être trouvé pour mettre au premier plan la dimension cognitive des processus anaphoriques, l'unanimité disparaît dès lors que l'on se tourne du côté des solutions proposées. Le renouvellement auquel nous assistons actuellement dans le domaine de l'anaphore, et, de façon plus générale, dans celui des expressions référentielles, met, en effet, aux prises des conceptions et des approches d'horizons et de tempéraments divers.

Il n'est pas exagéré de dire que Kleiber a donné du grain à moudre à toute une génération de linguistes, en France ou à l'étranger, en accordant un grand intérêt à l'expression anaphorique dans ses travaux. Selon lui (1994 : 7), « les processus anaphoriques constituent en réalité des mécanismes de construction référentielle discursive beaucoup plus complexes qu'il n'y paraît » et :

> Le principal problème que pose toute expression référentielle est, bien entendu, celui de la trouvaille du référent et, de préférence, celle du 'bon' référent. Le lieu de résidence de ce référent apparaît à cet égard comme un critère pertinent : si l'on sait où il est, on peut aussi, évidemment, le retrouver.

et c'est lors de cette quête du bon référent qu'une ambiguïté pourrait apparaître. Nous nous sommes intéressée alors au concept d'ambigüité qui est définit par Fuchs (2009) comme :

> insister sur le fait que les différents sens d'un constituant ambigu sont mutuellement exclusifs. Si c'est le sens A, ce n'est pas le sens B (et inversement) ; il faut donc nécessairement choisir entre les deux si l'on veut comprendre le message. En conséquence, ces différents sens donnent lieu à des représentations métalinguistiques distinctes et de même niveau. (Fuchs 2009 : 1)

Selon Gillon (1990, 2004), il existe trois formes différentes d'ambigüité, deux qui sont lexicales : soit polysémique, soit homonymique et une forme qui est structurale non lexicale comme dans l'exemple suivant :

> [2] Non contente d'avoir provoqué la chute de la maison Hulot, **la Cousine Bette** a décidé d'épouser le Maréchal Hulot, le frère ainé du baron. Elle espère ainsi obtenir une réussite sociale supérieure à celle de **sa cousine Adeline**, dont **elle** est toujours secrètement jalouse. (Résumé La Cousine Bette).

Ici, en effet, différentes structures peuvent correspondre :

> [2a] La Cousine Bette est jalouse de sa cousine Adeline.

[2b] Adeline est jalouse de la cousine Bette.

[2a] et [2b] ont d'ailleurs des interprétations différentes.

Il est important de noter que l'on peut se heurter dans certains cas à une anaphore ambiguë qui, faute de pouvoir être résolue, remet en question la compréhension textuelle, comme dans dans l'exemple [2] ci-dessus. Nous nous proposons donc ici d'identifier les mécanismes capables de résoudre cette ambiguïté. Ces mécanismes doivent être ainsi à la fois identifiables et réguliers. L'aboutissement de cette démarche est la création d'un outil informatique capable de résoudre automatiquement les anaphores pronominales.

Problématique

On peut regrouper les études qui traitent l'anaphore[1] en trois catégories principales : certaines considèrent la référence anaphorique comme phénomène complexe et tentent de mettre en place une terminologie afin d'identifier les expressions anaphoriques, d'autres s'intéressent à identifier les catégories des expressions anaphoriques et leur classification et d'autres enfin cherchent la résolution de l'anaphore. Suite à l'étude exploratoire de la bibliographie des concepts et notions liés au thème de l'anaphore, nous nous situons dans le prolongement de cette dernière catégorie afin de faire dialoguer les considérations théoriques et les résultats de l'analyse qui en découlent et de proposer un travail ouvrant la voie à un traitement automatique. Notre objectif scientifique général consiste à proposer un travail analytique en faveur d'une vision massivement textuelle du phénomène et nous envisageons une approche empirique développée à partir des théories conceptuelles soutenues par une linguistique de corpus pour créer un outil résolvant automatiquement les anaphores pronominales.

La plus grande difficulté dans ce projet apparait quand il existe une ambigüité anaphorique qui remet en question la continuité référentielle et rend difficile la compréhension d'un texte. Ainsi, quand une ambigüité existe, une solution serait de découvrir les mécanismes de relais entre les différentes expressions anaphoriques. Dans un cas d'ambigüité, nous supposons l'existence de mécanisme(s) capable(s) de préserver la continuité référentielle, ceci en suivant des règles objectives quant à la sélection de l'anaphore. En plus de l'étude de l'expression anaphorique, cette étude pluridisciplinaire implique l'étude de la cohésion et de la cohérence, mais aussi celle de la compréhension du

[1] Des études se sont aussi intéressées à la relation qui unit anaphore et deixis comme Kleiber (1991), Cornish (1995), Guillot (2007), Gardelle (2019).

texte ou encore de l'influence du contexte sur la sélection de l'anaphore. Cette dernière contrainte met en lumière les relations existant entre la cognition et le langage.

Outre la désambiguïsation de l'anaphore et l'étude de ses mécanismes, le but de ce travail est aussi la création d'un outil permettant de résoudre automatiquement les anaphores pronominales dans un corpus spécifiques dénommé $\text{RESUMAN}_{\text{C}}$. Nous relevons, quand il y a des cas d'ambiguïtés, les régularités de fonctionnement que l'on peut retrouver dans les chaînes anaphoriques. Afin de mettre en place notre stratégie, nous nous inspirons des théories de Hobbs (1989) et de Lappin/Leass (1994) et espérons nous distinguer de la littérature par la création d'un outil gratuit en se basant sur nos propres moyens : ce dernier, entraîné sur les textes bruts de notre corpus, procure un antécédent approprié à chaque pronom.

La subjectivité dans l'interprétation d'ambigüités apparaissant pour un humain à plusieurs niveaux, notamment au niveau syntaxique, peut devenir problématique en sémantique et en pragmatique. Il a été démontré, en utilisant des résumés de films, qu'il y a des différences d'interprétation selon les lecteurs, même si ces derniers sont linguistes (Charolles 2002). Ainsi, c'est la liste des antécédents possibles qui est variable d'un individu à l'autre, même dans le cas d'une lecture dirigée vers la recherche des ambiguïtés et des sous-déterminations : il est parfois difficile, voire impossible, d'attribuer un référent à une expression référentielle. L'exemple ci-dessous est révélateur des difficultés que l'on peut rencontrer :

> [3] Enfant naturel$_{?}$, **Candide$_{i}$** mène une existence heureuse dans cet univers idyllique : Le baron$_{j}$ et la baronne$_{k}$ de Thunder-ten-Tronckh possèdent en effet "le plus beau des châteaux". **Candide$_{i}$** est ébloui par la puissance de son$_{?}$ oncle$_{?}$, et par les sophismes lénifiants du docteur Pangloss$_{l}$, le précepteur$_{l}$. Il$_{?}$ admire également Cunégonde$_{m}$, la fille du baron$_{j}$. Tout bascule le jour des premiers ébats de **Candide$_{i}$** et de Cunégonde$_{m}$. La réaction du baron$_{j}$ est brutale, **Candide$_{i}$** est banni et chassé de cet$_{?}$ Eden. Il$_{i}$ se retrouve dans "le vaste monde".: **Candide$_{i}$** envoie Cacambo$_{n}$ racheter Cunégonde$_{m}$ au gouverneur$_{o}$ de Buenos Aires , tandis qu'il$_{?}$ ira l$_{?}$'attendre à Venise. (Résumé Candide)

Si nous considérons le pronom possessif *son*, celui-ci peut faire référence soit à Candide, soit au baron. Dans cet exemple c'est la lecture du roman dans son intégralité qui permet intuitivement de conclure que le référent est Candide. De la même façon, il y a une grande difficulté à identifier à qui réfère *il* et *l'* dans « tandis qu'il ira l'attendre à Venise ». En effet, on ne peut pas résoudre cette anaphore, puisque nous devons inclure dans le calcul des antécédents à la fois Candide et Cacambo, Cunégonde et le gouverneur de Buenos Aires.

En général, une série de coréférences, mise en place par des appartenances successives, est donnée quand un personnage apparaît à plusieurs reprises dans un texte,. Néanmoins, une expression référentielle peut correspondre, dans certains cas, à plusieurs référents en même temps qu'ils soient ambigus ou bien identifiés, et non à un seul référent. Il peut de ce fait y avoir une remise en question du référent tout au long de la lecture, ce que l'on peut comparer au phénomène des référents de type évolutif.

Dans le cadre de ce travail, l'ajout systématique d'une annotation aux pronoms et antécédents ne semble pas nécessaire, trois stratégies s'offrant à nous. Tout d'abord la stratégie de type linéaire ou encore linguistique qui ne considère pas les possibles ré-interprétations ultérieures et ne tient compte que de la forme linguistique. Dans une telle perspective, l'antécédent de *il* est Candide et celui de *l'* est (le) gouverneur de Buenos Aires. Ensuite, la stratégie appelée réaliste car elle tient compte des concepts : en ce sens, c'est après avoir intégré le sens du texte et effectué le calcul des références que l'annotation est effectuée. Selon cette stratégie, c'est Candide et Cacambo qui seraient les antécédents possibles. Enfin, la troisième stratégie est celle adaptée à l'étude du corpus. Elle permet de considérer les biais interprétatifs et les effets stylistiques : pour cela on doit élargir de manière ponctuelle à partir des concepts aux différentes interprétations. Selon cette stratégie, dans l'expression « il ira l'attendre (…) » nous procédons à une double annotation restreignant la référence et l'interprétation initiale (et temporaire) seulement aux antécédents Candide et Cacambo.

Devant de telles difficultés, les psycholinguistes se sont chargés de mettre en lumière l'existence d'au moins trois types de mécanismes qui peuvent remplir une fonction déterminante dans la levée des ambiguïtés :

- L'anticipation : le recours au contexte pour préparer l'analyse ultérieure met le sujet humain dans un état d'appréhension systématique de réception ; le sujet « s'attend » à un mot d'une certaine classe syntaxique, d'un champ sémantique précis.
- La polyvalence : Tout en sachant qu'il est aisé de définir les niveaux d'analyse d'un texte (syntaxique, sémantique, pragmatique), la principale caractéristique du système de compréhension humaine est sa capacité à exploiter une information de n'importe lequel de ces niveaux pour résoudre une difficulté à n'importe quel moment de l'analyse.
- Le traitement différé : à l'opposé, le système peut maintenir une ambiguïté aussi longtemps qu'il le faut et analyse « en parallèle » les diverses interprétations possibles, jusqu'à aboutir au choix de l'une d'entre elles.

Plusieurs réalisations informatiques récentes recourent aux mécanismes cités précédemment pour analyser les langues naturelles. Il est nécessaire d'admettre que même si des langages permettaient une implémentation plus rapide et plus pertinente de ces méthodes de traitement, on peut être confronté, comme c'est le cas dans les autres approches classiques de ce problème, à des difficultés dues essentiellement à leur mode d'application. Nous pouvons citer, pour expliciter ce qui a été avancé, l'exemple de l'anticipation. Au cas où ce à quoi « s'attend » le système de traitement automatique ne se produirait pas, que devrions-nous faire ? Si le programme refuse systématiquement la phrase, on obtient des systèmes défaillants, puisqu'ils ne seront capables de traiter qu'un sous-ensemble assez réduit en nombre de phrases « bien construites ». Si, dans le cas contraire, le programme accepte la phrase, toute la conception de l'analyseur doit être modifiée, et le mécanisme d'anticipation devient marginal, permettant au mieux de gagner un peu de « temps-machine », mais il ne pourra nullement constituer le principe organisateur du traitement. Nous sommes ainsi là face à une sorte de « dilemme » si tant est que cette expression soit appropriée à la situation. Cela dit, les méthodes classiques s'intéressent généralement aux phénomènes locaux et en conséquence n'engendrent pas chez l'annotateur un effort de réflexion ou des connaissances approfondies. Ainsi, les annotations utilisant des méthodes classiques, comme en morphologie, ne permettent pas d'aboutir à une solution. Il est nécessaire, dans les linguistiques de corpus, de donner des traits (ou propriétés, ou annotations) à des entités textuelles préalablement identifiées : nous utiliserons ainsi dans notre modélisation la troisième stratégie, celle du traitement différé, qui présente de plus deux avantages. Tout d'abord, elle permet de considérer la réalité de référence, ce qui présente un avantage certain pour le traitement automatique. Ensuite, elle prend en compte le mode d'introduction des personnages et d'écriture du texte. Néanmoins, cette stratégie a un défaut majeur : elle impose une lecture quasi intégrale du texte source du résumé préalablement à l'annotation de ce dernier. Elle rend ainsi difficile la méthode d'annotation, car elle doit lister les cas où une interprétation temporaire est nécessaire dans un manuel d'annotation.

Il existe cependant des techniques et des méthodes numériques, utilisées par les professionnels du TAL, qui ne nécessitent pas une analyse poussée ou une compréhension en profondeur du texte source pour annoter le corpus. Ces dernières ont aussi joué un rôle dans la mise en place d'approches pour résoudre l'anaphore de manière automatique. Ces

applications offrent deux avantages majeurs, la souplesse dans les traitements et une vitesse de productivité accrue.

Dans une activité d'annotation, le but est d'identifier ce que peut apporter un raisonnement quand celui-ci inclut des informations de natures différentes. Nous prenons le parti de travailler sur la problématique de l'anaphore pronominale qui, malgré qu'elle soit la plus étudiée, reste source de dissensions. Quand on s'intéresse aux différents travaux traitant de la linguistique de corpus, on observe qu'il n'y a peu de travaux portant sur l'annotation automatique de l'anaphore pronominale, dont l'origine pourrait être un malaise des linguistes travaillant précisément sur corpus. Nous remarquons que les études théoriques et qualitatives portant sur l'anaphore ambiguë, malgré leur aspect fort intéressant, doivent être réorganisées afin de pouvoir les valider et les reproduire. Le présent travail permet d'expliciter nos conceptions personnelles et d'illustrer nos prises de position en focalisant notre attention sur les relations unissant anaphore et emplois textuels dans le cadre de l'analyse d'un corpus en ligne rassemblant un ensemble de résumés d'œuvres littéraires.

Corpus

Etant donné que le traitement automatique de l'anaphore pronominale dans des textes écrits en français constitue l'axe de ce travail, nous proposons comme corpus des résumés écrits en ligne sur le site www.alalettre.com[2]. Cette combinaison nous a inspiré la nomination $RESUMAN_C$[3] pour désigner notre corpus d'étude. Le corpus sur lequel nous avons choisi de travailler est de type électronique. Ceci a deux avantages, tout d'abord, il donne la possibilité d'effectuer une étude anaphorique linguistique, ensuite, il nous permet de mettre en place une application numérique afin de résoudre l'anaphore. Ce corpus est composé de 82 résumés électroniques de classiques de la littérature française. Nous avons analysé les textes bruts grâce à deux outils : le concordancier AntConc[4], et l'analyseur morphosyntaxique Fips[5]. Ceux-ci permettent d'explorer et d'analyser le corpus.

---

[2] http://www.alalettre.com/oeuvres.php

[3] Nous utilisons l'indexe $_C$ pour désigner notre corpus. Ce choix est fait pour différencier le nom de notre corpus ($RESUMAN_C$) de celui de notre outil ($RESUMAN_O$).

[4] Laurence Anthony (Faculté des Sciences et de génie, Université de Waseda, Japon) est le professeur qui a développé *AntConc*. Ce logiciel est aisément téléchargeable à partir du site de son auteur (http://www.antab.sci.waseda.ac.jp/softxare.html) et tourne sous Windows et autres systèmes d'exploitation comme Macos X et GNU/Linus. Il ne demande plus d'installation, une double clique sur le fichier exécutable (.exe) permet de le lancer. Antconc est un logiciel concordancier qui permet de traiter des structures textuelles de langues différentes y compris celles asiatiques ; il traite aussi des textes bruts ou annotés à partir des fichiers de différentes extensions (.txt, .xml ou .html) et contenant des divers discours, notamment dans le

Au vu du corpus que nous utilisons et de la présence possible d'ambiguïtés, nous choisissons de nous intéresser aux chaînes d'anaphores où ce sont les pronoms de la 3ème personne qui interviennent (*il, ils/elle, elles*). Cette stratégie limitative permet de réduire les faux sens et la multiplicité des ambigüités. De plus, ce type d'analyse est assez fastidieux et nécessite de sélectionner des procédés pertinents de résolution d'anaphores.

Dans ce type de texte, on peut se heurter dans certains cas à une anaphore ambiguë et si cette dernière ne peut pas être résolue, alors c'est la compréhension textuelle qui est remise en question. Un objectif majeur du travail est ainsi l'identification de mécanismes capables de résoudre cette ambiguïté, ces mécanismes devant être, comme nous l'avons vu précédemment, à la fois identifiables et réguliers. Grâce à ce corpus, nous aurons une base afin d'explorer et d'étudier les anaphores ambigües et aussi d'extraire les anaphores pronominales. L'objectif principal étant la mise en place d'une application permettant de résoudre automatiquement ces dernières, deux tâches distinctes doivent être réalisées à partir du corpus. La première consiste à s'appuyer sur des exemples identifiés dans RESUMAN_C pour construire une étude sur corpus (on part en premier lieu de la linguistique sur corpus), la seconde permet de quantifier et de généraliser les résultats à d'autres textes et elle consiste à placer le corpus au centre de l'étude (on finalise par la linguistique de corpus).

Présentation du plan

Notre étude, répartie selon trois axes principaux, sera focalisée sur deux volets qui s'entrecroisent : le volet linguistique et le volet informatique. Nous partons des dimensions textuelles pour arriver aux dimensions automatiques et spécifiquement à la création d'un outil informatique, ceci après avoir défini ce que représente une ambigüité anaphorique. Notre thèse est divisée en trois parties :

La première présente les fondements théoriques qui ont servi à l'analyse de l'anaphore et au choix de notre corpus RESUMAN. Dans le premier chapitre, nous commençons par une réflexion sur la notion de référence, en mettant l'accent sur la nécessité de rendre compte de l'étroite relation entre la référence et l'anaphore. Nous

---

domaine de la diddactique. Il est, par ailleurs, largement utilisé dans le domaine de la linguistique, surtout pour l'analyse du discours ou l'analyse des corpus.

[5] Analyseur syntaxique multilingue (Laenzlinger et Wehrli 1991 ; Wehrli 1997, 2004)1 développé au LATL2 de l'université de Genève, sur la base de grammaires inspirées des théories chomskyennes. Il est accessible via ce lien : http://latlapps.unige.ch/Parser. Pour plus de détails : http://alpage.inria.fr/iwpt09/atala/fips.pdf

replaçons, ensuite, la notion d'anaphore dans son contexte linguistique (définition, types et approches). Dans un énoncé, l'anaphore est centrale, en effet sa résolution est essentielle à sa compréhension, c'est la raison principale qui a déterminé notre objet d'étude. Ainsi, après avoir défini les principales caractéristiques linguistiques de l'anaphore, nous exposons dans le deuxième chapitre les principales caractéristiques de la tradition de corpus et les différentes phases de l'élaboration de RESUMAN. Nous terminons par une présentation de la méthodologie utilisée dans notre corpus.

Tout au long de cette étude nous plaçons la résolution de l'anaphore au centre de la compréhension d'un texte. C'est pour cette raison que dans la deuxième partie, nous étudions les dimensions textuelle (premier chapitre) et automatique de l'anaphore (troisième chapitre) : notre choix d'étudier l'anaphore se justifie par le rôle essentiel et organisateur de celle-ci dans un texte pour maintenir sa cohérence et sa cohésion. En effet, résoudre correctement une anaphore représente une étape importante pour la compréhension. Le concept de l'ambiguïté étant central dans notre thèse, nous consacrons le deuxième chapitre aux facteurs de la résolution de l'anaphore pronominale en nous appuyant sur la théorie de l'Accessibilité d'Ariel (1990). Dans cette étude, nous partons du postulat de base que la continuité référentielle est maintenue dans la quasi-totalité des chaînes (sauf occurrences d'ambiguïté référentielle évidente). La question qui nous intéresse est de savoir comment, à travers les expressions anaphoriques utilisées, cette continuité référentielle est assurée. Nous allons alors observer comment se réalisent, dans RESUMAN_{C}, les relations anaphoriques, afin de découvrir comment le choix de l'expression permet la désambiguïsation des chaînes anaphoriques ambiguës.

Dans la troisième partie, nous répondons aux interrogations qui surviennent quand notre méthodologie de résolution automatique se met en place. Pour ce faire deux problématiques sont centrales : quelle est la solution pour que les méthodes classiques soient plus affinées et mieux exploitées et comment la linguistique de corpus influence-t-elle notre conception de l'anaphore et de sa résolution. Dans le premier chapitre, nous dressons un résumé historique de l'évolution de l'Intelligence Artificielle et son interaction avec la linguistique. Nous exposons ensuite, dans le deuxième chapitre, un descriptif panoramique des systèmes de résolution automatique de l'anaphore. Nous consacrons un troisième chapitre à la stratégie adoptée et aux dimensions retenues dans la phase d'élaboration de notre outil. Nous faisons remarquer que notre approche ne diffère pas de

façon significative des systèmes symboliques (reposant sur des connaissances linguistiques) cités. Cependant, notre système a quelques caractéristiques intéressantes. Le fait par exemple qu'il soit mis en œuvre en utilisant un outil unique pour l'analyse syntaxique et morphologique et l'élimination des pronoms impersonnels (cela démontre par ailleurs la puissance expressive de Fips) et le fait qu'il utilise un ensemble de préférences ordonnées pour choisir le meilleur antécédent parmi un ensemble des candidats. En prenant en compte différents facteurs syntaxiques et cognitifs, notre algorithme recourt à un modèle permettant d'évaluer d'une manière efficiente le poids d'un antécédent potentiel. Ces facteurs comportent chacun un indice différent en fonction de leur utilité dans la résolution. Cette dernière propriété permet une évaluation indépendante de chaque préférence, offrant ainsi une meilleure compréhension du processus de résolution des pronoms. Notre objectif étant de proposer une approche de résolution automatique d'un phénomène linguistique, l'intervention d'un informaticien était indispensable. C'est au sein de notre laboratoire de recherche TIL que nous allons pouvoir créer notre algorithme, écrit en langage de programmation logique et exécuté à l'aide d'une interface réalisée en Java[6].

Nous terminons notre thèse par une évaluation de la performance de l'outil : l'identification du référent à travers son expression anaphorique en cas d'ambiguïté serait-elle possible par notre schéma d'analyse linguistique (fonction morphosyntaxique du terme mentionné, catégorie et fonction syntaxique, poids de saillance) ? Et quelles sont les limites et les perspectives d'amélioration de notre outil ?

[6] Le travail informatique sera préparé en collaboration avec M[lle] Wided SELMI (doctorante en Informatique à l'Université de Sfax).

Première Partie
Cadre théorique

[...] nous nous inscrivons dans une linguistique cumulative, qui pense que l'on peut et que l'on doit progresser à partir des acquis antérieurs. Modestement sans doute, à petits pas le plus souvent. Penser et dire que tout a déjà été dit, ce qui est à traduire en fait dans notre tradition par Aristote a déjà dit... ou c'est déjà chez Aristote, n'a jamais fait pousser les tomates, – c'est un jardinier qui parle ! Inversement, comme l'habitude venue d'outre-Atlantique tend hélas à se répandre de plus en plus, ignorer les prédécesseurs, pour ne pas avoir une vue corrompue des choses, – c'est le prétexte généralement avancé –, et (re)construire à neuf avec sa propre truelle, son propre mortier et ses propres briques n'est pas la meilleure garantie d'une architecture qui soit nouvelle et, surtout, qui tienne debout. Il ne suffit pas d'un cheval et d'un regard de cow-boy béat pour découvrir de nouvelles terres vierges dans des paysages archi-fréquentés. Qui veut éviter le fossé en fermant les yeux est bien souvent le premier à (re)tomber dedans. C'est encore le jardinier qui vous le dit ! Cela ne signifie pas qu'il faille tout lire – tout le monde sait que c'est impossible et que ce n'est pas souhaitable – mais que l'on revienne peut-être un peu plus à cet état de la question qui servait de passeport d'entrée aux thèses et aux monographies d'antan. **(**Kleiber 1994 : 19)

Nous consacrons cette première partie à la littérature des concepts de base de ce travail afin de mieux nous positionner au sein des multiples études traitant l'anaphore. Dans notre étude, nous prenons comme support de travail un corpus constitué des résumés des œuvres littéraires françaises, $RESUMAN_{C,}$ pour repérer les ambigüités anaphoriques et y étudier leur fonctionnement.

Nous replaçons, dans le premier chapitre de cette partie, l'anaphore dans la problématique générale de la référence, de la coréférence et du texte, en mettant en perspective certains domaines et approches théoriques qui s'y intéressent. Ainsi, après avoir introduit certaines approches se rapportant à la philosophie du langage et au domaine de la logique, nous les associons par la suite aux approches linguistiques traitant de l'anaphore. Nous nous focalisons spécifiquement sur la caractérisation linguistique de la référence anaphorique, grâce aux études antérieures. Nous mettons aussi en exergue dans ce chapitre les relations associant l'anaphore au texte.

La sélection des résumés électroniques qui constituent le corpus utilisé dans notre travail est l'objet du second chapitre de cette partie. Nous y présentons la méthodologie que nous employons quant aux choix des relations et des phénomènes anaphoriques traités, du genre textuel et enfin des outils informatiques utilisés. Ce repérage s'avère le support à travers lequel nous tentons de résoudre[7] automatiquement les ambiguïtés anaphorique présentes dans $RESUMAN_C$.

[7] Notre approche proposée sera présentée au troisième chapitre de la dernière partie de ce travail.

Chapitre 1. L'anaphore : éléments de définitions, classification et approches linguistiques

De nombreux travaux en linguistique se sont intéressés à la question de l'anaphore en la traitant sous différents angles (processus interprétatifs et mécanismes référentiels mis en jeu, diachronique (moyen français[8]) et synchronique). Cette notion a connu de multiples définitions qui ont évolué et varié selon les auteurs et les approches. Étymologiquement, l'anaphore est un élément qui reporte en arrière. Un des premiers auteurs à en avoir proposé une définition est le grammairien grec Apollonius Dyscole. Selon lui, l'anaphore, une connaissance seconde mise en œuvre par l'esprit, s'oppose à la deixis, une connaissance première référant à ce qui n'est pas signalé par nos sens et qui est intériorisée. Il distingue ainsi les déictiques -se référant directement aux objets- et les anaphoriques -indirects- c'est à dire qui ne les désignent que par d'autres segments de discours. A partir de cette définition, des auteurs contemporains ont diversifié les approches et les aspects référant à ce domaine et ont vérifié le rapport entre les deux termes se rapportant à la problématique de l'anaphore. L'objectif premier de cette partie est de porter dans ce contexte l'attention sur certaines notions et théories sous tendant notre analyse des anaphores pronominales.

Si nous parlons des ambigüités anaphoriques, entre autres de leur identification et leur fonctionnement, il est important d'accorder de l'intérêt à la référence, la coréférence et l'anaphore (Sections 1-3). Le point de vue adopté dans notre thèse étant l'anaphore conçue comme un dispositif lié à la structuration, la progression et la dynamique textuelles, notre propos aborde, dans un deuxième temps, le fonctionnement phrastique de l'anaphore. En nous inscrivant dans une théorie de l'anaphore dans son fonctionnement textuel, nous rejoignons le point de vue des conceptions sémantique, cognitive et pragmatique affirmées par **Kleiber**, et **Charolles** qui préconisent le cadre d'une analyse de l'anaphore à partir de diverses approches et par plusieurs biais. Les diverses approches fixent notre cadre théorique (Section 4).

1. Classification et définitions de l'anaphore

La notion d'*anaphore* constitue la clé de voûte de notre étude. Pourtant, elle forme une unité de langue paradoxale : d'un côté, elle constitue une notion familière ; d'un autre,

[8] Comme : Perret (2000), Dupuy (2006), Combettes (2007) et Fournier (2008).

dès qu'il faut expliquer de manière précise ce qu'est une anaphore, les difficultés surviennent. L'anaphore est définit comme :

> un phénomène de dépendance interprétative de deux unités, dont la première, à laquelle se reporte la seconde, l'anaphorique, est appelée «interprétant» (Ducrot/ Todorov 1972)

Il y a des acceptions variées pour le terme *anaphore* car il est utilisé dans différents contextes : on peut y voir soit une notion rhétorique impliquant la répétition d'un syntagme potentiellement indépendant : citons comme exemple Roux, dans son introduction aux Actes de l'Atelier sur l'Anaphore[9], qui admet qu'une anaphore est un :

> procédé visant à un effet de symétrie, d'insistance, par répétition d'un même mot ou groupe de mots au début de plusieurs phrases ou propositions successives. (Roux 1987 : 7)

ou une notion linguistique, notion fort commode quand on l'évoque de loin chez les linguistes, fort suspecte dès qu'on l'examine de près. Il suffit d'ouvrir n'importe quel ouvrage linguistique traitant de l'anaphore pour être frappé par l'hétérogénéité des critères qui servent à la caractériser :

> il n'y a pas d'expression qui soit spécialisée pour la référence anaphorique. En effet, les anaphoriques constituent non seulement une classe hétérogène d'expressions (pronoms, SN démonstratifs, possessifs, adverbes, temps grammaticaux, etc.) − ce qui peut déjà surprendre − mais les expressions rassemblées sont également − ce qui est plus étonnant − des expressions dont (presque) aucune n'est destinée à être uniquement anaphorique. » (Kleiber 1992 : 2)

Certains auteurs ont parlé de la difficulté de l'identification des catégories d'anaphore. Citons le cas de Corblin (1987 : 35) qui voit qu' *:*

> il est fort difficile d'identifier la classe des groupes nominaux anaphoriques et celle des pronoms ; aucun de ceux qui considèrent la notion d'anaphore pour elle même ne parvient en fait à s'y résoudre. Il y a en effet des candidats au titre d'anaphorique aussi bien parmi les positions moins spécifiées que les pronoms (positions nulles), que parmi les positions plus spécifiées que les pronoms (groupes nominaux à la tête N définis et démonstratifs).

Face à une telle inflation de critères, nous avons examiné, tout d'abord, le statut de l'anaphore dans les livres de grammaire où nous n'avons pas rencontré de définition explicite de la notion. En effet, en se basant essentiellement sur ses propriétés grammaticales, les auteurs de ces ouvrages dressent un catalogue de différents dispositifs

[9] Atelier de linguistique, SAES Bordeaux, 1987.

anaphoriques et ne proposent pas de définition unitaire de l'anaphore. Nous pouvons citer, à titre d'exemple, l'ouvrage de Grevisse/Goosse (1995). Dans cet ouvrage consacré à la grammaire française, alors que son index dépourvu de référence à la notion d'anaphore nous a surprise, nous pouvons trouver un chapitre entier de 38 pages détaillant les différentes fonctions grammaticales de tous les types de pronoms (personnels, indéfinis, interrogatifs, démonstratifs, possessifs, numéraux et relatifs). Les auteurs distinguent les pronoms reprenant un terme qui est présent dans le contexte (leur antécédent) ou *pronoms représentants* de ceux dépourvus d'antécédent ou *pronoms nominaux*.

Dans l'exemple :

[1] Paris, automne 1819. Dans une pension miteuse de la rue Neuve-Sainte Geneviève, la maison Vauquer (du nom de sa tenancière), se côtoient des pensionnaires et des habitués du quartier **qui** ne viennent **y** prendre que le dîner. (Résumé Le père Goriot)

qui et *y* sont des représentants qui renvoient à des éléments qui font partie du contexte, respectivement aux *pensionnaires* et *des habitués du quartier* et à *une pension miteuse de la rue Neuve-Sainte Geneviève*. *Personne*, dans l'exemple 2, n'est pas un pronom nominal :

[2] Lorenzo avertit les républicains qu'il va bientôt tuer Alexandre. **Personne** ne veut le croire. (Résumé Lorenzaccio)

Le contexte impliqué dans la définition des pronoms représentants se réduirait, bien que les grammairiens ne précisent pas la notion de contexte, au contexte interne du discours. Prenons comme exemple les pronoms *je* et *tu* qui sont, grammaticalement parlant, des pronoms nominaux selon Grévisse/Goosse parce qu'ils renvoient à l'interlocuteur[10] et n'ont pas d'antécédents. Les linguistes décrivent ces pronoms comme déictiques ou exophoriques puisqu'ils trouvent généralement leurs références dans la situation d'énonciation du discours. Ainsi, nous pouvons penser que sur la base des pronoms *je* et *tu*, la distinction grammaticale entre pronoms représentants et nominaux renvoie à la distinction linguistique entre pronoms anaphoriques et déictiques. En se fondant sur la distinction entre la partie extra-discursive ou intra-discursive du contexte dans lequel se trouve leur référent, le déictique (déixis) est alors opposé à l'anaphorique[11].

[10] L'interlocuteur étant un élément appartenant à la situation de production du discours.

[11] Nous ne présenterons pas ici la dichotomie – anaphore *vs* deixis – et nous renvoyons aux travaux de Wiederspiel (1989), Kleiber (1991), Corblin, (1995 : 24-25) et Perdicoyanni-Paléologou (2001). Cette dichotomie ne rentre pas directement dans notre cadre de recherche.

Nous pouvons noter que les grammairiens, contrairement à certains linguistes, admettent qu'un pronom ait un autre pronom comme antécédent. C'est-à-dire que, dans l'exemple (3) ci-dessous, le pronom *la* ait comme antécédent le pronom *lui* et non pas le substantif *Marianne*. En effet, afin de déterminer les possibilités d'accord avec le verbe, c'est généralement les propriétés du nom (genre, nombre et personne de l'antécédent) transmises aux différents pronoms que présentent les grammairiens dans les ouvrages. Grévisse/Goosse indiquent dans leur ouvrage (1995 : 205), les règles d'accord en genre et en nombre des pronoms quand l'antécédent est un nom ou un autre pronom.

Exemple *:*

> [3] **Marianne**, **elle**, plus romantique, **se** laisse éblouir par le séduisant sir Willoughby. Ce Dom juan d'opérette **lui** fait de belles promesses avant de **la** rejeter d'une manière brutale et peu cavalière. (Résumé Raison et sentiments)

Dans un résumé, nous faisons souvent référence de façon répétitive à un même personnage, événement, action ou fait. Dans l'exemple précédent, on parle de *Marianne*. A chaque fois que l'on veut en dire quelque chose, on utilise un pronom afin d'assurer la cohérence du résumé et éviter la répétition inutile d'informations (***Marianne****,elle,…se.…..lui.……..…la.…..*). Pour la reprise d'éléments mentionnés précédemment dans un discours, la relation d'anaphore est mise en jeu par l'usage de pronoms. C'est quand l'interprétation d'une unité lexicale nécessite la présence de l'autre qu'il existe une anaphore entre les deux. De plus[12], le pronom est considéré comme neutre et renvoie à l'antécédent quand l'antécédent n'est pas un nom ou un autre pronom. Ainsi, un pronom peut référer à *l'idée* véhiculée par l'antécédent et ne renvoie pas forcément à l'antécédent identifiable à la surface du discours. Même si les auteurs ne l'indiquent pas explicitement, ce serait donc le sens qui serait transmis de l'antécédent au pronom. Et c'est sur ce point (le transfert de sens) que les linguistes insistent particulièrement pour caractériser l'anaphore. Ainsi, selon Reichler-Béguelin (1988), il y a une dépendance interprétative entre un segment d'énoncé appelé *antécédent* (ou *interprétant, référé, référent, source, contrôleur de l'anaphorique ou encore source sémantique* (Perdicoyanni-Paléologou 2001)) et un autre segment, l'*anaphorique*, généralement un pronom ou un syntagme nominal défini ou démonstratif, qui, s'il n'est pas mis en connexion[13], ne peut avoir un sens référentiel. Dans l'exemple :

---

[12]On rencontre aussi la désignation *anaphore* pour désigner non pas la relation mais l'unité anaphorique.
[13] *Cf.* définition (4) de Wiederspiel (1989 : 99).

[4] Le vieil demande à **Lorenzo** de délivrer Florence d'Alexandre de Médicis. **Lorenzo** évoque alors son destin. Lorsqu'**il** était jeune, **il** était bon et pur, puis un jour fatal, **il** a promis de libérer la patrie de ses despotes. Investi de cette mission, **il** a souhaité tuer le pape, mais **il** n'en a pas eu le temps. **Il** s'est alors infiltré dans l'entourage d'Alexandre de Médicis, c'est pour cela qu'**il** fut obligé de devenir son complice de débauche. Nul ne peut agir sans se compromettre et se salir les mains. (Résumé Lorenzaccio)

nous remarquons l'existence d'une relation liée à la référence entre deux unités où l'interprétation de la deuxième dépend étroitement de la première, au point qu'on peut la qualifier, comme le fait Milner (1982 : 18), de « répétition » (*Lorenzo…Lorenzo…il..il…etc*). C'est un des aspects centraux de la notion d'anaphore. C'est sur la capacité à faire apparaître l'antécédent (un segment du contexte) comme répondant aux conditions d'interprétation qu'exige l'anaphorique (un autre segment) que reposerait cette relation.

La relation linguistique entre *l'entité anaphorique* (appelé aussi *anaphore*) et *l'antécédent* (une autre unité du cotexte) représente le champ d'étude privilégié de la linguistique. Cette définition de l'anaphore, illustrée par l'exemple (5), dans son usage le moins technique et assez simple, est loin d'unir la communauté linguistique.

[5] La première fois qu'Aurélien$_i$ vit Bérénice, il$_i$ la trouva franchement laide. (Résumé Aurélien)

où le pronom *il* est anaphorique et réfère au groupe nominal *Aurélien*.

Des utilisations terminologiques abusives peuvent être sources de confusions néfastes à l'explication de la relation anaphorique. Ainsi, il convient tout d'abord de commencer par Ducrot/Todorov qui la définissent en termes d'interprétation :

> un segment de discours est dit anaphorique lorsqu'il est nécessaire pour lui donner une interprétation de se reporter à un autre segment du même discours. (Ducrot/Todorov 1972 : 358)

D'autres notions, notamment la cataphore et l'endophore, sont essentielles à la compréhension de cette théorie. Si la notion d'anaphore a été connue depuis le II^{è} siècle, la cataphore, quant à elle, est une notion relativement récente qui a été traitée par d'importants linguistes des années 1970 et 1980 (Maillard 1974, Halliday et Hasan 1976, Kesik 1989, etc)[14]. La cataphore est définit comme l'anticipation d'un segment d'une réalité non encore énoncée ou le renvoi au contexte subséquent (Halliday et Hasan 1976, cité par Kesik 1989 : 65), le terme générique désignant le phénomène général de la

[14] Les travaux récents (Lefebvre 1991, Combettes 2001, Bodelot 2004, Dupuy 2013, Kesik 2014, etc) partent des définitions avancées par ces précurseurs.

dépendance contextuelle abstraction faite de la position de l'antécédent tel que l'énonce Maillard (1974). Quand on s'intéresse spécifiquement à l'endophore, on peut noter qu'elle renferme aussi les notions d'anaphore et de cataphore. La première renvoie à une relation de présupposition de l'endophorique avec des « cibles » du texte précédent et la deuxième à la relation de l'endophorique avec le texte ultérieur.

L'anaphore relève elle-même de l'endophore en tant que mécanisme plus général. Grâce à l'environnement textuel d'un signe (ou d'un ensemble de signes, comme dans le cas d'un SN), on peut identifier son référent en utilisant ce dernier mécanisme. Quand c'est à l'environnement textuel antérieur à ce signe que l'élément appartient, on parle d'anaphore et quand c'est à l'environnement textuel postérieur à ce signe que l'élément appartient, on parle de cataphore. Prenons l'exemple suivant où il y a anaphore puisqu'*il* n'y est interprétable que par renvoi à sa source *Julien Sorel* :

> [6] **Julien Sorel**, ambitieux, rêve de gloire et s'évade dans la littérature. **Il** puise son imagination dans les Confessions de Rousseau, Les Bulletins de la Grande Armée, et Le mémorial de Sainte Hélène. (Résumé Le Rouge et le Noir)

Par ailleurs, nous avons une cataphore dans l'exemple :

> [7] Alors qu'**ils** se rendent chez les Pazzi, **Pierre et Thomas** sont arrêtés par un officier allemand et conduits en prison. (Résumé Lorenzaccio)

En effet, ce n'est que par renvoi à l'énoncé *Pierre et Thomas* que *ils* est interprétable. Nous illustrons ces relations sémantiques dans le schéma suivant :

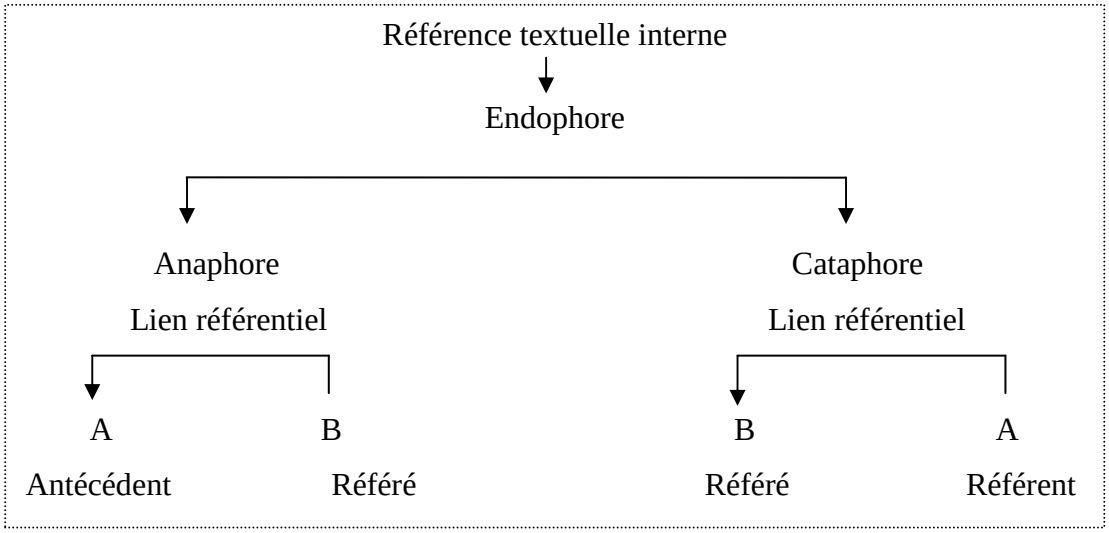


Figure 1 : Relations sémantiques référentielles[15]

[15] Dans le cas de cataphore, nommer l'entité A par antécédent nous a paru paradoxale. Nous préférons utiliser le terme *référent* au lieu *d'antécédent* pour conserver la distinction entre les deux notions *anaphore* et *cataphore*.

Après avoir mis l'anaphore dans son cadre textuel général, nous allons examiner quelques définitions de celle-ci. Nous avons constaté, à travers nos lectures, qu'il n'existe pas de consensus définitif sur la classification selon les unités linguistiques de l'anaphore. Par ailleurs, les auteurs ne se s'accordent pas sur la délimitation de la catégorisation des reprises anaphoriques et sur les critères de distinction de l'anaphore. En effet, l'étude de la définition de l'anaphore implique, notamment, deux positions qui s'opposent : certains la définissent à partir de sa fonction dans la phrase et d'autres, à partir de sa position dans les enchaînements souvent au-delà de la phrase[16]. Nous citons, dans ce qui suit, entre autres, des définitions évoluées dans le temps pour montrer la divergence notionnelle et définitoire de l'anaphore.

1) Il y a relation d'anaphore entre deux unités A et B quand l'interprétation de B dépend crucialement de l'existence de A, au point qu'on peut dire que l'unité B n'est interprétable que dans la mesure où elle **reprend entièrement ou partiellement** A. (Milner 1982 : 18)

2) Il me semble ainsi que le plus souvent, l'anaphore se ramène pour l'essentiel à **un simple passage à l'hyperonyme** (c'est en cela qu'on peut y voir, par rapport à la répétition intégrale, un procédé d'économie) (...) (Berrendonner 1983 : 236)

3) il est **fort difficile d'identifier la classe des groupes nominaux anaphoriques et celle des pronoms** ; aucun de ceux qui considèrent la notion d'anaphore pour elle même ne parvient en fait à s'y résoudre. Il y a en effet des candidats au titre d'anaphorique aussi bien parmi les positions moins spécifiées que les pronoms (positions nulles), que parmi les positions plus spécifiées que les pronoms (groupes nominaux à la tête N définis et démonstratifs). (Corblin 1987 : 35)

4) L'anaphore se caractérise (...) comme un **phénomène de rappel informationnel** relativement complexe où sont susceptibles d'intervenir :

- le savoir construit linguistiquement par le texte lui-même ;
- les contenus inférentiels qu'il est possible de calculer à partir des contenus linguistiques pris pour prémisses, et cela grâce aux connaissances lexicales, aux prérequis encyclopédiques et culturels, aux lieux communs argumentatifs ambiants dans une société donnée. ([Reichler]-Béguelin 1988 : 18)

[16] Cf. Milner (1982 : 363) : «La notion traditionnelle est celle d'une relation entre deux termes. On distingue de ce point de vue entre une anaphore libre, qui est insensible aux contraintes du sujet spécifié et des phrases finies, et une anaphore liée, qui est sensible à ces contraintes. La première relation ressortit au discours en tant qu'il excède les limites de la phrase. La seconde relation ressortit exclusivement à la phrase : ainsi, un pronom usuel tel que *il* peut avoir un antécédent situé dans une phrase distincte, ou même une réplique différente dans un dialogue. En revanche, le réfléchi ne peut avoir pour antécédent qu'un terme situé dans la même phrase.»

5) **une connexion anaphorique** qui met en relation un terme supérieur, l'antécédent, qui communique une valeur à un terme inférieur, qui n'est autre que l'élément anaphorique. » (Wiederspiel 1989 : 99)

6) On parle d'anaphore **lorsqu'un terme, souvent un pronom de troisième personne**, est utilisé pour reprendre une autre expression nominale qui le précède, appelée traditionnellement son antécédent, à laquelle il emprunte sa référence, c'est-à-dire l'objet qu'il désigne. (Moeschler/Reboul 1994 : 523)

7) La notion d'anaphore semble en fait utilisée pour distinguer **la nature du lien** qui **unit *il* au terme qui le précède** dans une chaîne de celui qui unit je et un nom propre à un éventuel prédécesseur dans une chaîne. (Corblin 1995 : 27)

8) L'anaphore est une **relation structurelle, définie comme asymétrique, intransitive, non réflexive** entre deux segments textuels dont l'un (l'anaphorisant) est dépendant de l'autre (l'anaphorisé) (Corblin 1985, 1989). (Perdicoyanni-Paléologou 2001 :57)

9) Nous définissons l'anaphore comme **une forme vicariante sémantiquement vide ou incomplète**, et qui est en même temps, d'un point de vue fonctionnel, une instruction explicite ou implicite visant à ce qu'on aille chercher dans le contexte gauche le matériel lexical (appelé *antécédent*) nécessaire à la reconstruction du syntagme qu'elle remplace. (Le Pesant 2002 : 39)

10) L'interprétation du lien anaphorique a successivement ouvert la voie à deux conceptions : **une version traditionnelle** où le texte est pris comme élément central de définition, puis **une version mentaliste** fondée sur des facteurs cognitifs donnant à la mémoire un rôle prépondérant. (Kara & Wiederspiel 2011 : §4)

11) Pour nous le moment est venu de tirer quelques conclusions de ce parcours définitoire. Des conclusions négatives tout d'abord. **II n'y a pas de catégorie générale unitaire stable de l'anaphore**, parce que:
(i) l'obligation de prendre en compte ä un moment donne de l'analyse les propriétés particulières des différents types d'expressions conduit soit ä postuler une catégorie de l'anaphore en des termes qui ne retiennent plus comme critère indispensable la présence d'un élément identificateur dans le contexte linguistique, soit ä faire éclater la catégorie anaphorique en une diversité de procédures de référence textuelle particulières,
(ii) parce que les paramètres de l'incomplétude et de la nécessite varient selon les expressions. (Kleiber 1988 : 13)

Tableau 1 : Quelques définitions de l'anaphore[17]

[17] Nous reviendrons sur ces définitions dans la section 4 de ce chapitre.

Ce tableau se termine sur des conclusions négatives que nous partageons. Définir l'anaphore, en dépit de la banalité apparente des exemples prototypiques, semble être une tâche difficile. On ne nie pas qu'il y a autant de problèmes liés à l'anaphore non résolus, mais ça n'empêche pas d'affirmer que certains auteurs, entre autres ceux cités dans notre tableau définitoire, ont réussi à établir des bilans clairs et des approches défendables pour traiter les chaînes anaphoriques et ses portées dans le texte. Nous pensons ainsi que les analyses avancées permettent de développer la thématique anaphorique et l'empêche de stagner.

Malgré la divergence ci-dessus, nous postulons que deux niveaux constituent la représentation des différents types d'anaphores : les unités linguistiques et les rapports sémantiques entre anaphore et source, fondant certaines classifications. Nous schématisons la définition (1) (Milner 1982 : 18) de la relation anaphorique comme suit :

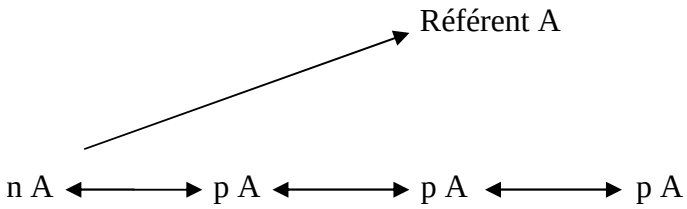


Figure 2 : Schéma de la relation anaphorique : n= nom / SN

Généralement, les définitions linguistiques de l'anaphore postulent que (1) le dispositif anaphorique repose sur un transfert de sens de l'antécédent, élément possédant sa référence propre, vers l'anaphorique, un élément qui par *l'emprunt* (*Cf.* définition (5) de Moeschler/Reboul (1994 : 523)) de la référence de cet antécédent acquière sa référence ; et (2) c'est sur une interprétation que repose la mise en relation de l'anaphorique avec son antécédent. Il est nécessaire de signaler que malgré cette opposition, la plupart des auteurs s'accordent à dire que l'anaphore[18] se définit comme la dépendance interprétative d'une unité B par rapport à une unité A.

Depuis plusieurs années, diverses théories s'intéressent au concept d'anaphore et à ses mécanismes référentiels, ainsi ce concept fait partie de la thématique générale de la référence. Dans la section suivante, nous n'aborderons pas toutes les théories mais seulement celles qui se rapportent à notre problématique.

[18] Nous nous livrerons à une analyse détaillée de notre choix dans la section 4.

2. Anaphore et référence

Il nous semble opportun de consacrer cette section à la thématique de la référence car l'étude de l'anaphore est étroitement dépendante de celle de la référence. Tout d'abord, nous donnerons quelques conceptions de la référence au sein de la réflexion linguistique au sens large. Ensuite, nous exposerons la théorie milnérienne de la référence et justifierons notre choix pour étudier, finalement, la position de l'anaphore au sein de cette théorie.

2.1. Quelques conceptions de la référence

De manière évidente, la problématique de la référence demeure, et a toujours été, centrale en philosophie en général et en philosophie du langage en particulier. Toutefois, elle est aussi abondamment traitée en linguistique, notamment en sémantique et en pragmatique. Une caractéristique très commune des langues naturelles est l'utilisation d'expressions référentielles qui, selon le contexte, la sémantique du discours et le message prévu du texte, ont une grande variété de formes. L'une des principales fonctions du langage est de communiquer sur le monde, qu'il soit réel ou imaginaire. Ainsi, l'étude de la référence est nécessaire à l'analyse des langues naturelles. Nous partons de la conception de Kleiber (1997) : selon lui, la référence est la relation par laquelle certaines unités linguistiques désignent certains objets du monde :

> Accepter que les expressions linguistiques réfèrent à quelque chose, qu'elles ont un référent, revient à accepter l'existence de ce référent. […] La référence repose crucialement sur « un axiome d'existence ». (Kleiber 1997 : 9)

Afin de mieux comprendre l'importance et les difficultés de la thématique de la référence en relation avec l'anaphore en linguistique contemporaine, nous pensons qu'il est important de présenter quelques approches de la référence. De la philosophie du langage médiévale jusqu'à la linguistique contemporaine, la référence se voit opposer à la dénotation, signification, signe, sens, existence, *etc*. Une bibliographie abondante avance des propositions variées afin d'éclaircir au mieux cette opposition. Nous remontons à l'époque médiévale, au XII^{e} siècle plus précisément, où l'on parle des Terministes (Guillaume de Sherwood, Pierre d'Espagne, Gautier Burley, Guillaume d'Ockham, Albert de Saxe, Marsile d'Inghen, Pierre d'Ailly,…) philosophes du langage qui se focalisaient sur les propriétés des termes[19]. Leurs débats s'articulaient principalement autour de la nature exacte du *significatum*, objet de la signification. Deux conceptions se sont ainsi opposées :

[19] Cf. Böhner (1952) pour une analyse détaillée.

- L'interprétation selon laquelle le langage, ou les termes, signifiaient uniquement des concepts mentaux et ne se rapportaient que par l'intermédiaire des concepts aux objets extérieurs. (C'est la conception de Boèce du Peri Hermeneias qui s'imposait jusqu'au XIII^{e} siècle).
- D'autres (comme Guillaume d'Ockham), *a contrario*, voyaient que le langage signifie les choses elles-mêmes et non plus les concepts. Nous rejoindrons l'affirmation de Biard (1997) et disons qu' :

> A la différence toutefois des terministes parisiens du XIII^{e} siècle, Guillaume d'Ockham ne définit jamais, dans la *Somme de logique*, la *signification* en tant que propriété du terme. L'idée de signification doit être reconstituée à partir de ce qui est dit du signe, d'une part, et de *signifier*, d'autre part. (Biard 1997 : 54)

Les Terministes médiévaux ont inspiré de nombreux philosophes contemporains, et surtout Frege (1892[20] et 1971), dont les travaux ont élaboré une véritable charnière dans la réflexion sur la thématique de la référence. Il traitait la dichotomie *Sinn/ Bedeutung* (traduite aussi par sens/référence ou sens/signification) en proposant que ce couple soit associé à un signe[21] :

- La référence d'une unité correspond à l'objet du monde désigné. Par exemple, dans :

> [8] Antigone rentre chez elle, à l'aube, après une escapade nocturne. Elle est surprise par sa nourrice qui lui adresse des reproches. L'héroïne doit affronter les questions de sa nounou. (Résumé Antigone)

L'héroïne désigne *Antigone*, la personne dont il est question.

- Le sens d'une telle unité est « le mode selon lequel l'objet est donné » (Frege 1971 : 105).

Nous envisageons ainsi que les expressions *L'héroïne* et *Antigone* ont toutes deux même *Bedeutung* (la personne nommée Antigone et qui est le protagoniste féminin de l'œuvre) mais présentent des *Sinn* différents, le cheminement cognitif menant à l'objet désigné n'étant pas identique. Frege illustre l'opposition sens/signification par le couple *étoile du matin* et *étoile du soir*, des expressions dont le nom, la signification est la planète *Vénus* mais de sens différents. Nous partageons l'interrogation de Porge (2003 : 36) :

[20] «Comme le dit Carnap, cet article n'a pas reçu en son temps l'attention qu'il méritait, exception faite de Russel qui le rejette en 1905. Il faut attendre Church en 1940 pour que l'importance de l'article de Frege soit pleinement reconnue. » (Porge 2003 : 34)
[21] Ou « nom propre » selon Frege.

« S'agit-il d'une identité entre les choses ou entre les signes de ces choses ? » et répondons en citant Frege :

> [L'image sur la lentille] est nécessairement partiale parce qu'elle dépend du point de vue d'observation, et pourtant elle est objective, parce qu'elle peut servir à plusieurs observateurs. (Frege 1971 : 106)

Ayant passé en revue quelques conceptions liées à la thématique de la référence, nous observons que la description de la façon dont les mots de la langue font référence au monde qui nous entoure restera une grande question de philosophie du langage. Ce monde peut être celui des idées et des concepts abstraits ou celui que nous percevons au moyen de nos sens. Notre objectif est d'apporter quelques éclairages sur la notion de *référence* au sens large. Lorsque nous parlons d'un objet concret ou d'une idée abstraite, comment définir la relation mise en jeu et comment décrire la manière dont elle relie nos mots à la réalité ? Nous rejoignons Neveu (2004 : 250-251) et disons que :

> la référence est généralement définie comme la relation qui unit une expression linguistique en emploi dans un énoncé avec *l'objet du monde* qui se trouve désigné par cette expression. On appelle *référent* cet 'objet du monde'.

Le référent d'une unité lexicale, qui peut être de nature concrète ou abstraite, représente l'entité ou la partie du monde à laquelle cette unité est associée. Par exemple, des entités différentes sont désignées par les groupes nominaux *une femme* et *un homme*. Des contraintes sur la référence de ces unités lexicales sont imposées par la langue et elle spécifie le type d'entité que chacune peut désigner afin de les distinguer. Pour être le référent d'une unité lexicale donnée, un type doit correspondre à l'ensemble des propriétés qu'une entité représente. Le type d'une unité lexicale ne réfère pas directement au monde mais reflète la réalité. En effet, une entité doit posséder un ensemble de propriétés comme *être humain* et *de sexe masculin* afin d'être le référent de l'unité lexicale *homme* par exemple.

Selon les divers travaux concernant la référence associée à une anaphore, le but final est de repérer le *bon* référent, c'est-à-dire de repérer une caractéristique précise d'un signe linguistique désignant une entité du monde, réelle qu'elle soit ou imaginaire. Pour y parvenir, toute notre explication va se fonder sur l'antécédent, bien évidemment, en donnant intérêt au contexte textuel où il se trouve. Le décryptage du sémantisme phrastique est une étape capitale pour le repérage du référent. En revanche, concernant l'anaphore pronominale, le contexte de notre corpus $\text{RESUMAN}_{\text{C}}$ n'est pas satisfaisant pour

l'expliquer, d'autres indications précises seront utilisées pour décoder la valeur référentielle[22]. Ainsi, la résolution d'une anaphore pronominale nécessite de suivre l'enchaînement référentiel proclamant un même objet ou un même individu. En outre, lorsque nous évoquons la détermination ou l'identification de la façon par laquelle les termes renvoient aux choses du monde ou réfèrent à des éléments du réel, à des *pensées* ou à des *idées*, nous évoquons inéluctablement une portée philosophique qui entre en jeu dans l'interprétation de la relation référentielle, et donc clarifiant le lien terme/réalité. Or, notre corpus ne réfère pas à un univers réel mais plutôt un fictif/virtuel, qui ne doit pas nécessairement être « spatio-temporel » (Milner 1982 : 9) et qui n'existe que par convention établie entre l'auteur et son locuteur. Pour Milner, la réponse est simple : le réel est contraint par le linguistique :

> Cela posé, il suffit de réfléchir un instant pour observer que n'importe quelle séquence nominale n'est pas associée à n'importe quel segment ; autrement dit, une langue naturelle comporte un lexique, et l'une des propriétés de ce dernier, c'est de distinguer des unités d'après le type de segment qu'elles peuvent désigner. Une unité lexicale étant choisie, certains segments sont d'emblée éliminés en tant que références possibles ; en ce sens, à chaque unité lexicale individuelle, est attaché un ensemble de conditions que doit satisfaire un segment de réalité pour pouvoir être la référence d'une séquence où interviendrait crucialement l'unité lexicale en cause. Cet ensemble de conditions décrit donc un *type* (ou si l'on veut une *classe*) de référence possible ; il est distinct des segments de réalité, mais pèse sur eux. (Milner 1982 : 10)

Dans cette optique, nous choisirons, alors, l'approche de Milner (1976, 1978, 1982 et 1989) qui constitue un excellent exemple d'une théorie de modélisation strictement linguistique, excluant le recours à des éléments situationnels des phénomènes liés à la référence en général et à l'anaphore en particulier, ce qui est conforme au cadre de notre recherche. Nous en proposerons, dans ce qui suivra, un compte rendu en nous basant particulièrement sur la première partie de l'ouvrage de Milner (1982), *Ordres et raisons de la langue*, qui rassemble et développe les travaux antérieurs de l'auteur. L'auteur, dans un premier chapitre, y évoque la thématique de la référence, puis, dans les chapitres II et III de cette même première partie, il s'attache plus particulièrement au problème de l'anaphore. En suivant ce mode de présentation, nous aborderons dans un premier temps les propositions de Milner concernant la référence, et traiterons ensuite plus spécifiquement l'anaphore au sein de la théorie milnérienne.

[22] Nous y reviendrons dans la section 4 de ce chapitre.

2.2. Théorie milnérienne de la référence

La théorie de la référence développée par Milner en 1982 permet d'expliquer comment une unité lexicale peut avoir différentes références. Plus précisément, l'auteur, dont la vision de la référence est à la fois sémantique et réaliste, s'intéresse au problème de la référence des séquences nominales dans le chapitre intitulé « Réflexions sur la référence et la coréférence ». Dans le passage suivant, Milner affirme que les séquences nominales ont pour fonction de désigner des portions du monde réel :

> On s'accorde à reconnaître que dans certaines conditions les séquences linguistiques peuvent être associées à certains segments de réalité, qu'elles sont dites désigner et qui sont leur référence. […] Une séquence nominale a […] une référence, qui est le segment de réalité qui lui est associé. (Milner 1982 : 9)

Cependant, écartant toute possibilité d'interprétation matérialiste de ses propos, l'auteur pense que, de la même manière qu'un nom concret, un nom abstrait peut être associé à un segment du réel. Il insiste en effet sur le fait que le segment de réalité désigné par une séquence nominale ne doit pas être nécessairement *spatio-temporel*. Deux concepts fondamentaux représentent la dualité de sa vision de la référence (conditions et segment du réel) :

- la *référence virtuelle* qui correspond au segment de réalité associé à une unité lexicale ;
- la *référence actuelle* qui correspond au segment de réalité qu'elle désigne en usage, c'est-à-dire son « sens lexical ».

Par analogie avec la conception frégéenne[23], Milner attribue *sinn*/sens à la *référence virtuelle* et *Bedeutung*/dénotation à la *référence actuelle* :

- La référence virtuelle représente la signification lexicale (ce que représente la définition du dictionnaire) : en d'autres termes, elle correspond à la description de l'ensemble des propriétés d'une unité lexicale. Même si toute unité lexicale possède une référence virtuelle, ce n'est qu'une fois employée, mise en contexte, qu'elle possédera une référence actuelle. S'il est isolé, le mot *table* ne fait que délimiter un type d'entité, alors qu'une fois employé, il désigne une entité dans le monde, cette dernière étant associée à des combinaisons d'unités lexicales (séquences composant des groupes nominaux ou verbaux). Par exemple, la référence virtuelle que possède le mot *table* ne

[23] En note de bas de page dans Milner (1982 : 10).

désigne aucune entité particulière mais représente un ensemble de propriétés telles que, *objet, quatre pieds, outil* etc. Maintenant, si on ajoute un déterminant, le groupe nominal *cette table* désignera une entité particulière dans le monde qui possède l'ensemble des propriétés d'une table.

La référence virtuelle spécifie les conditions auxquelles une entité doit satisfaire pour être un référent acceptable de ce terme référentiel. Ces conditions peuvent être en partie fixes, mais aussi variables. Pour le terme *chien* par exemple, les propriétés [être un animal], [être un mammifère] ou encore [aboyer] sont fixes, alors que le chien a un pelage de couleur variable parmi un ensemble fini de couleurs possibles dont est, à priori, exclu le vert par exemple. D'autres conditions peuvent être facultatives, c'est-à-dire non obligatoirement réalisées lors de toutes les acquisitions d'une référence actuelle. Ainsi le fait de posséder une queue pour un chien est une propriété facultative, car même si sa queue est coupée, on pourra donc toujours y référer avec le terme *chien*.

- La référence actuelle renvoie, contrairement à la référence virtuelle, à un objet particulier de la réalité extralinguistique. Une référence actuelle est attribuée à un terme référentiel seulement quand celui-ci apparaît dans un énoncé. Ainsi, les termes *président de la République* ont pour référence actuelle *Macron* dans l'exemple suivant :

[9] Macron, le président de la République a visité la Chine avec son épouse.

Cette référence actuelle ne correspond pas à chacune des unités de la séquence mais à leur combinaison : les références virtuelles *président* et *république* peuvent se combiner et former le groupe nominal *le président de la république* qui, grâce au déterminant, possède une référence actuelle.

> si l'on considère les emplois en eux-mêmes, ce ne sont pas aux unités lexicales comme telles que sont associés les segments de réalité, mais bien aux groupes nominaux pris dans leur ensemble. Dans ces groupes, plusieurs unités lexicales peuvent intervenir, et les références virtuelles de chacune se combinent pour contraindre une référence actuelle possible ; mais une référence actuelle donnée n'est associée qu'à la combinaison d'ensemble et non pas à chacune des unités combinées. (Milner 1982 : 10-11)

La référence des verbes peut être décrite de la même manière que celle des noms. Selon Milner, les verbes n'acquièrent de référence actuelle que lorsqu'ils ont un temps et qu'ils sont accompagnés d'arguments (au moins un sujet *syntaxique*). Les verbes ne sont pourvus que d'une référence virtuelle quand ils sont isolés. Comme pour les noms, la référence actuelle se rapporte à la combinaison d'unités qui forment un groupe verbal. Par exemple, *marcher* est un verbe qui désigne un mouvement particulier des pieds pour se déplacer d'un endroit à un autre, un type d'action. Mais *Paul marche*, est une action particulière avec une référence au monde réel. En effet, elle désigne plutôt une instance particulière de ce type d'action. Celle-ci est située temporellement dans le présent avec un actant qui s'appelle *Paul*. Les groupes nominaux constituant les actants du verbe et le verbe lui-même (éventuellement doté de propriétés de temps, d'aspect ou de mode) ont chacun leur propre référence virtuelle dans le groupe verbal. Les verbes imposent, en plus de ces propriétés, des contraintes sur le type de leurs arguments. Par exemple, dans son interprétation habituelle, le verbe *accoucher* se réfère à un sujet humain de sexe féminin. Ainsi, la référence virtuelle d'un verbe impose des contraintes sur les références virtuelles de ses arguments.

2.3. L'anaphore dans le cadre de la théorie milnérienne

Milner définit l'anaphore ainsi :

> il y a relation d'anaphore entre deux unités A et B quand l'interprétation de B dépend crucialement de l'existence de A, au point qu'on peut dire que l'unité B n'est interprétable que dans la mesure où elle reprend entièrement ou partiellement A. (Milner 1982 : 18)

Cette définition regroupe l'anaphore pronominale et l'anaphore nominale. Nous nous intéressons à l'anaphore pronominale puisque c'est le cas étudié dans cette thèse. Milner présente les pronoms comme dépourvus de référence virtuelle. Ils ne désignent pas de façon indépendante des entités du monde et ne sont pas capables de déterminer leur propre référent. En effet, il considère qu'ils sont des unités *référentiellement non-autonomes* :

> Bien évidemment, le cas des pronoms de troisième personne entre dans cette catégorie : la référence virtuelle d'un tel pronom ne peut être définie en elle-même hors emploi, aucune condition n'étant requise d'un segment de réalité pour être désigné par *il* (*elle*, *ils*, etc.), sinon des conditions tenant à l'énoncé singulier où *il* est employé. (Milner 1982 : 19)

Les pronoms anaphoriques sont privés d'autonomie référentielle car ils n'ont pas de référence virtuelle selon Milner. En d'autres termes, lorsqu'ils apparaissent dans un

énoncé, ils ne peuvent pas déterminer par eux-mêmes leur référence actuelle. L'auteur définit la saturation sémantique comme la capacité à posséder une référence virtuelle claire. Selon ce dernier, les pronoms anaphoriques sans référence virtuelle sont totalement insaturés sémantiquement. En effet, plus la référence virtuelle d'un terme est précise, plus sa saturation sémantique est élevée. C'est par un processus de saturation sémantique qu'ils acquierraient une référence actuelle. Dans une relation anaphorique, les pronoms ne peuvent pas être l'antécédent et ne peuvent pas fournir une référence virtuelle. Ils entretiennent néanmoins une relation de coréférence quand plusieurs pronoms anaphoriques ont le même antécédent, c'est-à-dire dans les chaînes de référence. Rappelons que la relation d'anaphore et celle de coréférence[24] sont bien distinctes. Quand il s'agit d'une anaphore pronominale, il y a toujours une coréférence virtuelle puisque les deux relations peuvent coexister. Dans une relation anaphorique, nous pouvons interpréter le pronom grâce à la coréférence virtuelle.

Dans ce processus, le pronom adopte la référence virtuelle de son antécédent. Ainsi, c'est sur l'établissement d'une coréférence virtuelle entre l'antécédent et l'anaphorique, qui entraîne l'identité des références actuelles des deux termes de la relation, que repose la relation anaphorique pronominale. Ainsi, le pronom anaphorique et son antécédent doivent avoir le même sens en emploi, en d'autres termes la même référence actuelle. Dans l'exemple :

> [10] A Saumur, Félix Grandet (le père Grandet) s'est constitué, grâce à de nombreuses spéculations foncières, une fortune qui n'a d'égal que son avarice. Il règne en tyran sur son entourage. (Résumé Eugénie Grandet)

Les deux références actuelles sont identiques et le nom donne une référence virtuelle au pronom, qui en était dépourvu. En effet, c'est avec le nom propre *Félix Grandet*, qui a une référence actuelle, que le pronom *il* est en relation de coréférence actuelle.

Les pronoms qui, selon Milner, ne possèdent pas de référence virtuelle portent tout de même, selon Kleiber (1994), un contenu sémantique propre. Ce contenu est composé de deux parties : une partie procédurale (ou instructionnelle) qui indiquerait à

> l'allocutaire comment procéder pour trouver la bonne interprétation, et en l'occurrence pour les expressions référentielles, comment accéder au référent (Kleiber 1994 : 14)

[24] Nous y reviendrons dans la section suivante.

et une partie descriptive (ou représentationnelle) qui correspondrait à la signification lexicale ou à ce que Milner (1982) définit comme référence virtuelle. Ces pronoms portent un sens descriptif même s'il est « ténu » (Kleiber 1994 : 14), bien que le sens procédural soit la partie la plus développée de la signification des pronoms anaphoriques. La partie descriptive procurerait essentiellement des informations sur le nombre et le genre des référents compatibles. Soient :

[11] Paul a acheté une Toyota, parce qu'elles sont économiques. (Résumé d'après Kleiber 1992 : 50)
[12] J'ai voulu chercher Pierre. Tu sais, ils n'habitent plus à X. (Résumé d'après Kleiber 1994 :71)

Le pronom *elles* dans l'exemple (11) ne réfère pas à la Toyota achetée mais à l'ensemble des Toyota, et le pronom *ils* dans l'exemple (12) réfère à la famille de *Pierre* et non à lui-même. Il n'y a pas d'accord en nombre entre le pronom et son antécédent (dans ce cas le terme de *déclencheur d'antécédent* semble plus approprié que le terme d'antécédent utilisé par Kleiber (cf. Apothéloz 1995)) ce qui permet de déterminer qu'il n'y a pas de coréférence entre les deux termes. Ceci montre que c'est bien le nombre véhiculé par la signification du pronom qui permet l'apparition du référent de ce pronom dans le discours et non pas l'antécédent qui confère la marque du nombre au pronom (Kleiber 1994 : 71).

Le sens procédural des pronoms anaphoriques représente une autre partie de leur signification. Par exemple, l'indication présente dans le sens du pronom *il* est, selon Kleiber (1994 : 82-83), qu'il faut rechercher son référent dans une proposition ou une situation qui est manifeste ou saillante[25] par le contexte antérieur ou par une perception directe de la situation, dans laquelle le référent joue un rôle d'argument. L'hypothèse de Milner se distingue de celle de Kleiber sur deux points majeurs. En premier lieu, la relation anaphorique est essentiellement pragmatique selon Kleiber, alors que selon Milner, elle est exclusivement sémantique. En deuxième lieu, Kleiber, qui parle de sens descriptif ou représentationnel, soutient que les pronoms sont pourvus de référence virtuelle propre, même si ce sens descriptif est moins riche que celui des expressions nominales, alors que Milner soutient qu'ils n'en possèdent pas.

L'analyse de l'anaphore de Milner semble centrée sur les cas d'anaphores coréférentielles. Dans ces dernières, la même référence actuelle est partagée par les deux termes de la relation anaphorique. L'analyse de Kleiber, par ailleurs, porte également sur les cas d'anaphores non coréférentielles. Lorsqu'il n'y a pas de relation de coréférence, la

[25] Nous analyserons la saillance, avec plus de précision, dans le deuxième chapitre de la deuxième partie.

partie du discours qui permet la présence d'un anaphorique est le déclencheur d'antécédent. Dans l'exemple :

[13] J'ai été au **concert** hier. **Ils** jouaient la neuvième symphonie. (Résumé Kleiber 1994 : 65).

Le déclencheur d'antécédent de *ils* est le mot *concert*. En effet, la forme pronominale renvoie à des *musiciens* dans ce discours. Le déclencheur d'antécédent et l'anaphorique ne partagent ni la même référence actuelle ni la même référence virtuelle. Ainsi, aucune forme de coréférence ne s'établit entre les deux termes de la relation anaphorique. On ne pourrait pas attribuer un référent à ces pronoms en se basant uniquement sur une relation anaphorique sémantique (proposition de Milner). Ainsi, en introduisant dans son analyse une composante pragmatique dans la relation anaphorique, la proposition de Kleiber nous semble plus adaptée pour décrire cette relation.

Avec l'introduction des concepts de référence virtuelle et actuelle et dans le cadre d'une approche logico-sémantique de la référence, l'apport de Milner (1982) est exploité dans l'étude du problème de l'anaphore que nous développerons par la suite. Nous procèderons à l'examen de la coréférence et de l'anaphore, deux relations qui dépendent de la définition de la notion de référence.

3. L'anaphore pronominale : une anaphore coréférentielle

La coréférence est une relation linguistique s'établissant entre deux unités lexicales qui ont la même référence. Selon Pepin (2009 : 335) :

> La coréférence est un procédé de reprise de l'information qui contribue à la cohérence du texte en indiquant au lecteur que l'on continue à parler de la même chose, d'une phrase à une autre. La coréférence met en relation deux termes : un substitut et son antécédent. Comme ces deux termes réfèrent à la même réalité, on dit qu'ils coréfèrent, d'où le nom du procédé.

Deux types de coréférence sont distingués par Milner (1976 et 1982) :

- La *coréférence actuelle* : c'est une identité référentielle entre deux termes référant au même objet du réel et désignant une même entité dans le monde. C'est une relation symétrique entre deux unités qui n'implique pas forcément l'identité des unités elles-mêmes, mais l'identité matérielle absolue des référents (Milner 1976 : 65). Dans l'exemple :

[14] Antigone$_i$ rentre chez elle, à l'aube, après une escapade nocturne. Elle est surprise par sa nourrice qui lui adresse des reproches. L'héroïne$_i$ doit affronter les questions de sa nounou. (Résumé Antigone)

c'est entre les deux groupes nominaux *Antigone* et *l'héroïne* qu'il y a une coréférence actuelle. En effet, même si les unités lexicales sont différentes, ces deux groupes réfèrent à la même entité.

- La *coréférence virtuelle* : Deux unités lexicales distinctes sont en relation de coréférence virtuelle quand elles ont les mêmes propriétés. « Elle équivaut à la synonymie lexicale absolue » (Milner 1976 : 65).

Il est possible de schématiser la relation de coréférence ainsi:

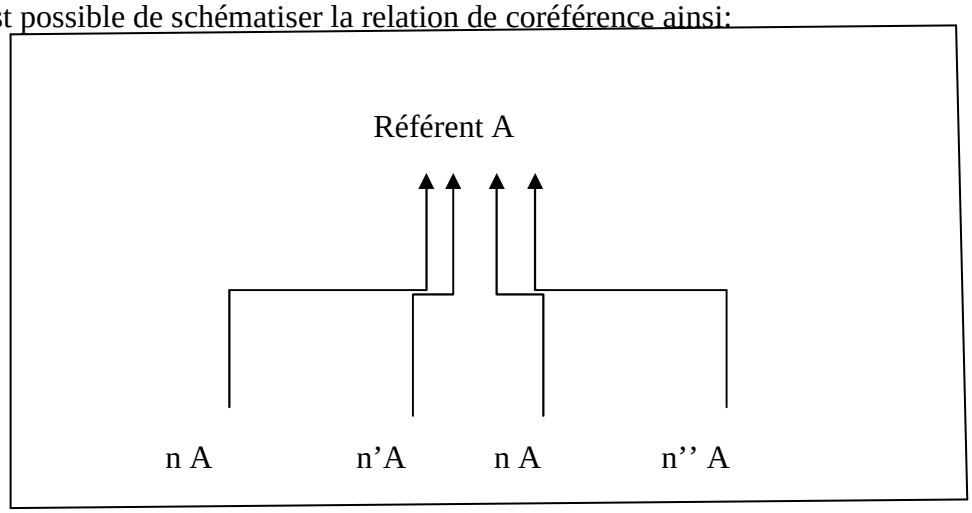


Figure 3 : Schéma de la relation de coréférence : n = nom/SN

L'expression anaphorique établit un lien direct avec le référent, nous qualifierons ainsi ce type de coréférence de *simple*. Nous confondons, de manière très fréquente, la coréférence avec l'anaphore alors que la première implique une relation symétrique et la deuxième une relation asymétrique. Si deux expressions désignent la même entité alors elles *coréfèrent*. Une *chaîne de coréférence,* qui est un élément essentiel de la *cohésion* du texte et de sa *cohérence*[26] est une suite d'expressions qui coréfèrent dans un texte. Résoudre des anaphores pronominales équivaut à rattacher un nouveau segment pronominal à une chaîne de coréférence. Un même objet du monde réel ou imaginaire est désigné par les entités textuelles dans une relation de coréférence. Dans l'exemple :

[15] Meursault, le narrateur, est un jeune et modeste employé de bureau habitant Alger. Le récit commence le jour de la mort de sa mère. Au petit matin, il reçoit un télégramme de l'asile de vieillards de Marengo. (Résumé L'étranger)

[26] Nous détaillerons ces deux notions dans le premier chapitre de la deuxième partie de ce travail.

Meursault et *il* réfèrent tous deux au narrateur. Cet exemple, où la relation de coréférence est aussi anaphorique, montre que l'anaphore et la coréférence sont liées. Cependant, certaines relations peuvent être coréférentes sans être anaphoriques et *vice et versa*. En effet, elles ne sont pas forcément identiques : considérer deux textes décrivant une même personnalité politique permet de trouver un exemple de relation de coréférence qui ne soit pas anaphorique. En effet, puisqu'il s'agit de deux documents différents, les entités textuelles ne sont pas anaphoriques et réfèrent bien au même objet du monde. Dans le but de distinguer les notions d'anaphores et de coréférence qui sont souvent confondues, nous allons discuter le statut de la coréférence. Selon Kleiber (1992 : 8) on pourrait croire qu'il y a une relation de coréférence entre l'antécédent et l'expression anaphorique. Cependant, ces deux notions représentent deux phénomènes linguistiques non identiques. Reprenons l'exemple (10) de Kleiber, cité plus haut :

[16] Paul a acheté une Toyota, car ces voitures sont économiques. (Résumé Kleiber 1992 : 50)

Une *relation anaphorique non coréférentielle* est établie entre une expression anaphorique qui renvoie à un référent générique et un antécédent qui renvoie à un référent spécifique comme dans l'exemple suivant :

[17] **Macron** est parti en voyage en Chine. **Le président de la République Française** a emmené avec lui son épouse.

Nous avons une *relation coréférentielle non anaphorique*. Il s'agit uniquement d'une identité référentielle (une relation de coréférence) entre les deux éléments de la relation et non d'une relation anaphorique *Macron/Président de la République Française*. La saturation référentielle du deuxième élément de la relation *président de la République Française* explique cette relation. Ce substantif séparé de son cotexte gauche, est facilement interprété par l'interlocuteur sans même être complété par un élément du cotexte antérieur. Il n'est pas considéré comme expression anaphorique du fait de son autonomie référentielle. Nous considérons la thématique de l'anaphore comme intégrée à la thématique générale de la référence : en effet, elle serait représentée par un sous-ensemble de la référence limité à un ensemble d'expressions linguistiques d'une langue. Dans ce même cadre de questionnement, Apothéloz (1995 : 307-311) propose trois conceptions de l'anaphore : la conception *substitutive*, la conception *antécédentiste* et la conception *mentaliste*. Nous nous y intéressons à présent dans la section suivante.

4. Approches de l'anaphore

A partir de notre étude de la bibliographie très riche concernant la thématique de l'anaphore, dans ses différentes approches et cadres théoriques, nous avons toutefois décidé de présenter, dans cette section, trois[27] grandes approches dans la mesure où elles nous permettront de mieux situer notre étude. Tout d'abord, l'approche syntaxico-formelle-substitutive[28], où l'anaphore constitue la substitution d'une expression linguistique par une autre, ensuite, l'approche fonctionnelle-textuelle-antécédentiste où l'interprétation de l'antécédent est possible grâce à sa dépendance à l'anaphore et enfin, l'approche cognitive-mémorielle-mentaliste où l'anaphore se présente comme une représentation mentale et où le segment textuel n'est plus pris en compte. Dans ce qui suit, nous allons d'abord présenter les idées maîtresses de chaque approche, et mettre ainsi en avant la complexité de l'étude de l'anaphore. Nous allons souligner les problèmes et limites de chaque approche étudiée pour finalement émettre notre positionnement et notre choix méthodologique.

4.1. Approche substitutive

4.1.1. Présentation

Cette conception de l'anaphore pronominale se fonde sur l'identité référentielle entre l'antécédent et l'anaphorique. Un pronom acquiert toutes les propriétés référentielles du nom quand il est lié à son antécédent par une relation d'anaphore. La substitution d'un syntagme nominal par un pronom peut être justifiée par des contraintes stylistiques comme éviter la répétition (Achard-Bayle 2001). Nous retrouvons cette justification chez Vargas (1992 : 172) qui pense que le pronom « permet d'éviter la disgrâce d'une répétition » et qui ajoute que c'est « un outil économique ».

Selon De Week (1991), le phénomène anaphorique est beaucoup plus restreint que la définition première de la substitution. Nous retrouvons dans cette définition, en plus des déictiques, les signes conventionnels *oui* et *non*, des auxiliaires modaux, ou non, des modificateurs (*peut-être, certainement*...) et les interrogatifs, selon le sens plus au moins strict que l'on y attache. Une importante confusion est entrainée par la grande diversité des

[27] Nous en avons déterminé cinq (*approche pragmatique et approche (pragma-)sémantique* en sus.) Pour des raisons opératoires, nous analysons seulement les trois premières approches. Ces dernières nous semblent préférables puisque notre étude s'appuie exclusivement sur la langue écrite dans un contexte bien défini. Autrement dit, le contexte d'énonciation des expressions référentielles ne peut être analysé qu'endophoriquement. Reboul (1989, 1990) est une des plus ardents défendeurs de *l'approche pragmatique* ; Kleiber (1994) est sans doute l'avocat principal de l'approche *(pragma-)sémantique*.

[28] Nous avons cumulé, pour chaque approche, les termes utilisés par les linguistes pour la désigner. Nous avons séparé ces termes par un trait. Cette variété terminologique confirme encore une fois la diversité anaphorique.

acceptions attribuées à la notion de substitution, alors même que les définitions de cette notion s'accordent sur la substitution d'unités sources par des unités lexicales appartenant à d'autres catégories grammaticales. Dans les ouvrages de grammaire, il semble toutefois que c'est au fonctionnement des dispositifs anaphoriques pronominaux que la substitution se limite. Ces dispositifs sont qualifiés de *pronoms représentants* par Grevisse/Goosse (1995 : 205). Selon eux, avec la notion de substitution, il est possible de remplacer, tout en maintenant le sens de départ, l'antécédent (une partie du discours) par un pronom anaphorique (lexème de catégorie grammaticale spécialisée). C'est sur le postulat que les pronoms anaphoriques doivent acquérir une référence selon la situation discursive, car ils ne possèdent pas leur propre référence, que l'hypothèse de l'identité référentielle entre l'antécédent et le pronom repose. Ainsi, selon cette hypothèse, l'anaphore est une relation essentiellement sémantique (Kleiber 1994).

La conception substitutive de l'anaphore est en désaccord avec les données linguistiques et c'est sur la confusion entre l'anaphore et la coréférence, que nous avons présentée dans la section précédente, qu'elle repose. Toutefois, l'une n'implique pas forcément l'autre même si ces deux relations sont souvent associées.

4.1.2. Limites de l'approche substitutive

Etablir une relation syntaxique entre l'anaphore et son antécédent pour rendre compte de la relation anaphorique n'est pas suffisant puisque l'aspect sémantique et pragmatique est tout aussi important. Pour prendre en compte cet aspect, il a été proposé deux alternatives à la conception de la substitution conçues pour la résolution d'anaphores non coréférentielles dans lesquelles c'est le contenu instructionnel de l'anaphore qui conditionne l'interprétation (Kleiber, 1994). Ces deux approches ont été développées successivement par Brown/Yule (1983 ; Yule 1982) : la conception *cumulative* et celle du *vague référentiel*.

Selon Brown/Yule (1983), dans la conception cumulative, le pronom anaphorique ne se substitue pas à l'antécédent identifiable dans le discours mais à une entité enrichie par tous les prédicats qui lui ont été appliqués jusqu'au point d'apparition du pronom. C'est sous la forme d'adjoints venant déterminer le syntagme nominal antécédent initial que les prédicats se trouvant entre le pronom et l'occurrence de l'antécédent sont récupérés et cumulés. Dans le célèbre exemple de Brown/Yule :

[18] Tuez **un poulet actif et bien gras.** Préparez-le[1] pour le four. Coupez-le[2] en quatre et rôtissez-le[3] avec du thym pendant une heure. (Résumé Brown/Yule 1983 : 202)

le premier référent de *le*[1] correspond au groupe *[un poulet actif et bien gras, mort]*, alors que celui de *le*[2] est le groupe *[un poulet actif et bien gras, mort, et préparé pour le four]* et que celui de *le*[3] coïncide avec le groupe *[un poulet actif et bien gras, mort, préparé pour le four et coupé en quatre*]. Ainsi, les pronoms *le* réfèrent à des entités qui diffèrent par les prédicats appliqués à l'entité de départ. Dans un contexte évolutif, la solution proposée par Brown/Yule permet d'expliquer l'interprétation des pronoms anaphoriques et de mieux comprendre le nombre de retours en arrière réduit au cours de la lecture de texte avec des référents évolutifs. C'est à partir de son occurrence précédente, dont la référence contient celle qui est initiale et enrichie par des prédicats appliqués au cours du discours, que l'on construit le référent du pronom. Il existe toutefois quelques arguments qui remettent en cause la solution cumulative[29]. Le plus pertinent est que cette solution n'est pas compatible avec un système cognitif dont les ressources sont limitées, comme le reconnaissent Brown/Yule (1983). En effet, il faut, dans cette solution, maintenir en mémoire de travail les prédicats qui apparaissent entre deux occurrences référentielles. Ceci est aussi valable pour l'augmentation de la distance séparant deux occurrences référentielles. L'élaboration de modèles linguistiques innovants traitant du fonctionnement des pronoms anaphoriques a été engendrée grâce au fait que le système cognitif de traitement du langage est incompatible avec l'analyse linguistique du problème des référents évolutifs (Brown/Yule 1983), même si cette analyse permettait d'éliminer les objections à la conception substitutive.

Le *vague coréférentiel* (décrit par Yule 1982) est la seconde alternative à la conception substitutive. Les pronoms anaphoriques seraient, selon l'auteur, des expressions qui servent de support aux prédicats en indiquant que ces prédicats s'appliquent aux entités saillantes du discours et ne seraient pas toujours des expressions référentielles. En effet, le caractère thématique donné de l'information référentielle à transmettre inciterait les interlocuteurs à se concentrer sur la partie prédicative et à porter un minimum d'attention à cette partie du message. Cette faible attention accordée au contenu référentiel des anaphoriques pourrait permettre d'expliquer pourquoi, alors qu'il n'y a pas de coréférence stricte ente les différents pronoms *le* et leurs antécédents, des exemples comme l'exemple 18 peuvent être produits et acceptés :

> [19] Tuez **un poulet actif et bien gras.** Préparez-le[1] pour le four. Coupez-le[2] en quatre et rotissez-le[3] avec du thym pendant une heure. (Résumé Brown/Yule 1983 : 202)

[29] cf. Charolles/Schnedecker (1993).

Ces pronoms indiquent simplement pour l'auteur que le locuteur, sans qu'il existe nécessairement une relation de coréférence stricte ente le référent du pronom et celui de son antécédent, entend maintenir en position focale un référent fortement accessible et déjà introduit dans le discours. Yule (1982) accepte qu'il se met en place en général entre le pronom et son antécédent une relation de coréférence, même si, selon lui, aucun contenu référentiel n'est véhiculé par les pronoms anaphoriques dans quelques cas extrêmes. L'auteur parle de *coréférence vague* quand la relation de coréférence repose sur une relation d'identité sémantique entre le référent de l'antécédent et celui du pronom qui est approximative. Selon Charolles/Schnedecker (1993 : 10), c'est là le point faible de la solution proposée par Yule qui se sert de l'exemple ci-dessous pour justifier la notion de coréférence vague et préciser cette affirmation :

[20] Oh, tout ce qu'ils font à Edimbourg, ils le font vite mais très lentement. (Résumé Yule 1982 : 319)

Dans cet exemple, Charolles/Schnedecker soulignent qu'on peut affirmer que les pronoms *ils* renvoient bien à la même entité même s'ils semblent référer à une entité floue (les Ecossais, les habitants d'Edimbourg, ...). Une autre remarque est soulevée par cet exemple : c'est par leur rôle d'indicateur du maintien de la saillance référentielle que Yule justifie l'utilisation des pronoms. Or, avant l'apparition du premier pronom *ils* dans cet exemple, aucun référent n'a été explicitement introduit. Puisqu'à partir des prédicats contenus dans la proposition et de son contenu sémantique propre, on peut associer à ce pronom un référent, même assez imprécis[30], il a forcément un rôle référentiel. Alors, on ne peut pas parler de coréférence vague. Quand on prend en compte les réalités linguistiques, la solution proposée par Yule du vague coréférentiel n'y serait pas adaptée comme la conception cumulative.

Trois déductions peuvent être tirées des solutions proposées par Brown et Yule, même, si aucune n'explique le fonctionnement du système anaphorique. Tout d'abord, pour expliquer la transmission de la référence d'un antécédent vers un pronom anaphorique, la solution substitutive est inadaptée. L'usage, dans des contextes d'anaphores associatives, de pronoms anaphoriques ou de référents évolutifs montre que c'est sur un effet de coïncidence que la définition substitutive de l'anaphore pronominale repose ;

[30] Le rôle d'indicateur du maintien de la référence thématique semble toutefois particulièrement adapté au fonctionnement de l'ellipse, même si, pour les pronoms anaphoriques, il n'est pas toujours pertinent. Cette conjecture attire notre attention sur l'importance de considérer, dans les théories modélisant leur fonctionnement, le sens procédural des pronoms anaphoriques.

même si la substitution paraît adaptée aux cas les plus simples d'anaphores pronominales. En opposition à la conception substitutive, si une identité de référence actuelle entre le pronom et son antécédent est souvent observée, on ne peut généraliser cette propriété à l'ensemble des occurrences pronominales de l'anaphore. Ensuite, c'est l'ensemble du discours qui contribue à modeler l'antécédent d'une anaphore pronominale. Ainsi, on ne peut résumer l'antécédent d'un pronom ni à sa référence actuelle au moment de son apparition dans le discours, ni à l'occurrence lexicale présente à la surface du discours. Dans la détermination de l'antécédent, c'est la totalité des informations discursives, celles non explicites mais inférables comprises, qui semblent intervenir ; même si Brown et Yule (1983) soulignent l'importance de l'influence de la partie prédicative du discours.

Une nouvelle conception du fonctionnement des anaphores pronominales a été proposée Brown/Yule (1983) devant l'incompatibilité de la solution du vague coréférentiel et celle de la cumulation. Les auteurs pensent que, vu le coût cognitif de la solution cumulative, l'antécédent des pronoms anaphoriques est une entité présente non pas dans la forme présente à la surface du texte mais dans la représentation mentale[31]. Toutefois, les idées majeures de ces deux solutions ne sont pas délaissées par les auteurs. Ils retiennent de la solution cumulative la nécessité de prendre en considération les prédicats appliqués à l'antécédent.

4.2. Approche textuelle

4.2.1. Présentation

Cette approche, considérée comme traditionnelle, s'intéresse à la distinction entre les différentes localisations du référent et est défendue par les auteurs qui prônent l'approche textuelle. Halliday/Hasan (1976) la qualifie de fonctionnelle[32], Apothéloz (1995) se veut antécédentiste[33], quant à Kleiber (1994), il la désigne par textuelle. Pour définir les phénomènes anaphoriques, cette approche s'appuie sur la dichotomie anaphore/deixis qui entraîne une opposition entre texte et situation. Notons toutefois un point commun entre ces deux phénomènes : ils possèdent tous deux une faible saturation sémantique. En effet,

[31] L'approche mentaliste sera abordée ultérieurement avec plus de détails.
[32] D'après Halliday/ Hasan (1976 : 52), la référence n'est pas uniquement étudiée pour elle-même, mais avec l'ellipse, la substitution, la conjonction et la cohésion lexicale, elle est considérée comme étant responsable de la cohésion et de la cohérence d'un texte.
[33] Apothéloz (1995) privilégie la conception antécédentiste et met l'accent sur la dépendance de l'expression anaphorique vis-à-vis de l'expression qui permet son interprétation (nommé « antécédent », « source », *etc*.). L'anaphore est ainsi réduite à un lien interprétatif unissant « deux segments textuels univoquement délimitables » (Apothéloz 1995 : 310).

dans l'approche textuelle, afin de distinguer l'anaphore, la reconnaissance d'une expression est nécessaire et de ce fait, il faut faire intervenir un antécédent ; en d'autre termes un autre élément du contexte linguistique. La distinction des deux termes se retrouve dans la source de leurs références. En ce sens, l'expression anaphorique trouve son exhaustivité dans l'antécédent, référentiellement autonome, alors que c'est dans l'environnement physique que l'expression déictique trouve sa référence.

Selon les auteurs de cette approche, un objet peut se localiser linguistiquement en discours ou en dehors du discours, il peut être alors défini par ces deux lieux d'existence : Halliday/Hasan (1976) ont postulé la différence entre l'endophore ou « référence textuelle » *(textual reference)*, lorsque le référent se retrouve dans le discours, et l'exophore ou « référence situationnelle », lorsque le référent se retrouve dans la situation extralinguistique d'énonciation (l'univers non discursif). La théorie de Maillard (1974) met aussi en relief cette opposition endophore/exophore. Néanmoins, l'auteur ne parle pas d'*endophore* mais de *diaphore*, notion qu'il caractérise par non « vectoriellement orientée ». Selon lui un référent est:

> le segment qui doit être mis en rapport avec une autre partie de la chaîne et référé ce qui est impliqué par le référent. Le lien référentiel est le rapport sémantique d'identification qui s'établit entre les deux. (Maillard 1974 : 56)

> un fragment énonciatif quelconque est soit « aphorique » soit « anaphorique » et / ou « cataphorique », relativement au contexte. Il est « aphorique » s'il est parfaitement clos sur lui-même et n'implique pas le texte. Il est « anaphorique » s'il suppose l'énoncé antécédent et « cataphorique » s'il se rapporte à l'énoncé subséquent. (Maillard 1974 : 57)

Relevant toutes les deux de la diaphore, l'anaphore et la cataphore peuvent soit référer à un simple segment et sont alors segmentales, soit référer à un énoncé qui peut être plus ou moins long et sont alors résomptives. Nous voyons pourquoi Apothéloz (1995) qualifie cette approche textuelle d'antécédentiste. En effet, elle favorise la dépendance de l'expression anaphorique envers l'antécédent qui représente l'expression permettant sa résolution. Ainsi, l'anaphore peut se définir comme la relation de dépendance de deux unités : la seconde renvoyant à la première qui est *l'anaphorique* (ou *antécédent*).

De Weck (1991 : 35) propose trois conditions dont dépend l'anaphore : « la dépendance cotextuelle, la présence d'une source identifiable dans la portion du texte qui précède et la reprise de la source ». Concernant la *dépendance cotextuelle*, notons que le texte représente le contexte interprétatif de l'unité linguistique : en effet, un certain degré d'incomplétude accompagne la définition de l'anaphore (Corblin, 1985 et 1987). Pour une

efficacité de la dépendance au co-texte et pour que l'anaphore soit référentiellement saturée, il est parfois nécessaire d'envisager des unités linguistiques, notamment les pronoms, comme des formes vides nécessitant un rattachement de l'anaphore au co-texte. Ainsi, l'anaphore doit être mise en rapport avec une unité que l'on peut identifier de manière claire dans le texte, *l'antécédent*. De ce fait, l'antécédent indique la direction de la mise en rapport des unités. L'anaphore est ainsi définie, dans l'approche antécédentiste, comme

> une connexion anaphorique [qui] met en relation un terme supérieur, l'antécédent, qui communique une valeur à un terme inférieur, qui n'est autre que l'élément anaphorique. (Wiederspiel 1989 : 99)

Selon Milner (1982 : 18, 32, 33, et 34), trois propriétés caractérisent cette relation structurelle : elle est définie comme asymétrique, non transitive et non réflexive. Tout d'abord, il existe une relation asymétrique entre l'anaphorisant et le terme anaphorisé. L'anaphore pronomiale est distinguée de la nominale, elle est définie par Milner (1982 : 32) comme « ouvertement hétérogène, du point de vue catégoriel ». Dans un cas d'anaphore pronominale, l'anaphorisant (un pronom de la troisième personne), qui est dépourvu de référence virtuelle, reçoit cette dernière grâce à l'anaphorisé (un nom défini ou indéfini) qui doit en avoir obligatoirement une. Concernant l'anaphore nominale, Milner parle d'« homogénéité catégorielle ». En effet, dans un tel cas, l'anaphorisant et l'anaphorisé sont des noms, de plus à la différence du premier cas, « la différence de l'indéfini au défini est cruciale : le pivot de la relation est en effet que le référent de l'anaphorisant soit tenu pour identifier un seul fait de la relation qu'il entretient avec le référent de l'anaphorisé » (Milner 1982 : 33). Ensuite, la non transitivité constitue la deuxième propriété de la relation anaphorique. Cette propriété se manifeste dans l'anaphore pronominale. En effet, « un pronom anaphorique ne peut fonctionner comme premier terme d'une relation d'anaphore. » (Milner 1982 : 33), ainsi un pronom par définition ne peut pas être anaphorisé, il est nécessairement anaphorisant. Enfin, vis-à-vis de sa source, l'anaphore n'est pas autonome, ainsi une connexion anaphorique est non réflexive. De ce fait, selon Milner, il est impossible qu' « un terme soit anaphorisant ou anaphorisé de lui-même » (Milner 1982 : 34).

Ce que nous retenons de cette approche est que la compréhension de certains éléments discursifs, qui peuvent avoir des dimensions variables, n'est possible qu'avec la prise en compte de leur relation, appelée relation anaphorique, avec d'autres éléments

antérieurs. Ainsi, cette relation entre l'anaphorisé et l'anaphorisant dont il dépend est structurelle. Elle est par définition asymétrique, intransitive, et non-réflexive. Quand on s'intéresse à la localisation, cette relation présente deux types de références : une textuelle, endophorique, où le référent se situe dans le discours et l'autre, situationnelle, exophorique, où le référent se situe dans l'espace extra-linguistique, en dehors du discours. Il est intéressant de noter que l'approche textuelle présente quelques limites exposées dans ce qui suit.

4.2.2. Limites de l'approche textuelle

Comme précédemment évoqué, la dualité deixis / anaphore est à l'origine de la l'approche textuelle. Les deux termes, déictique et anaphorique, ont en commun une faible saturation sémantique. En revanche, la source de leur référence est distincte, le premier la retrouve dans la situation d'énonciation et le deuxième dans l'antécédent, terme référentiellement autonome. La conception textuelle de l'anaphore implique le caractère endophorique de l'expression anaphorique où le référent est présent dans le discours. Néanmoins, parfois, une expression peut ne pas opérer sur un segment textuel mais sur une représentation. De plus, le contexte linguistique joue un rôle important dans cette conception puisque c'est à partir de ce contexte que les anaphores sont interprétées (Milner 1982 : 20). Ajoutons à cela qu'il existe une relation de coréférence entre l'anaphore et son antécédent dans la conception textuelle. Citons tout d'abord la notion de contexte linguistique. Selon Kleiber (1994b : 8) :

> la situation est aujourd'hui beaucoup moins sereine. On ne tient plus pour totalement satisfaisante, quelque vertu qu'elle pût avoir, la conception localisante classique en matière d'anaphores et de pronoms. Les travaux de ces quinze dernières années (…) ont en effet clairement montré que la réponse qui consiste à dire que le référent d'une expression anaphorique se trouve dans le contexte linguistique s'avérait de toute façon trop « courte » puisque rien n'était dit sur la façon de retrouver le « bon » antécédent et que, beaucoup plus grave, le « morceau » de solution proposé n'était même pas adéquat.

Dans l'exemple :

[21] Attention ! ne t'approche pas. Il est dangereux. (Résumé Kleiber 1992a)

le pronom *il* est employé sans l'intervention d'un antécédent et selon la conception textuelle, il y a ici une utilisation déictique du pronom. De ce fait, Kleiber remet en cause ici la pertinence de la conception textuelle de l'anaphore. La deuxième limite que l'on peut relever est la dualité deixis / anaphore qui représente un critère opératoire pour cette

conception. Néanmoins, notons que l'on peut employer de manière anaphorique ou déictique une même expression (SN démonstratif ou défini, un pronom de troisième personne). Dans l'exemple :

[22] M. Lantin ayant rencontré **cette jeune fille**, dans une soirée. (Résumé Charolles 1991 : 206)

L'auteur montre que le SN démonstratif *cette fille* utilisé ici ne peut pas être interprété ni déictiquement ni anaphoriquement et de la sorte, il pose des difficultés d'interprétation. Ces dernières sont dues dans le premier cas, au fait de ne pas montrer ostensivement la personne en question (même dans le cadre du contexte imaginaire où elle est insérée) et dans le second cas à l'apparition du démonstratif en première position. Selon Charolles, ni l'interprétation exophorique, ni l'interprétation endophorique ne sont admissibles.

La troisième limite de la conception textuelle est la remise en question de l'homogénéité de la catégorie de l'anaphore. En effet, cette conception oblige le recours à un antécédent. Dans certaines situations la localisation de la source est difficile à déterminer, ce qui rend difficile l'identification de l'antécédent dans le cas de la cataphore. Afin de pallier ce problème, Corblin (1985 : 178) empreinte le terme de *source* à Tesnière afin de remplacer celui d'antécédent. Cela permet ainsi de ne plus prendre en compte la localisation du terme anaphorisé.

Une dernière limite de la conception textuelle, que nous citons, réside dans les relations coréréférentielles qu'entretiennent les expressions anaphoriques avec leurs antécédents. Selon Kleiber (1988 : 3), l'anaphore est « un processus référentiel où une expression anaphorique renvoie à un référent déjà mentionné dans le discours ». Dans ce type d'approche, l'anaphore se cantonne aux rapports coréférentiels. Par ailleurs, elle implique l'assimilation des deux notions et ce phénomène est rejeté par certains auteurs linguistes. Par exemple, l'anaphore et la coréférence se distinguent chez Milner qui met en avant leurs propriétés divergentes. En effet, comme précédemment évoqué, la relation anaphorique entre deux segments textuels s'assimile à une relation asymétrique, non réflexive et intransitive. En ce qui concerne la coréférence, Milner (1982 : 32) affirme que dans le cas de coréférence « la relation est manifestement symétrique et transitive ; il n'est pas dépourvu de sens de la tenir pour réflexive ». Ainsi, dans le cas de la coréférence, la relation entre les deux unités A et B est symétrique, réflexive et transitive. En effet,

> les deux termes concernés peuvent être homogènes ou non quant à leur nature catégorielle : on aura en particulier des paires homogènes N''/ N'' ou pronom / pronom et des paires hétérogènes N''/ pronom- et réflexive : une unité référentielle peut être dite coréférentielle d'elle-même. (Milner 1982 : 32)

Kleiber (1988 : 4) précise que « l'absence de coréférence ne signifie pas absence de relation référentielle entre l'expression anaphorique et son antécédent ». Il ajoute que

> cette dernière remarque invite à ne pas séparer totalement coréférence et anaphore. Il est pertinent de considérer qu'il y a des coréférences anaphoriques et non anaphoriques et des anaphores coréférentielles et non coréférentielles. (Kleiber 1988 : 4)

Comme une anaphore n'est pas toujours coréférentielle, cette dernière remarque prend tout son sens en distinguant anaphore et coréférence. A partir de cette définition, une relation coréférentielle ne peut pas comporter une relation anaphorique, alors que l'inverse est possible. Au vu des ambiguïtés engendrées par la coréférence, l'approche textuelle a délaissé cette notion. Ainsi, Kleiber met en avant de nombreuses difficultés quant à la localisation du référent (dans le discours ou dans le site d'énonciation immédiate). Ceci implique la redéfinition des notions d'anaphore et de deixis et de leur interprétation, où la localisation de l'unité (textuelle ou non) n'est plus un critère « non subsumant ».

4.3. Approche mémorielle

4.3.1. Présentation

La conception mémorielle a été élaborée par des auteurs issus du courant cognitiviste, notamment par Bosch (1985), Cornish (1986, 1990) et Ehlich (1982). Cette conception s'intéresse aux situations de communication réelles où le locuteur et l'interlocuteur interagissent. Dans le but de faire ressortir les propriétés communes aux expressions référentielles, qu'elles soient anaphoriques ou exophoriques, cette conception s'intéresse à l'organisation intrinsèque des expressions référentielles. Ainsi, cette nouvelle approche, dite mémorielle ou mentaliste, donne la priorité à l'opération de maintien du référent et non plus à l'opération de renvoi vers l'antécédent. L'acte de communication, qui implique des expressions anaphoriques déclenchant des procédures distinctes par nature, exige ici de distinguer l'anaphore de la deixis. Selon les concepteurs de cette approche, il existe deux procédures dans des situations de communication réelle : l'une anaphorique et l'autre déictique. Selon Ehlich 1982 (cité par Corblin 1995), la première est :

> un instrument linguistique pour faire maintenir au récepteur une focalisation antérieurement établie sur un item spécifique vers lequel il a orienté son attention

> auparavant. La procédure anaphorique est réalisée au moyen d'expressions anaphoriques.

Tandis que la seconde : la procédure déictique

> est un instrument linguistique pour réaliser la focalisation de l'attention du récepteur sur un item spécifique qui appartient à l'espace déictique pertinent. La procédure déictique est réalisée au moyen d'expressions déictiques.

Ainsi, une classe d'expression linguistique est attribuée à chaque procédure ce qui entraine la résolution de certaines limites rencontrées dans la conception textuelle. En effet, en associant des procédures particulières à des catégories spécifiques d'expressions linguistiques (utilisation des pronoms de première et deuxième personnes et de noms propres au niveau du discours), elle est en opposition avec l'approche textuelle puisque cette dernière ne s'intéresse pas aux termes linguistiques qui rendent possible la distinction entre les relations anaphoriques et exophoriques. Même si dans certains cas, comme par exemple des dialogues où l'on retrouve des pronoms personnels de première et deuxième personnes, quelques expressions sont relevées dans le discours, ce manque n'a pas été relevé par l'approche textuelle et ces expressions ne sont en rien envisagées comme anaphoriques. Selon Ehlich, le rôle de l'expression déictique est de focaliser le récepteur sur un élément particulier alors que l'expression anaphorique maintient son attention sur un élément de la situation de communication qu'elle soit linguistique ou contextuelle. L'intention du locuteur est un point essentiel selon l'auteur, quelque soit le mode de focalisation employé : segment du discours ou situation. En ce sens, l'auteur confère une dimension cognitive à la relation existant entre procédures et expressions linguistiques. Le tableau ci-dessous confirme cette idée ;

Procédure	Expression	Champ	Focalisation
Déictique	Démonstratif	Situation Texte	Instaurée
Anaphorique	Pronom Défini	Situation Texte	Maintenue

Tableau 2 : Relation entre procédures et expressions linguistiques d'après Ehlich (1982).

La conception mémorielle implique la réinterprétation de la distinction de l'anaphore et de la deixis, ce qui n'est pas le cas de la conception textuelle. Autrement dit, les critères

de texte et de situation immédiate pour définir l'anaphore et la deixis disparaissent au profit de ceux de saillance[34] et de nouveauté. Ainsi, ce sont les éléments saillants, dans ce cas le référent est déjà connu, qui définissent le processus d'anaphore. Dans le cas où l'introduction d'un référent en mémoire immédiate serait obligatoire, il s'agit de la deixis. Notons que la saillance est assimilée au processus de l'anaphore, alors que la nouveauté est assimilée à celui de la deixis. C'est alors l'accessibilité du référent qui est utilisée comme critère dans la conception mémorielle.

Dans cette nouvelle proposition qualifiée de « solution mentaliste » par Charolles et Schnedecker (1993), à mesure que le discours progresse, le contenu référentiel de la représentation mentale élaborée est accessible grâce aux expressions référentielles qui seraient des déclencheurs. La rencontre avec une expression référentielle anaphorique ne renverrait pas au déclencheur d'antécédent présent à la surface du discours, mais plutôt à la forme de l'antécédent présente dans la représentation mentale, qui évolue en fonction des prédicats qui lui sont progressivement associés. L'attribution de la référence passerait seulement par leur « corrélat » (Charolles/Schnedecker 1993 : 11) dans la représentation mentale du texte et non pas par les formes de surface. En effet, la localisation du référent est abandonnée par l'interlocuteur au profit du mode de connaissance de celui-ci. Il y a une redéfinition complète des anaphoriques et des déictiques dans cette nouvelle approche et les expressions déictiques deviennent anaphoriques et vice-versa. Selon Ehlich (1982) l'anaphore est « une expression qui, à l'intérieur de toute action linguistique (texte ou discours) opère un renvoi sur un élément préalablement mis en focus et connu des deux interlocuteurs. » et permet ainsi de mettre en évidence la transformation de l'opposition anaphore / deixis. L'auteur illustre ces propos avec l'exemple suivant où l'anaphore est dépourvue d'antécédent :

> [23] Il est 5 h 20, A et B sont assis dans une pièce. Ils attendent C qui leur a promis de venir à 5 h précise. Ils attendent en silence depuis que 3 h a sonné. Soudain, A entend des pas dans la cage d'escalier. Il dit alors à B :
>
> - **Il** arrive. (Résumé Ehlich 1982 : 330)

Dans le cas d'un référent saillant dans la situation d'énonciation de l'occurrence référentielle, l'exophore mémorielle devient anaphorique dans la conception mémorielle. C'est la continuité de la saillance qui permet l'apparition du caractère anaphorique de l'occurrence. Quand le contenu propositionnel de la référence n'est pas encore saillant,

[34] Nous reviendrons sur cette notion, en détail, dans le deuxième chapitre de la deuxième partie.

alors l'anaphore résomptive devient déictique. La deixis attire l'attention sur un référent nouveau. Ainsi, la définition mémorielle de l'anaphore reste applicable aux cas paradigmatiques d'anaphore et de deixis de la version textuelle. Ceci est aussi applicable même si le référent est saillant, dans le sens où il est déjà identifié par l'interlocuteur. L'espace endophorique, dans la conception mémorielle, ne représente aucune contrainte. En effet, c'est sur des facteurs cognitifs que se base le caractère anaphorique de cette conception. Si nous nous intéressons à l'opposition anaphore / deixis, c'est la saillance du référent qui est mise en évidence dans cette approche, que celui-ci soit employé dans le texte ou non. Notons que la saillance peut avoir deux origines, soit par une évocation antérieure du référent soit par sa perception antérieure dans la situation d'énonciation.

Selon Reboul (1989 : 84), c'est à partir de la représentation mentale construite à partir du discours que l'on peut désigner un référent à l'expression anaphorique. Ceci veut dire que c'est dans la mémoire immédiate de l'interlocuteur, qui peut être enrichie par la situation extralinguistique ou le texte, que la complétude référentielle s'acquière. L'approche mémorielle, à l'opposé de la textuelle, offre une vue unitaire de certains marqueurs référentiels. En effet, comme c'est l'interlocuteur qui a une représentation mentale de l'assignation du référent à l'expression anaphorique, le changement d'état du référent dans le discours ne pose pas de problème dans cette approche. Quant aux déictiques, en effet, elle préconise de ne pas considérer le caractère nouveau du référent. Notons par ailleurs que cette conception mémorielle présente certaines limites.

4.3.2. Limites de l'approche mémorielle

Dans cette nouvelle approche, la dichotomie anaphore / deixis a vu un changement de catégorisation et malgré cela, la distinction entre texte et situation peut être qualifiée de pertinente, en permettant de déterminer l'origine de la saillance du référent. Effectivement, la saillance trouve son origine soit dans un texte quand il y a mention antérieure du référent, soit dans un site non textuel dans le cas de la situation d'énonciation. Le processus de l'anaphore peut trouver son origine dans l'environnement extralinguistique immédiat et le texte. Ces deux localisations alimentent ce qui est appelé *mémoire discursive*.

Dans le cas de certaines expressions référentielles (citons par exemple les emplois inférentiels, le nom propre, ou les embrayeurs comme *je, tu, etc.)*, l'échelle d'accessibilité ou encore les analyses non unitaires, l'approche mémorielle présente quelques limites bien que la perspective de cette démarche soit différente de celle textuelle. Kleiber (1992a)

montre, en utilisant des illustrations traitant des inférences, qu'il n'est pas toujours évident de reconnaitre les cas d'anaphores grâce au critère de la saillance. Cela est démontré par les exemples où il montre que, dans la représentation mentale du discours, la présence d'un antécédent n'est pas suffisante pour autoriser une reprise pronominale :

[24] Nous arrivâmes dans **un village**. **L'église** était située sur une butte. (Résumé Kleiber 1992a : 6)
[25] Paul a acheté **une Toyota** car **elles** / **ces voitures** sont économiques. (Résumé Kleiber 1992a : 50)

La principale critique de cette solution selon Charolles/Schnedecker (1993) est le fait que des contraintes linguistiques gérent l'emploi des reprises anaphoriques. Les pronoms devraient en effet pouvoir donner accès à des entités présentes dans le modèle mental (qui ne sont pas indiquées explicitement à la surface du texte) seulement s'ils sont interprétés en fonction du modèle mental élaboré au cours de lecture, une idée soutenue par Reichler-Beguelin (1989). En effet, dans les cas où l'inférence est indirecte, comme dans les exemples précédents où les situations correspondant aux entités inférables, le critère de saillance n'est plus pertinent (Kleiber 1992a). Le SN *l'église* de l'exemple (24) est une partie inférée du tout *village* et représente un exemple d'anaphore associative. Par ailleurs, le SN *ces voitures* (une sous classe générique) de l'exemple (25) est inférée du SN *une Toyota* et représente quant à lui une illustration d'anaphore hyperonymique. Ainsi, ces deux exemples illustrent le manque de pertinence du critère de la saillance et les limites de la conception mémorielle. Une entité différente de la première mention de référent infère l'objet de référence, ainsi celui-ci ne peut pas être déjà saillant. Le critère connu / nouveau n'est pas suffisant pour définir les marqueurs référentiels anaphoriques et déictiques, au vu des limites de l'opposition anaphore / deixis. Notons cependant, que malgré cette carence, la pertinence du critère d'accessibilité de l'approche mémorielle reste plausible :

> i) en montrant que l'anaphore (textuelle) est avant tout un phénomène de mémoire immédiate, où le texte sert simplement d'introducteur et non de champ de recherche pour le bon antécédent ;
> ii) en orientant, par l'intermédiaire du trait nouveau ou saillant, la description des expressions référentielles d'entités « nouvelles » vers la référence indexicale. (Kleiber 1991 : 15)

Dans l'approche textuelle, l'hypothèse de la référence indexicale des déictiques n'est pas autant contrastée que dans l'approche mémorielle. C'est à partir du texte que l'on déduit la référence, ceci est montré par la détermination démonstrative (exemple 24). Rappelons que dans la conception mémorielle, le pronom *il* est traité de manière homogène relativement à la conception textuelle et ceci représente un des avantages de cette

approche. Ainsi,

> toutes ces difficultés rendent impraticables une définition générale des marqueurs référentiels en anaphoriques et déictiques à l'aide du critère connu ou donné / nouveau. L'opposition mémorielle ainsi conçue amène à la version standard à des analyses non unitaires indésirables des marqueurs référentiels et, surtout, se révèle trop pauvre pour maîtriser la diversité des situations de saillance référentielle rencontrées. (Kleiber 1992a : 15)

Il existe une grande diversité de mécanismes référentiels, et la définition de la conception mémorielle est trop insuffisante pour tous les maitriser. C'est grâce à l'environnement spatio-temporel de l'occurrence du référent et à l'aspect cognitif nouveau du référent désigné qu'il est possible de pallier à la difficulté de la saillance préalable du référent (Kleiber 1992b). Enfin, à la surface du discours, la représentation mentale ne détermine pas complètement les occurrences référentielles, même si les pronoms anaphoriques s'interprètent au niveau de cette représentation. Quand elle se base seulement sur le fonctionnement des représentations cognitives, une analyse du fonctionnement des anaphores pronominales (ou non pronominales) est incapable d'expliquer la totalité des propriétés que l'on observe.

Le bien-fondé d'une solution faisant intervenir la représentation mentale du discours est cependant admis par les auteurs pointant l'inadéquation entre la solution mentaliste et les contraintes linguistiques qui contrôlent les différentes formes de reprises anaphoriques (Charolles/Schnedecker 1993). Ils sont aussi en accord avec le fait que, dans la forme de surface du discours, cette représentation n'est pas restreinte aux informations explicites. Par ailleurs, ils affirment que la structure référentielle de la forme de surface obéit aussi à des règles de formation linguistiques qui sont conditionnées par la structure lexicale et syntaxique de cette forme de surface. Elle n'obéirait pas uniquement à des règles de formation dépendant du contenu du message comme c'est le cas dans la représentation mentale.

Nous avons souligné les limites et les divergences des approches textuelle et mémorielle, néanmoins, ces deux approches restent complémentaires ; de ce fait les critères textuel et d'accessibilité conservent leur propre pertinence. Afin d'allier les deux approches, Kleiber (1992 et 1994) propose une approche pragma-sémantique, de nature sémasiologique. Cette approche propose de

> s'occuper prioritairement des expressions anaphoriques elles-mêmes et non remonter à la mention antérieure qui a permis de les interpréter. Cela suppose

> évidemment que l'on règle au préalable le problème du fonctionnement particulier de chacune de ces expressions anaphoriques. (Kleiber 1992 : 18)

4.4. Bilan

Notre étude parallèle de la littérature souligne l'importance de la mise en place d'une approche pluridimensionnelle. Nous allons opter pour les trois approches, accompagnées de leurs limites (citées plus haut). Les approches traitant de la notion d'anaphore se basent sur des modalités complémentaires dans le but d'expliquer l'anaphore. L'essence de notre approche est dans la prise en considération de facteurs cognitifs qui ont amené certains linguistes à rendre compte des anaphoriques.

En effet, l'introduction de facteurs cognitifs permet de mieux caractériser l'anaphore, puisque l'approche syntaxique, à elle seule, entrave l'étude complète du phénomène anaphorique. Selon Kleiber (2001 : 34), il existe une distinction cognitivo-discursive qui est à l'origine de la séparation des deux approches textuelle et mémorielle. La structure informationnelle du modèle contextuel est balisée de manière prégnante par un fait discursif réel qui est l'introduction explicite d'une entité par le texte. En revanche, une introduction implicite non textuelle n'a pas de poids coercitif. En effet, elle ne reçoit sa légitimé de saillance que plus tard, étant donné que ce n'est qu'un fait potentiel. Par ailleurs, notons que les trois approches sont complémentaires et notre étude de l'approche textuelle[35] n'enlève rien de la pertinence de l'approche mémorielle[36] et l'approche substitutive[37]. Nous utilisons les trois approches quand cela s'avère nécessaire[38]. Notre travail se place dans une approche cognitive et fonctionnelle mais il ne met pas de côté pour autant les aspects syntaxiques et sémantiques.

Un des principaux objectifs de notre recherche consiste à décrire et expliquer le comportement ambigu des anaphores pronominales dans les textes français. En effet, nous prenons en compte les effets cognitifs et textuels de l'emploi de *il* dans notre corpus $\text{RESUMAN}_{\text{C}}$. Dans la mesure où nous souhaitons mettre en évidence les heuristiques de résolutions des cas ambigus, nous optons pour une étude sur corpus. C'est ce que nous allons présenter au chapitre suivant.

[35] Premier chapitre de la deuxième partie.
[36] Dont la notion de saillance sera abordée dans le deuxième chapitre de la deuxième partie.
[37] Troisième chapitre de la deuxième partie.
[38] Troisième partie de notre étude.

Chapitre 2 : Méthodologie et présentation du corpus

S'il est un homme tourmenté par la maudite ambition
de mettre tout un livre dans une page,
toute une page dans une phrase,
et cette phrase dans un mot, c'est moi.

JOUBERT (8 février 1815)

La constitution de corpus, élément primordial pour une thèse en sciences du langage, est une démarche nécessitant un travail préparatif rigoureux. Comme notre étude a des fins automatiques, nous avons constitué un corpus automatisé permettant d'analyser le fonctionnement des anaphores pronominales ambigües dans un cadre textuel bien délimité selon des critères que nous exposerons ultérieurement. Nous avons nommé notre corpus $RESUMAN_C$, ce qui reflète à la fois le genre textuel choisi (les résumés) et le phénomène linguistique étudié (l'anaphore en général).

Avant de brosser le tableau des différentes phases de l'élaboration de $RESUMAN_C$, nous nous proposons de donner, dans la première section, les principales caractéristiques de la tradition des corpus. La linguistique de corpus se définit comme une branche de la linguistique reposant sur l'utilisation raisonnée de technologies informatiques. Au vu des difficultés existant auparavant pour rassembler et annoter manuellement en contexte des mots et des exemples de leurs emplois, la linguistique de corpus apparait aujourd'hui, plus que jamais, comme un outil indispensable dans le domaine des sciences humaines. Ainsi, les premiers corpus sont apparus sous version papier et le premier corpus *papier* était représentatif de l'anglais britannique. Par la suite, on est passé du papier à l'octet et la transformation des corpus sous forme électronique a représenté un progrès considérable. Cette nouvelle technologie informatique a largement facilité l'analyse des divers faits ou problèmes auxquels l'utilisateur pouvait se confronter pour ensuite les évaluer statistiquement. La seconde section exposera une analyse de terrain des corpus anaphoriques existants. Nous y expliciterons l'évolution de ces corpus depuis le tout premier, à nos connaissances, jusqu'à celui actuel en détaillant leurs importances et leurs limites. Cette section montrera les fondements de notre choix d'étudier les anaphores

pronominales, malgré leur réputation comme du *déjà beaucoup étudié*[39]. Dans la troisième section, enfin, nous présenterons $RESUMAN_C$: les choix effectués pour définir sa source, sa taille, sa qualité et son genre. Nous détaillerons aussi notre méthodologie de fouille des textes et les outils utilisés.

1. Le rôle du corpus en linguistique

Il est intéressant de noter que le champ d'action des corpus est très large et que son usage n'est pas réservé exclusivement aux linguistes et aux lexicographes. Il intègre ainsi d'autres disciplines telles que la sociologie, la sociolinguistique, la traduction, le domaine des médias et de la communication de masse, la psychologie parmi d'autres.

1.1. Définition du corpus en linguistique

Afin de mener une analyse pertinente sur le corpus, nous pensons qu'il est indispensable, avant tout, de le définir. Le corpus est une compilation, soumise à des principes, de données du langage empirique à travers des textes (ou des fragments de textes) considérés comme échantillons d'un discours donné pour lequel ils révèlent une valeur représentative. En effet, la notion de corpus soulève des problèmes, connus des sciences du langage, qui n'ont toujours pas abouti à un consensus qui permettrait de converger vers une définition univoque faisant figure de référence lors de chaque justification d'un corpus d'analyse.

Pendant longtemps, certains linguistes – notamment les génératives – ne privilégiaient pas les corpus dans leur pratique scientifique, voire la confrontation de leur démonstration ou de leur modèle avec des données attestées, recueillies et structurées en corpus. Cela pourrait s'expliquer selon Bouquet (2005) par une interprétation – on ne peut plus stricte – de la dichotomie saussurienne, qui soutient que « la linguistique a pour unique et véritable objet la langue envisagée en elle-même et pour elle-même » (Saussure 1916 : 314). De surcroît, le corpus était présenté comme un recueil d'énoncés, un recueil de performances individuelles produites par des locuteurs dans une réalité sociale et historique donnée et pouvait être négligé, voire ignoré. En effet, la tâche principale des linguistes était d'étudier la langue, le système et la compétence linguistique. Il est à noter que les corpus, tout particulièrement les corpus textuels, s'inscrivaient plus dans une linguistique de la parole considérée comme de la sociolinguistique ou de la

[39] L'anaphore pronominale est un fait beaucoup étudié certes, mais nous verrons que cette abondance est contrecarrée par la rareté des corpus annotés en anaphores d'où l'existence de ce travail.

psycholinguistique, une forme d'analyse de discours au même titre que la pragmatique ou la stylistique.

La déclaration de Chomsky (1999), pour qui « la linguistique de corpus n'existe pas[40]» (Chomsky, entretien avec Baas Aarts, cité par Rastier 2005 : 40), est une illustration des grandes polémiques existant sur la question d'une linguistique sans ou hors corpus. Cette dénégation s'inscrit dans la dissension opposant théoriciens et descriptivistes que connaissait le monde anglo-saxon. En corpus, la grammaire universelle est sujette à de nombreux facteurs tels que la culture, les fluctuations d'humeur du locuteur, les choix, les sélections de l'analyste, etc. Les corpus de données attestées déroutent le théoricien, car ils brouillent le système et n'ont pas toujours le mérite de le révéler.

A l'inverse, d'autres linguistes pensent que la linguistique peut se réduire à l'observation de données réelles, voire au simple recueil desdites données : ainsi, l'utilisation des corpus deviendrait obligatoire voire suffisante. Dans ce contexte précis, Scheer (2004a et b), a mis en relief les défaillances de cette linguistique empirique dominée par le béhaviorisme dans le troisième numéro de *Corpus*. Les descriptivistes sont allés plus loin en soulignant que le système se réduisait à des réalisations multiples, variées, imprévisibles qu'il fallait répertorier dans des macro-corpus. En résumé, il était seulement question d'écarts par rapport aux règles, de cas particuliers que l'on trouvait dans des corpus oraux ou écrits assez typiques, et non de règles ou de structures.

Un corpus écrit est constitué des mots qui possèdent déjà une signification mais n'acquièrent de sens et de valeur qu'à l'intérieur d'un contexte. Par exemple, le TLF (le Trésor de la Langue Française), issu de cette tendance, est un dictionnaire tout à fait original qui attribue des définitions à des mots en fonction d'usages et d'exemples trouvés dans la littérature française des origines modernes à nos jours et non à partir d'un sens déjà-là ou construit de manière logique (par l'étymologie par exemple). Le chercheur rassemble dans des concordanciers automatiques toutes les phrases ou tous les paragraphes contenant le mot étudié dans le corpus en étant assisté de l'outil informatique et des facilités qu'offre l'hypertextualité. Par ailleurs, grâce à la statistique textuelle, il fait ressortir les contradictions des occurrences des unités linguistiques du corpus après les avoir relevées et sélectionnées[41].

[40] Texte original "corpus linguistics does not exist".
[41] Nous reviendrons sur les outils utilisés pour ce faire à la 3ème section de ce chapitre.

Devant ces attitudes assez disparates et afin de dissiper toute confusion quant à l'emploi du terme *corpus* dans des acceptions bien différentes ou encore dans son instrumentalisation, il nous semble nécessaire de parvenir à une réflexion de synthèse. En effet, le terme est désormais largement répandu et l'éventualité que la *linguistique de corpus* telle qu'elle est définie depuis plusieurs années dans le monde anglo-saxon ou en France et qui a œuvré – il faut l'admettre – à limiter la réflexion et à imposer le corpus en objet, se fond aujourd'hui dans une linguistique générale qui n'a pas, dès le départ, les mêmes centres d'intérêts.

En effet, nous tenons à présenter l'étude de Mellet (2002), car elle met en perspective comparative le corpus « clos et exhaustif » et le corpus « échantillonné ». Mellet (2002) a essayé d'interroger les pratiques du corpus. D'après elle, un corpus ne peut être considéré comme *clos et exhaustif* :

> que dans le cadre d'une monographie, auquel cas il sera étudié en tant que tel, sans prétendre à être représentatif d'autre chose que de lui-même ni à ouvrir sur aucune forme de généralisation ou modélisation. (Mellet 2002 : §3)

Ainsi, l'une des questions les plus récurrentes posées dans la littérature porte sur la définition du corpus et sa clôture. Un tel corpus, généralement très homogène, se retrouve le plus souvent dans les études stylistiques ou en analyse du discours. Il faut tout de même souligner qu'il est difficile d'avoir un corpus qui respecte ces critères (clôture et exhaustivité). Ceci est particulièrement vrai dans le cas où la progression de la recherche et de l'analyse elle-même nécessite souvent d'introduire petit à petit au corpus initial des notes d'auteur, des commentaires critiques, et même d'autres œuvres littéraires à des fins de comparaison.

Il existe un autre type de corpus appelé « corpus échantillonné » (Mellet 2002) où le problème de « la représentativité » du contenu du corpus prime dans ce cas sur celui de son exhaustivité car sa forme n'est ni homogène ni complète. Son objectif est alors de constituer des exemples qui rendent compte d'une réalité plus étendue et renfermant plusieurs enjeux. Tout d'abord, celui de cerner et mettre en exergue cette réalité trop vaste pour être appréhendée dans sa totalité (par exemple une étude de l'oral français contemporain). Ensuite, celui de mettre en place des bases empiriques permettant de répondre à un questionnement théorique ou soutenir une hypothèse structurale (par exemple décrire, comprendre et unifier les emplois du conditionnel dans le système verbal français contemporain). Finalement, celui d'ériger les bases de connaissances nécessaires

au développement des nouveaux outils essentiels pour la progression du TALN (annotation des corpus pour des résolutions automatiques).

Pour le traitement du corpus, nous pouvons retenir les principaux problèmes suivant soulevés par Charaudeau (2009 : 6-37) :

- celui relatif au recueil des données : ce recueil peut se faire de différentes manières : exploration de terrain, procédés d'enregistrement libres ou imposés, à l'instar des acteurs de parole, etc. Le choix du procédé est soumis au choix de la matière langagière (acte de parole oral ou écrit), du support qui rend compte de ces paroles et leur relation avec une situation de communication (pour l'écrit : lettres, rapports, journaux, affiches, etc. ; pour l'oral : radio, télévision, conférences, conversations du quotidien, etc.) ;

- celui relatif à l'importance du matériel recueilli et à la valeur représentative que l'on peut lui reconnaitre. Il est essentiel de préciser si ce matériel peut être considéré comme un objet en soi ou comme un simple outil, en d'autres termes, s'il est ouvert ou clos. On note que dans le dernier cas l'exploitation des données peut se faire même si la clôture du corpus est revendiquée, à titre expérimental, par certains analystes. Ainsi, l'hypothèse de l'exhaustivité, très défendue par l'attitude positiviste, n'a plus lieu d'être ; ceci en dépit du développement auquel assiste ladite *linguistique de corpus* introduite dans le monde anglo-saxon. Si on suppose que le corpus est partiel, alors se posent les problèmes de sa valeur comme échantillon et de la possibilité de le décomposer en sous-corpus. Enfin, la possibilité qu'il soit un objet en soi ou un instrument nous renverrait à la question du contexte que nous analyserons plus loin ;

- celui relatif, à l'intérieur du matériau langagier recueilli, aux catégories qui vont faire l'objet de l'analyse : grammaticales (connecteurs, pronoms, verbes, etc.), lexicales (par champs ou de façon aléatoire), syntaxiques (selon divers types de construction) ; mais aussi les variables dépendantes de la production des actes langagiers, comme les types de locuteurs, les procédés de communication, ainsi que les variables qui se rapportent au temps (l'historicité) et à l'espace (les cultures) ;

- celui relatif à l'outil de traitement des données : analyse, traitement informatique au moyen de logiciels conçus spécialement, mise en place d'échantillons à partir de bases de données.

En résumé, dans tous ces cas, la question de la pertinence des choix et de leur rapport avec les présupposés théoriques, et l'éventuelle circularité qui peut s'établir entre ceux-ci et le corpus, est de mise. Ce descriptif nous permet d'avoir des critères de choix efficaces

pour l'élaboration de notre corpus, notamment pour y mettre en relation le type de textes choisi et le phénomène étudié. Nous partageons en partie les conclusions de Mellet (2002) concernant la sélection des données qui pourra être soumise à une orientation en fonction de l'objectif fixé en ce qui concerne le corpus échantillonné. Notre objectif étant la résolution automatique de l'anaphore pronominale, y compris ambigüe, dans $RESUMAN_C$, nous oriente vers un choix de méthodologie automatisée. Cela nous a permis de nous positionner et d'adopter la création d'un corpus en ligne et par la suite exclure l'hypothèse d'utiliser un corpus en papier qui nécessiterait d'ample travail de traitement. De notre côté, il ne s'agit pas d'un travail réalisé dans une perspective définitoire représentative du phénomène anaphorique mais plutôt d'une étude menée dans une perspective informatique outillée, partant d'une analyse sur corpus.

1.2. Linguistique de corpus ou Linguistique sur corpus ?

Si certains chercheurs privilégient la linguistique de corpus, d'autres proposent de parler, plutôt, de la linguistique sur corpus. Deux auteurs retiendront particulièrement notre attention : Mayaffre et Tognini-Bonelli.

Mayaffre (2005a), qui distingue dans son article, d'une « manière hiérarchique », trois grandes catégories de corpus : les corpus lexicographiques qui sont, à la fois, des corpus clos et finis, car très exhaustifs, et qui peuvent englober de nombreux éléments traités par les dictionnaires ; les corpus phrastiques de grammairiens ou de syntacticiens, qui ont la particularité de recueillir des exemples élaborés, c'est-à-dire non authentiques et les corpus textuels réunissant toujours des données attestées, du fait que l'on ne peut concevoir artificiellement un texte pour en appréhender le sens et qui ne sont alors ni exhaustifs ni représentatifs. Ainsi, comme le souligne Mayaffre (2005a : §9), le corpus s'apparente à un « lieu linguistique où se construit et s'appréhende le sens» des textes. Notons que pour certains linguistes, comme Mayaffre, le corpus a un rôle déterminant dans la linguistique contemporaine puisque la matière du linguiste est non seulement le texte, mais aussi le corpus textuel.

Différentes attitudes scientifiques, permettant d'appréhender la discipline, peuvent être adoptées face à un corpus. Il faut alors nous interroger sur les termes exacts que l'on est appelé à utiliser et se demander à propos de son statut : est-il un outil permettant d'examiner un phénomène transcendant ou est-il lui-même un phénomène observé, examiné, vu son intérêt en lui-même ? En d'autres termes, il s'agit de savoir si le corpus contribue à construire tout simplement un sens et serait alors une méthode permettant de

dégager des règles d'étude, autrement dit, une heuristique permettant d'aboutir à des modèles sémantiques à envisager ou alors s'il permet de rendre compte d'un sens qui existerait à priori comme le présenterait un quelconque document de type recueil d'exemples, base de données, échantillons de langue. Les pionniers de cette branche de la linguistique ont été surtout dans les pays anglophones. La linguistique de corpus, selon la conception de Habert *et al.* (1997) ou Rastier (2001), privilégie le corpus par rapport aux textes dont il est composé : Rastier, « l'un des rares auteurs à avoir théorisé, par devant le texte, les corpus (textuels) en linguistique » (Mayaffre 2005b : 5), déclare que :

> tout texte placé dans un corpus en reçoit des déterminations sémantiques, et modifie potentiellement le sens de chacun des textes qui le composent. (Rastier 2001 : 92).

À travers ces différents types de corpus se pose la question polémique de l'objet pertinent de la linguistique. En linguistique de corpus, au sens le plus étroit, le texte devient fondamental puisque l'idée est que l'objet du linguiste est le texte et non la phrase. A ce propos, Rastier souligne que :

> la linguistique a non seulement pour objet empirique mais pour objet théorique cette unité de communication-interaction langagière qu'on appelle un TEXTE (ou un DISCOURS). (Rastier 2001 : 92)

Quant à la linguistique sur corpus (*corpus-based*), elle se distingue de la linguistique de corpus (*corpus-driven*), selon Tognini-Bonelli (2001), par le fait qu'elle utilise généralement des données de corpus afin d'explorer une théorie ou une hypothèse, visant à la valider, la réfuter ou l'affiner. Elle souligne que le corpus, lui-même, incarne une théorie du langage (Tognini-Bonelli 2001 : 84-5). Ainsi, la principale différence entre les deux est que la linguistique sur corpus part d'une théorie préexistante, qui est testée à l'aide des données du corpus, tandis que la linguistique de corpus construit la théorie étape par étape, en présence de la preuve. L'observation de certains modèles conduit à une hypothèse, qui à son tour conduit à la généralisation des termes de règles d'usage et trouve enfin l'unification dans un statut théorique (Tognini-Bonelli 2001 : 17). Selon Tognini-Bonelli, la linguistique sur corpus donne la priorité à la déclaration théorique préexistante et, malgré l'apport de la variabilité de la langue naturelle, elle tente de « l'isoler, la normaliser et la réduire » (Tognini-Bonelli 2001 : 67). L'auteure reconnaît qu'il n'existe pas de différence claire entre la linguistique de corpus et la linguistique sur corpus. Ainsi, en partant de la sélection de l'objet de l'enquête jusqu'à l'interprétation des résultats, l'intuition

joue inévitablement un rôle dans tout type de recherche. D'ailleurs, comme Tognini-Bonelli le suggére (2001 : 84), la recherche sur corpus pourrait être appliquée à l'intégrité des données dans l'ensemble du corpus ou pourrait viser à être exhaustive en ce qui concerne le corpus comme preuve. Plusieurs traductions de linguistique sur corpus (*corpus-based*) représentent des contre-exemples : Saldanha (2004) montre que, aussi longtemps que nous sommes prêts à revoir nos théories à la lumière des données si cela est nécessaire et que les exceptions à la norme sont également prises en compte, l'utilisation d'hypothèses préexistantes n'est pas un problème en soi[42].

Nous n'avons pas évoqué tous les traits définitoires de la linguistique de/sur corpus dans cette section. Ce qui nous concerne, est de préciser dans quel cadre s'inscrit notre recherche. Nous partons du fait qu'il y a des anaphores ambigües dans $\text{RESUMAN}_{\text{C}}$, qui nécessitent une résolution soit humaine soit automatique[43]. Nous avons choisi des résumés électroniques d'une centaine d'œuvres de la littérature française et nous examinerons les cas d'ambiguïté anaphorique pronominale : les facteurs, les impacts et les solutions de résolution. Il s'agit, donc, de linguistique sur corpus, puisque nous savons ce que nous cherchons à analyser. Nous nous pencherons donc, dans la section suivante, sur l'examen des particularités des corpus anaphoriques existants, leurs évolutions et leurs impacts sur le choix de $\text{RESUMAN}_{\text{C}}$.

2. Ressources en corpus anaphoriques

Muzerelle *et al.* (2013 : 2) montrent dans leur article que l'apparition de plusieurs applications opérationnelles, qui sont au centre de notre recherche, visant aussi bien le grand public que les professionnels, est la conséquence de l'essor considérable de l'ingénierie des langues de ces deux dernières décennies. Dans ce contexte et dans la lignée des diverses technologies langagières, la recherche d'information et l'indexation de documents représentent incontestablement des champs applicatifs à l'avenir très prometteur. Cela dit, il faut mettre en place des dispositifs de structuration et d'interrogation automatique intelligents afin de développer exponentiellement les ressources textuelles ou multimédias accessibles sur Internet. Cette masse documentaire consultable sur Internet nécessite d'être indexée afin de faciliter son exploitation. En revanche, nous assistons de nos jours à une explosion colossale de textes numériques

[42] Voir aussi comme exemple Kenny (2001).
[43] Nous reviendrons, dans la troisième partie, avec plus de détails sur les raisons de la sélection du type de la résolution.

publiés chaque instant sur la Toile. Malgré l'indexation automatique de certains textes, qui restent tout de même minoritaires, les tâches d'analyse automatique ou de fouille de texte restent difficiles à mettre en œuvre.

La capacité des outils d'indexation ou d'interrogation (destinés à la réalisation de ces tâches) à détecter des entités anaphoriques, unités linguistiques qui renvoient à un élément précis de l'univers du discours, conditionne leur performance. Ainsi, du fait qu'elles répondent aux questions principales (qui ? quoi ? où ? quand ?) nécessitant une information précise, la détection des entités anaphoriques est indispensable dans les applications d'extraction ou de recherche d'information textuelle[44].

Actuellement, avec des taux de précision relatifs, les systèmes les plus performants sont capables de détecter et d'annoter les relations anaphoriques. Les travaux actuels en TALN ambitionnent le suivi des chaînes référentielles dans un document donné. Examinons le texte suivant, dans lequel sont soulignées les entités lexicales de la relation anaphorique :

> [1] L'histoire débute durant l'été 1922. Un enfant découvre une vipère. Il la saisit et l'étouffe de ses mains. Cet acte lui vaut d'être comparé à Hercule, le personnage de la mythologie grecque, qui dans son berceau étrangla deux serpents. 25 ans plus tard, Jean Rezeau, "l'enfant de 1922" est le narrateur de l'histoire. Son surnom est Brasse-Bouillon. Il évoque la propriété de sa famille, La Belle Angerie. (Résumé Vipère au poing)

Dans cet exemple, il est nécessaire de résoudre les relations anaphoriques existantes entre le pronom *il* et les lexiques *Un enfant, Jean Rezeau, l'enfant de 1922 et Brasse-Bouillon* pour dégager la chaîne référentielle qui les unit. Dans ce contexte, quand l'interprétation d'une de ces entités dépend de l'autre, nous tâcherons de nous pencher sur l'anaphore pronominale qui gère la transmission de l'information lexicale d'une maille à une autre. *Brasse-Brouillon* isolé ne contient aucune indication référentielle, il faudrait remonter dans le texte et chercher son premier antécédent qui est le personnage principal *Jean Rezeau*. Nous annotons, ci-dessous, manuellement les anaphores pronominales et nous constatons l'insuffisance de cette tâche sur le niveau lisibilité :

> [2] L'histoire débute durant l'été 1922. Un enfant découvre une vipère. Il la saisit et l'étouffe de ses mains. Cet acte lui vaut d'être comparé à Hercule, le personnage de la mythologie grecque, qui dans son berceau étrangla deux serpents. 25 ans plus tard, Jean Rezeau, "l'enfant de 1922" est le narrateur de l'histoire. Son surnom est Brasse-Bouillon. Il évoque la propriété de sa famille, La Belle Angerie. (Résumé Vipère au poing)

[44] Nous reviendrons sur ces notions au chapitre 3 de la deuxième partie.

Pour pallier aux limites de la résolution manuelle, l'intérêt que revêt l'annotation automatique des anaphores pour les technologies langagières a ces dix dernières années inéluctablement mené à l'élaboration de plusieurs travaux, qui ont particulièrement ciblé les documents et messages électroniques en langage écrit et « qui ont fait l'objet de campagnes d'évaluation internationales telles MUC [45] et SemEval [46] ou nationales comme DEFT [47] au cours de la dernière décennie.» (Muzerelle *et al*. 2013 : 3). Dans les sous-sections suivantes, nous examinerons successivement les ressources des corpus anaphoriques internationales et nationales.

2.1. Ressources internationales pour d'autres langues que le français

Nous nous sommes inspirée de plusieurs articles antérieurs cités dans la littérature, en particulier : Salmon-Alt (2002), Gardent/Manuélian (2005), Loáiciga (2013) et Désoyer, Landragin/Tellier (2015).

Sur le plan international, ces dernières années ont vu une grande effervescence au niveau des campagnes ayant contribué, entre autres, au développement de ressources linguistiques annotées en relations anaphoriques. « Ce domaine a donné lieu à de nombreux travaux, mais les données sur lesquelles ils se sont fondés étaient jusqu'à présent essentiellement de l'anglais écrit. » (Désoyer *et al.* 2015 : 1) Elles ont pour objectif d'évaluer « les systèmes de compréhension de textes (MUC-1 à MUC-7)[48], de détection automatique de thème (TDT)[49] et d'extraction automatique d'information (ACE)[50] » (Salmon-Alt 2002 : 164). Les corpus MUC-6 et MUC-7[51], qui sont issus des deux conférences MUC-6 et MUC-7 organisées entre les années 1987 et 1998, ont conduit à l'élaboration d'un schéma d'annotation pour la coréférence et à l'évaluation des systèmes de résolution des chaînes de coréférence. Ces corpus sont annotés en vue du traitement de la coréférence entre noms, groupes nominaux et pronoms (personnels et démonstratifs). Nous pouvons « citer ici l'exemple des données créées pour l'évaluation des résolveurs de coréférence dans MUC qui s'élèvent à environ 65 000 mots et comportent des listes de chaînes coréférentielles relatives à des sujets journalistiques. » (Salmont-Alt 2002 : 165).

---

[45] www.itl.nist.gov/iaui/894.02/related_projects/muc/
[46] http://semeval2.fbk.eu/semeval2.php
[47] http://deft.limsi.fr/index.php?id=1&lang=fr
[48] Message : Understanding Conferences : http://www.itl.nist.gov/iaui/894.02/related_projects/muc/)
[49] Topic Detection and Tracking : http://www.nist.gov/speech/tests/tdt/
[50] Automatic Content Extraction : http://www.itl.nist.gov/iad/894.01/tests/ace/
[51] http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/muc_7_proceedings/marsh_slides.pdf

Nous pouvons affirmer, avec Salmont-Alt (2002), que le genre journalistique, portant sur les mouvements de dirigeants, a mené indirectement à l'utilisation de 7 catégories réparties en 3 types adoptées lors du balisage des textes bruts :

– Enamex : désigne les noms de personne, d'organisation et de lieu. Les sous-types sont : Person, Organization et Location ;

– Timex : désigne les expressions temporelles. Les sous-types sont : Date et Time;

– Numex : désigne les expressions numériques, de monnaie et de pourcentage. Les sous-types sont : Money et Percent.

La figure 4 présente un exemple de texte annoté de MUC-6.

Mr. <ENAMEX TYPE="PERSON">Dooner</ENAMEX> met with <ENAMEX TYPE="PERSON"> Martin Puris</ENAMEX>, president and chief executive officer of <ENAMEX TYPE="ORGANIZATION">Ammirati Puris</ENAMEX>, about <ENAMEX TYPE="ORGANIZATION">McCann</ENAMEX>'s acquiring the agency with billings of <NUMEX TYPE="MONEY">400 million</NUMEX>, but nothing has materialized.

Figure 4: Exemple de corpus balisé MUC-6

Nous mentionnons, pour ce qui est de l'anglais, outre les MUC où les annotations se limitaient particulièrement à des phénomènes de coréférence, le corpus de Lancaster (Garside *et al.* 1997) dont l'accès est payant. Il compte 100 000 mots, annoté pour divers phénomènes anaphoriques (pronoms, anaphores nominales, ellipses, pronoms génériques). Il est inspiré de l'étude sur la cohésion de Halliday/Hasan (1976).

Le corpus de Poesio *et al.* (1998), extrait de la Penn Treebank[52] comportant l'annotation anaphorique (coréférence et anaphores associatives), est un corpus en anglais constitué d'articles de presse où sont annotées environ 1400 descriptions définies. Son annotation possède deux intérêts : mieux répertorier les différentes catégories de descriptions définies et cerner les difficultés soulevées par l'annotation et la résolution de celles-ci. Il faut reconnaître que le taux d'accord entre annotateurs est relativement bas pour les classifications assez serrées et ne permet ainsi qu'une classification binaire. Les conclusions relatives aux résultats obtenus sont donc assez insatisfaisantes. Il semble très épineux, pour ne pas dire impossible, d'annoter, pour les corpus Poesio *et al.*, les descriptions définies d'un corpus de façon suffisamment fiable au point de servir à

[52] http://www.cis.upenn.edu/~treebank/

l'évaluation et à l'entraînement de modules d'interprétation des expressions référentielles.[53]

Par ailleurs, le corpus GNOME[54] (Poesio 2004) a l'avantage de mettre en place un manuel d'annotation précis pour ces différents phénomènes ainsi que d'entraîner et évaluer des résolveurs d'anaphores (Poesio 2003, Poesio 2004) et des générateurs (Cheng *et al.* 2001, Cheng 2001 et Karamanis 2003)[55]. Il compte environ 18 000 groupes nominaux (pronoms, démonstratifs, possessifs et définis), annotés selon 14 dimensions syntaxiques et sémantiques.

Dans le corpus GNOME, l'information anaphorique est marquée au moyen d'un élément *(ante)* spécial. L'élément *(ante)* spécifie l'index de l'expression anaphorique (un élément *(ne)*) et le type de la relation sémantique (*identité* par exemple), alors que l'un ou plusieurs éléments *(anchor)* indiquent les antécédents possibles : si on a un seul élément *(anchor)*, l'anaphore est résolue. La présence de plus d'un élément *(anchor)* signifie un cas d'ambiguïté anaphorique.

```
<unit finite='finite-yes' id='u227'>
 <ne id='ne546' gf='subj'>  The drawing of
   <ne id='ne547' gf='np-compl'>the corner cupboard
   </ne></ne>
 <unit finite='no-finite' id='u228'>,or more probably
   <ne id='ne548' gf='no-gf'> an engraving of
      <ne id='ne549' gf='np-compl'>it </ne></ne>
 </unit>,
...
</unit>
<ante current="ne549" rel="ident"> <anchor ID="ne547">
</ante>
```

Figure 5 : Exemple annoté de GNOME (Poesio 2004 : 6)

[53] *cf.* Poesio/Vieira. (1998). A corpus-based investigation of definite description use. Computational Linguistics, 24(2):183–216, June : http://www.aclweb.org/anthology/J98-2001

[54] http://cswww.essex.ac.uk/Research/nle/corpora/GNOME/index.htm

[55] Texte source en Anglais de Poesio (2004) : "These results, and the annotated corpus, were used in the development of both symbolic and statistical natural language generation algorithms for sentence planning (Poesio 2000a, Henschel *et al.* 2000, Cheng et al. 2001), aggregation (Cheng 2001) and text planning (Karamanis 2003)." (Poesio 2004 : 1)

La figure 5 illustre la résolution automatique du pronom *it* dans l'exemple :

[3] The drawing of the corner cupboard, or more probably an engraving of it, …

Le pronom *it*, qui est identifié par l'élément *ne549*, a comme antécédent l'élément *ne547* qui correspond au groupe nominal *the corner cupboard*. Nous sommes face à une résolution automatique d'une anaphore pronominale sans ambiguïté. La présence d'un seul élément *(anchor)* en est la preuve[56].

Les corpus, que nous venons de citer, servent de ressources uniquement pour l'anglais. Il n'est pas facile de trouver des informations sur les corpus annotés en relations anaphoriques en d'autres langues. Nous pouvons citer un exemple de corpus pour l'espagnol-catalan : Ancora-Co (Recasens/Martí 2009), « il s'agit d'un corpus disponible tant pour l'espagnol que pour le catalan et dont chaque partie est composée d'articles journalistiques » (Loáiciga 2013 : 685) et quelques corpus multilingues de taille modeste : COREA (Heindrickx *et al.* 2008) en hollandais, NAIST Text (Idia *et al.* 2007) en japonais, et Salmon-Alt *et al.* (2002) en français-portugais. Nous classons, dans le tableau 2 ci-dessous, les principaux corpus annotés en anaphore disponibles au niveau international.

Corpus	**Phénomène annoté**	**Taille**	**Langue**
MUC-7 (1997)	Noms, GN, Pronoms Coréférence	Info indisponible	Anglais
Poesio et Vieira (1998)	Desc. définies Toutes relations	1400 DD	Anglais
ARCADE multilingue (Tutin *et al.* 2000)	Pronoms pers. et dem. Coréférence	80000 mots 800 pronoms	Français Anglais
Salmon-Alt et Vieira (2002)	Desc. Déf. et Dém Toutes relations	500 DD 300 Dem	Français Portugais
GNOME (Poesio 2004)	GN Coref. Et Assoc	18000 SN dont 554 DD	Anglais
TüBa-D/Z (Hinrichs *et al.* 2005)	Coréférence	800 000	Allemand
OntoNotes (Pradhan *et al.* 2007)	Coréférence	50 000	Anglais
NAIST Text (Idia *et al.* 2007)	Coréférence	970 000	Japonais

[56] Nous reviendrons sur ces aspects au chapitre 2 de la troisième partie.

COREA (Heindrickx *et al.* 2008)	Coréférence	325 000	Hollandais
PDT (Nedouluzzhko *et al.* 2009)	Coréférence	800 000	Tchèque
PCC (Ogrodniczuk *et al.* 2013)	Coréférence	514 000	Polonais

Tableau 3 : Principaux corpus mondiaux annotés en anaphore

2.2. Corpus annotés en anaphores en français

Afin d'assurer une analyse linguistique des relations anaphoriques dans un corpus, les annotations doivent présenter au moins un découpage structurel et un étiquetage morpho-syntaxique. La mise en place de ressources-clés nécessite, à vrai dire, de telles informations permettant l'évaluation du traitement automatique des anaphores. De surcroît, ce traitement repose essentiellement sur des connaissances structurelles (découpage du texte en titres, paragraphes et phrases), morphologiques (genre et nombre), syntaxiques (identification des groupes nominaux ainsi que leur portée interne, structures phrastiques, fonctions grammaticales) voire sémantiques (restrictions sélectionnelles, traits sémantiques). Nous examinerons, dans les sous-sections suivantes, les corpus anaphoriques en français.

2.2.1. Corpus écrits

Devant cette évolution des corpus électroniques annotés, aussi bien sur le plan quantitatif que qualitatif, nous partageons avec Lefeuvre et *al.* (2014 : 2691) qu' « il nous apparaît ainsi important de disposer de données d'observations suffisamment représentatives pour mener des études quantitatives de corpus utiles aussi bien à la linguistique qu'au TAL. Malheureusement, il n'existe pas en français de corpus d'envergure annoté en coréférence. » Cette insuffisance est confirmée par Désoyer *et al.* (2015 : 440) :

> Les systèmes de résolution automatique de la référence sont encore aujourd'hui extrêmement rares s'agissant du français, et même, à notre connaissance, inexistants – à l'exception de systèmes à base de règles tels que celui décrit dans (Trouilleux 2001) ou RefGen présenté dans (Longo 2013).

Cette rareté est aussi confirmée par Lefeuvre *et al.* (2014 : 4) :

> A notre connaissance, le seul corpus disponible en français est DEDE, centré sur l'étude des descriptions définies. Il ne comporte malheureusement que 48 kMots (Gardent/Manuelian 2005), ce qui limite sa représentativité et le rend inutilisable pour les besoins de l'apprentissage automatique. De même, le corpus

> du CRISTAL, de grande envergure, ne peut qu'être partiellement utilisé car il ne code que certaines formes particulières d'anaphore1*[Cf.]* (Tuttin *et al.* 2000).

ainsi que par Longo (2010 : 250) :

> Malgré l'existence de quelques corpus français annotés en relations de coréférence (Manuélian 2003, Salmon-Alt 2001, Tutin 2002), force est de constater qu'il est difficile d'utiliser ces ressources dans un système d'apprentissage automatique. En effet, ces corpus sont insuffisants en nombre, en taille (Salmon-Alt 2001) et ils se révèlent hétérogènes quant aux schémas d'annotation adoptés ainsi qu'aux choix des phénomènes annotés (pronoms personnels, descriptions définies, coréférence, anaphores associatives). Ainsi, la tâche#1 de la campagne SemEval 2010 Coreference Resolution in Multiple Languages propose-t-elle des données d'apprentissage pour plusieurs langues (anglais, espagnol, catalan, italien, allemand, néerlandais) mais aucune pour le français.

Malgré les confirmations de Landragin (2015), Lefeuvre *et al.*, (2014) et Longo (2010), nous avons pu trouver d'autres corpus non mentionnés par eux. Le tableau 3 résume les corpus trouvés. Nous avons remarqué qu'un très faible pourcentage de ressources françaises, pré-annotées partiellement en anaphores, était libre de droits et aucun schéma d'annotation n'était soumis à un quelconque partage. De plus, les phénomènes annotés ne présentent pas de points communs et seul un des corpus (Popescu-Belis, 1999) comprend une annotation s'étendant à plusieurs niveaux systématiques (structure, morphologie et syntaxe)[57].

Le corpus français ARCADE (Tutin *et al.* 2000), distribué par l'ELRA, compte près d'un million de mots et expressions anaphoriques annotées dont : les pronoms personnels, possessifs, démonstratifs, indéfinis et les adverbes anaphoriques. A l'exception des descriptions définies non annotées, le corpus présente tous les types de relations anaphoriques. Par ailleurs, plusieurs articles citent des corpus annotés mais non diffusés (exemple Salmon-Alt *et al.* 2002) contenant environ 500 descriptions définies et 300 descriptions démonstratives.

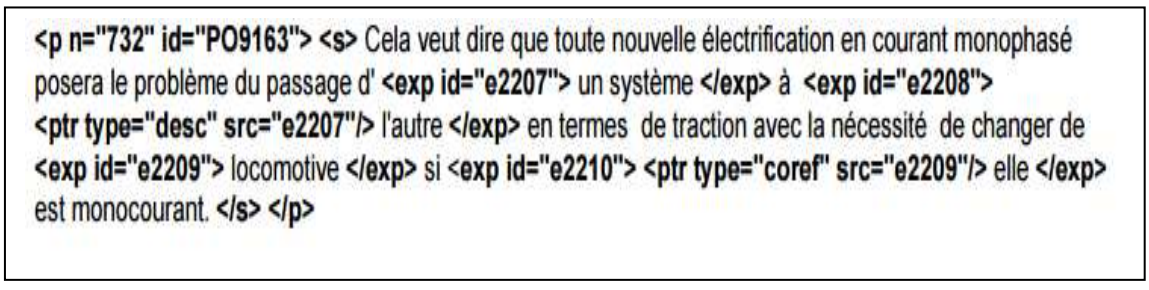


Figure 6 : Extrait d'ARCADE (Tutin *et al.* 2000)

[57] Voir http://andreipb.free.fr/textes/apb-these-1999.pdf pour plus d'informations.

```
<verbal_action id="xsd:20" who="l">
   <seg> ensuite prendre <de id="re:7" det="NP_indef"> une autre barre petite barre verticale </de> </seg>
   <comment type="speech_act"> requête </comment>
</verbal_action>
   <seg> et <de id="re:8" det="PR_pers"> la </de> mettre à peu près quatre centimètres à droite de
   <de id="re:9" det="NP_def"> la euh première </de> </seg>
</verbal_action>
<link_coref coref_type="classic" re="re:8" anchor="re:7"/>
<link_codom cd="ordinal" re="re:9" anchor="re:7"/>
```

Figure 7 : Extrait d'annotation de Salmon-Alt (2001)

Nous constatons, à partir des deux exemples ci-dessus, qu'il y a une différence entre les schémas d'annotation et les phénomènes annotés. Dans la figure 3, l'extrait de Tutin *et al.*, (2000) montre l'annotation de la coréférence. Quant à la figure 4, Salmon Alt (2001) a annoté la relation anaphorique. Les deux annotatrices ont utilisé des balises qui délimitent chaque élément. Elles sont formées de balises syntaxiquement valides en XML. En revanche, les deux annotations diffèrent sur le contenu des balises (élément <*ptr*> vs. élément <*link*>) et sur le type des relations anaphoriques (*desc*(ription) vs. *codom*(anialité)). Nous remarquons que, outre les relations anaphoriques, d'autres informations sont aussi annotées comme les paragraphes <*p*> et les phrases <*s*>. Nous partageons, avec les deux auteures, l'importance de l'annotation des relations anaphoriques pour une meilleure réutilisabilité des ressources disponibles. En revanche,

> parmi ces ressources, celles qui sont libres de droits pour la recherche ne bénéficient pas d'une infrastructure de diffusion ou de pérennisation particulière ce qui limite leur diffusion et leur réutilisation effective. » (Salmon Alt 2002 : 166)

Nous tenons à présenter ces études car il s'agit, à notre connaissance, des seuls travaux portant sur l'annotation automatique de l'anaphore en français. Certains aspects de ces études se retrouvent dans notre recherche. Ainsi, il s'agit bien de résoudre et d'annoter l'anaphore pronominale dans des textes en français. Nous partageons en partie les conclusions de Salmon Alt, concernant la limitation de diffusion des ressources citées dans le tableau ci-dessous. Néanmoins, notre travail diffère du sien sur plusieurs points[58].

[58] Nous reviendrons sur ces points, avec plus de détails, dans la troisième partie.

Corpus	**Auteurs**	**Nb de mots**	**Expressions annotées**	**Nb d'expressions**	**Liens annotés**	**Schéma d'annotation**	**Accès libre pour recherche**
Père Goriot, (Balzac)	Bruneseax *et al*. (1997)	30.000	GN pour personnages, lieux et objets principaux	3359	Coréférence, anaphores associatives	Compatible MATE	Oui
Vittoria Accoramboni (Stendhal)	Popescu-Belis (1999)	10.000	Tous les GN	638	Coréférence	MUC-compatible MATE	Oui
Le Monde Diplomatique	Clouzot *et al.* (2000)	95.000	Pronoms personnels 3ème personne	1316	Coréférence	TEI + format propriétaire	Négociable ?
ARCADE Différents genres	Tutin *et al.* 2000	1.000.000	Expressions anaphoriques sauf descriptions définies	?	Coréférence, anaphores associatives	format propriétaire	Non
ANANAS	Salmon-Alt (2001)	11.000	Tous les GN	1344	anaphores associatives en « autre »	TEI + compatible MATE	Oui

La Tribune (finances)	Trouilleux (2001)	45.000	Pronoms personnels 3ème pers. (sujet et objet), pronoms possessifs 3ième pers.	886	Coréférence, anaphores associatives	format propriétaire	Non
DEDE	Gardent C et Manuélian H. (2005)	48.360	Corpus de descriptions définies	4 910	Coréférence, anaphores	schéma Multext	Oui
EvalRefGen	Laurence Longo, Amalia Todirascu (2010)	Librement disponible sur demande à l'auteur	Corpus annoté en chaînes de référence (multigenre)		Coréférence, anaphores	Librement disponible sur demande à l'auteur todiras@unistra.fr	Oui

Tableau 4 : Corpus avec annotation anaphorique pour le français écrit

2.2.2. Corpus oraux

En plus des quelques corpus écrits en français annoté en relations référentielles, des corpus oraux sont apparus récemment, citons en exemple les principaux corpus CO2[59] et ANCOR[60], deux projets d'actualité qui ont réalisé des annotations sur des sous-corpus d'ESLO[61] (Enquête SocioLinguistique d'Orléans). Le corpus ESLO est en fait un projet du laboratoire LLL (Laboratoire Ligérien de Linguistique de l'université d'Orléans) et comporte deux sous-corpus : ESLO1 et ESLO2.

Avec l'objectif didactique précis qu'est l'intégration de l'enseignement du français, langue étrangère dans le système éducatif public anglais, un groupe d'universitaires anglais avait pour tâche de réunir des enregistrements sonores à Orléans entre 1968 et 1974, s'ensuivit alors la création d'ESLO1 (ELSO2, quant à lui, présente les enregistrements à partir 2008). Près de 200 interviews composaient cette étude, soit au total plus de 300 heures de parole variant entre des interviews directs et des enregistrements variés (conversations quotidiennes, discours officiels, entretiens médicaux, etc.). Le laboratoire LLL a poursuivi l'exploitation de ces matériaux en annotant en coréférences et en anaphores associatives trois fichiers extraits de ELSO1 et diffuse le corpus annoté CO2 sous licence Creative Commons CC-BY-NC-SA[62].

Corpus	CO2
Version	1.0 (juin 2013)
Type de dialogue	Dialogue oral peu interactif (interview socio-linguistique)
Locuteurs	Adultes hommes ou femmes francophones
Enregistrement	Voir distribution corpus ESLO
Contenu	Transcription orthographique + annotation en coréférence
Superviseurs	Jean-Yves Antoine (LI, Université de Tours), Emmanuel Schang (LLL, U. Orléans)
Annotateurs	Judith Muzerelle (LLL, U. Tours) et Aurore Pelletier (LLL, U. Tours)
Diffusion	libre sous réserve du respect de la licence Creative Commons CC-BY-NC-SA

Figure 8 : Présentation du corpus CO2

[59] http://www.info.univ-tours.fr/~antoine/parole_publique/CO2/Pres_CO2.pdf
[60] http://www.info.univ-tours.fr/~antoine/parole_publique/ANCOR_Centre/Pres_ANCOR_Centre.pdf
[61] http://eslo.huma-num.fr/
[62] http://eslo.huma-num.fr/index.php/pagelarecherche/pageprojetspartenaires/pageco2

L'objet du projet CO2 est l'étude des relations anaphoriques en s'intéressant spécifiquement à l'anaphore nominale. La méthode d'annotation utilisée consiste en :

- La délimitation des groupes nominaux et des pronoms dans le corpus.
- La description des groupes nominaux, entités nommées et pronoms compris.
- La détermination des relations anaphoriques :

> Les chaînes anaphoriques seront annotées par paires de relations entre une reprise anaphorique et la première mention de l'élément du discours concerné. Par exemple, si un nouvel élément du discours est introduit par l'élément A, puis est repris par l'élément B et l'élément C, on annotera deux relations anaphoriques : la relation (B ⟶ A) puis la relation (C ⟶ A). Les relations d'une même chaîne anaphorique pointeront donc toutes vers la première mention du référent concerné : celui-ci doit toujours porter la mention NEW = YES. (Muzerelle et *al.* 2013 : 13)

- La description de chaque relation anaphorique par quatre propriétés : type, genre, nombre et reprise par le locuteur. Nous revenons sur cette description dans notre troisième partie lors de l'approche que nous proposons.

Les fichiers d'ESLO n'ont pas seulement donné naissance au projet CO2 ; mais aussi au projet ANCOR (Anaphore dans les Corpus Oraux) qui est un corpus de langue orale et vise à s'intéresser aux procédés anaphoriques en oral spontané, en vue d'une résolution automatisée. Ce corpus a englobé quatre fichiers oraux d'ELSO, préalablement transcrits, pour faire preuve d'une certaine variété dans les situations de communication spontanée. La diffusion d'ANCOR se limite aux fichiers de transcription et d'annotation en coréférence, en d'autres termes, elle est réservée uniquement à l'annotation du corpus. Les personnes désirant obtenir en plus les fichiers audio relatifs à ces corpus, devront consulter les sites de distribution des corpus oraux originaux (ELSO). Ces corpus sont également accessibles librement sous licence Creative Commons[63]. L'objectif d'ANCOR était la description de toutes les relations référentielles existant dans le corpus, en réalisant une double annotation sous le logiciel GLOZZ.

> Le schéma d'annotation du corpus ANCOR cherche de manière classique à identifier pour chaque entité référentielle (ou mention) si elle introduit une nouvelle entité du discours, puis si elle réfère à une entité précédemment mentionnée (coréférence) ou si la référence a une entité précédemment mentionnée dans le texte est nécessaire pour son interprétation (anaphore associative). (Lefeuvre et *al.* 2014 : 3)

[63] http://www.info.univ-tours.fr/~antoine/parole_publique/ANCOR_Centre/index.html

Corpus	ANCOR_Centre (en abrégé : ANCOR)
Version	1.1 (novembre 2014)
Type d'oral	Parole spontanée : interview ou dialogue oral homme-homme finalisé
Taille	488 000 mots – 30,5 heures d'enregistrement
Locuteurs	Adultes hommes ou femmes
Enregistrement	Conditions réelles
Contenu	Transcription orthographique + annotation en anaphore et coréférence
Concepteur(s)	Laboratoires LI (Université François Rabelais de Tours) et LLL (CNRS) Jean-Yves Antoine (LI), Emmanuel Schang (LLL)
Annotation	Judith Muzerelle (LLL), Aurore Pelletier (LLL)
Révision	Judith Muzerelle (LLL), Anaïs Lefeuvre (LI)
Format intégré	Adèle Désoyer (LATTICE), Frédéric Landragin (LATTICE), Isabelle TELLIER (LATTICE)
Evaluation fiabilité	Jeanne Villaneau (IRISA), Iris Eshkol (LLL), Denis Maurel (LI), Judith Muzerelle (LLL), Anaïs Lefeuvre (LI), Jean-Yves Antoine (LI), Emmanuel Schang (LLL)
Diffusion	Licence Creative Commons CC-BY-NC-SA

Dans cette étude, notre intérêt portera sur l'interprétation automatique des anaphores pronominales et plus spécifiquement les cas ambigus. Cela dit, comme Landragin (2015) l'a confirmé, il n'y a pour le français aucun corpus annoté exhaustif, assez pertinent et de qualité suffisamment fiable qui assure ainsi l'entraînement et l'évaluation des résolveurs de descriptions définies. Devant la vaste étendue du terrain des anaphores, il a fallu penser à restreindre la relation anaphorique sur laquelle nous nous penchons dans cette recherche. Cette restriction s'est faite par une sélection d'un corpus adéquat réalisé selon des critères particuliers présentés dans la section suivante.

3. Critère de choix des textes de RESUMAN$_C$

Nous présentons, dans cette section de notre étude, une réflexion relative à l'application des principes de la linguistique sur corpus et à l'utilisation de ses méthodes. Nous commencerons par cerner notre corpus : des résumés en ligne d'ouvrage de la littérature française. Nous présenterons tout d'abord quelques définitions de l'objet d'étude, nous étudierons, par la suite, l'activité humaine de résumer qui peut donner lieu à des produits de natures différentes : clair/ambiguë, compréhensible/ non compréhensible, etc. Nous exposerons, après, les différents types de résumés. Le choix de ce genre textuel sera ensuite développé en partant du constat qu'il présente une importante densité en anaphore pronominale. En tenant compte de ce constat et les études sur l'anaphore ne répondant pas généralement aux exigences des méthodes d'analyse de corpus, nous

procèderons à l'analyse des difficultés spécifiques rencontrées. Il est à noter que les éléments de cette section ont trait au résumé produit par un agent humain.[64]

3.1. Définition du résumé

Plusieurs définitions du résumé sont présentes dans la littérature. Une d'entre elles définit un résumé comme une « présentation abrégée, orale ou écrite, qui rend compte de l'essentiel » (TLFi, 9ème édition). En 1979, l'ANSI (*American National Standards Institue*) le caractérise par « une présentation abrégée et exacte du contenu d'un document. Les résumés sont utiles pour l'accès aux publications et dans les bases de données informatisées »[65]. Chartrand (2011 : 1) définit, quant à elle, l'acte de résumer un texte comme :

> reformuler l'essentiel de son sens en un nombre plus réduit de mots, autrement dit, […] garder la pertinence communicationnelle d'un texte tout en réduisant sa quantité informative.

Pour ce faire, l'utilisation de l'anaphore est pertinente. Les chaînes référentielles garantissent la cohésion textuelle et maintiennent ainsi sa cohérence (*cf.* premier chapitre de la deuxième partie).

Ainsi d'après ces définitions, nous pouvons déterminer un texte résumé comme une condensation de l'information importante qui provient d'un ou de plusieurs documents sources, le but étant de fournir une version concise pour un ou plusieurs utilisateurs et une ou plusieurs tâches tout en contenant l'essentiel. L'activité de résumer est définie comme «une transformation réductive d'un TS[66] à un texte résumé réalisée par une réduction de contenu par sélection et/ou généralisation de ce qui est important dans la source » (Sparck 1998 : 2). L'auteur propose trois étapes principales pour aboutir à un résumé :

- Le texte-source doit être interprété pour donner une représentation de ce dernier ;
- Cette représentation est transformée en une représentation du résumé ;
- Le résumé est généré à partir de sa représentation.

Le type du texte source ainsi que les connaissances du producteur du résumé régissent l'activité de résumer : « Les résumés sont l'expression de la macrostructure d'un

---

[64] La structure de cette section est inspirée d'une Thèse pour obtenir le titre de Docteur en Informatique présentée par Mohamed Hédi Maâloul *Approche hybride pour le résumé automatique de textes. Application à la langue arabe*. Soutenue le 18 décembre 2012.

[65] [An abstract] is an abbreviated, accurate representation of the contents of a document. Such abstracts are useful in access publications and machine-readable databases.

[66] TS : texte source

texte tel qu'il est interprété par un individu à la lumière de ses connaissances» (Hutchins 1987 : 151). L'activité de résumer[67] recoure à des procédés de réduction des segments du texte retenus comme essentiels et désigne ainsi un procédé cognitif complexe qui implique non seulement une sélection, mais aussi une structuration des informations contenues dans le texte source. Nous détaillerons dans ce qui suit quelques caractéristiques du résumé en adoptant la définition de Masson (1998) :

- La concision : Elle est liée directement au rapport entre la taille du texte source et celle du résumé, autrement dit le pourcentage de réduction. Le caractère restrictif des critères mis en œuvre pour produire le résumé définit, généralement, ce pourcentage. Nous remarquons que la condensation anaphorique produit des résumés plus courts que les textes sources.
- La couverture : Elle correspond au rapport entre les thèmes (ou éléments) présents dans le texte d'origine et ceux contenus dans le résumé.
- La fidélité : Elle atteste de la qualité globale du résumé. En effet, elle réside en la relation d'homologie objective entre le résumé et le texte source. Les notions de fidélité et de couverture sont en rapport étroit : généralement, si la couverture est correcte, le résumé sera assez fidèle au texte d'origine ;

Enfin, les critères de cohésion et de cohérence seront abordés dans le deuxième chapitre de la deuxième partie. L'anaphore assure la cohérence et la cohésion du résumé, elle évite la répétition, son recours est donc indispensable. Regrouper les éléments retenus consiste à les enchaîner les uns aux autres, en d'autres termes, à leur donner une articulation d'ensemble.

3.2. L'activité de résumer

Produire un résumé nécessite des capacités à comprendre et à synthétiser. Il requiert donc une analyse pointilleuse du contenu textuel, ce qui n'est pas toujours une tâche aisée. Un résumé résulte d'une sélection précise d'informations du texte source selon une compréhension et des critères particuliers. Le processus que suit un humain afin de produire un résumé est présenté dans ce qui suit. Nous nous sommes appuyée sur des travaux dans le domaine de l'activité de résumer, qui ont analysé les processus mis en œuvre dans cette tâche : Kintsch/van Dijk (1978), Hidi/Anderson (1986), Giquel (1990), Fayol (1985 et 1997) et Mandin (2009).

[67] Nous analyserons cette activité dans la section suivante.

Afin d'élaborer un résumé par compréhension, d'après Giquel (1990) (voir figure 3), il faut passer par cinq étapes distinctes, qui sont : la lecture complète du texte source, l'analyse, la hiérarchisation et la synthèse de l'information source et, enfin, la rédaction du texte du résumé. Nous détaillerons chaque étape :

– Lecture complète du texte source

Afin de prendre connaissance du texte et de s'en imprégner, l'auteur du résumé doit procéder plusieurs fois à sa lecture complète et totale. Afin de maîtriser son contenu et donc son sens, il doit avoir une bonne compréhension de la langue du texte source. Suite à cette lecture, il portera en général son attention sur le début et la fin et moins sur les autres parties.

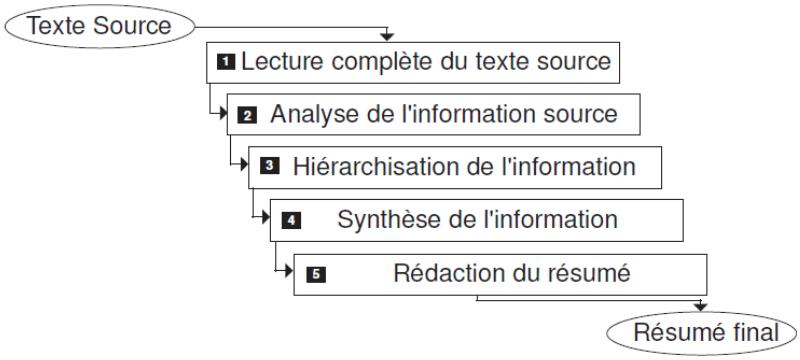


Figure 9 : Processus humain pour le résumé par compréhension Giquel (1990)

– Analyse de l'information source

Le lecteur analyse ensuite l'information (certains passages, mots-clés, connecteurs, anaphores, etc.) et établit uniquement un inventaire des informations figurant dans le texte d'origine. Il doit classer les éléments informatifs les plus importants qui constitueront le résumé. Pour cela, tout en restant neutre et objectif, il cherche à connaître le sens de chaque notion développée. Il peut également identifier les grandes parties du texte ou leur donner un titre.

– Hiérarchisation de l'information

Résumer un texte nécessite, au préalable, « de hiérarchiser l'information perçue par ordre d'importance, afin de décider du minimum d'informations à transposer dans le résumé » (Mandin 2009 : 5). Suite à un tri des informations présentées et pour s'assurer de la bonne compréhension du texte, le sujet classe les principales idées pour dégager ainsi les articulations référentielles et sémantiques assurant la cohérence des idées. Cette

compétence à hiérarchiser les informations d'un texte dépend de facteurs multiples comme les connaissances antérieures du texte à résumer, mais aussi des compétences à comprendre.

– Synthèse de l'information

Cette tâche nécessite que la personne, qui résume soit capable de sélectionner les informations indispensables au maintien des idées importantes du texte, « celles qui sont moins importantes, mais utiles pour comprendre le contenu du texte et qui peuvent être absentes du texte, mais inférées » (Van Dijk/Kintsch 1983). Le sujet réduit et regroupe toutes les informations sélectionnées visant la réduction du nombre de mots et de phrases : « Le passage du texte à la macrostructure nécessite l'application de règles de transformation textuelle (appelées macrorègles) dans le but de réduire l'information à l'essentiel » (Mandin 2009 : 7). C'est grâce aux opérations suivantes, distinguées par Van Dijk/Kintsch (1978) que nous pouvons synthétiser de l'information : les généralisations, les constructions et les suppressions.

– Rédaction du texte du résumé

Dans un style correct, neutre et clair, la dernière étape consiste en la rédaction d'un texte concis par rapport au texte initial. En restant objectif (se gardant de tout commentaires, illustration, etc), le « résumeur » reprend et synthétise les principales informations du texte : c'est la macrostructure qui est mise en texte (Fayol 1985, 1997 ; Van Dijk/Kintsch 1978). La question est de savoir si une compréhension correcte du texte source engendre de façon automatique un résumé clair ? Cette clarté influence la formation des chaînes référentielles tout au long du texte car d'autres éléments peuvent surgir et rendre la compréhension du lecteur ambiguë. Cette problématique sera examinée dans la deuxième partie de notre travail.

Par ailleurs, pour un meilleur impact sur le lecteur et afin de répondre à ses principales attentes, les résumés de texte ont été répartis selon leurs contenus et leurs objectifs. Ainsi, en fonction de l'utilisation qui lui est destinée mais surtout selon le besoin préalable de l'utilisateur, le résumé comporte différentes visées. Bodineau (1996), Fayol (1997), Minel (2002) et Hasler (2007) ont proposé une classification des principaux types de résumés selon leur fonction. Ceux-ci, loin d'être indépendants les uns des autres, renferment des propriétés communes comme, par exemple, la condensation considérable d'informations, qui renforce l'emploi dense de l'anaphore, comparé à un texte de même taille.

Nous distinguons trois types :

– Résumé informatif : le résumé s'applique surtout à englober toutes les informations pertinentes du texte original et c'est ce qui explique en quelque sorte que tous les principaux sujets doivent être rapportés. En fait, ce type de résumé offre une vue large du contenu d'un texte en fournissant un ensemble d'informations. Ainsi, pour assurer un juste aperçu du texte source, les sujets principaux rappelés dans le résumé sont répartis de manière fidèle à l'organisation initiale. Ainsi, ce « modèle propose de réorganiser et de contracter le texte-source seulement après un travail de sélection et d'élimination des informations triviales » (Fayol 1985). Cette phase d'élimination facilite au « résumeur » la rédaction. « La sélection et l'élimination peuvent elles mêmes être amorcées par une hiérarchisation préalable des informations » (Bodineau 1996).

– Résumé indicatif : ce type de résumé implique la notion de thématisation, c'est-à-dire qui se limite aux thèmes développés dans le document source indépendamment des commentaires. Il fournit suffisamment d'informations au lecteur qui décide s'il est nécessaire de retourner au document source ou pas. Comparé au résumé informatif, il ne contient que des éléments partiels mais pertinents afin de répondre à sa fonction, ce type de résumé s'apparente alors dans une certaine mesure à une table des matières. Ce type de résumé est intéressant pour la description des documents plutôt longs en fournissant un bref aperçu au lecteur de leur contenu, on le retrouve ainsi dans les fonds documentaires. Il demeure, néanmoins, utile pour les textes courts en fournissant un aperçu direct du contenu.

– Résumé scolaire ou fonctionnel : ce type de résumé a une visée essentiellement pédagogique, il permet d'évaluer les compétences cognitives d'un élève dans l'analyse et la compréhension d'un texte et dans la rédaction. Tout en respectant les grandes lignes et la structure générale du sujet traité, il doit être fidèle au texte original. Ainsi, et selon Paquay/Lauwaers (1992), l'élaboration de ce type de résumé passe par quatre sous-tâches : lecture du texte source, compréhension du vocabulaire, distinction des informations essentiels et enfin établir le plan du texte.

En examinant de près ces différents types de résumé, nous avons choisi le résumé informatif comme corpus pour notre travail. C'est un texte bref et cohérent, dense en occurrences anaphoriques. De plus, il contient tous les éléments éventuels pour une résolution d'anaphores ambiguës : données syntaxiques, contextuelles, pragmatiques et

logico-sémantiques. Ainsi, ce type de résumé obéit à nos critères de recherche sur l'anaphore.

3.3. Présentation de RESUMAN$_C$

La production de résumés ne peut être envisagée sans tenir compte des éléments qui assurent les relations existant entre le résumé et le lecteur (Masson 1998). En effet, tout résumé est doté d'une fonction particulière. Même s'il n'est pas aisé de trouver des critères qui définiraient *à priori* les utilisateurs, on peut par exemple les classer en fonction du ou des thèmes traités dans le texte source selon leur degré de spécialisation : s'ils sont spécialistes de ces thèmes ou non spécialistes.

Dans une telle distinction, on peut aisément comprendre que le niveau de compréhension du lecteur constitue un biais dans la résolution d'une anaphore ambiguë. En effet, le niveau des connaissances, de langage et le concept même d'un résumé constituent des facteurs essentiels pour la résolution de l'anaphore. Il est certain qu'un linguiste par exemple (un spécialiste), ayant connaissance des méthodes et techniques précises relatives au domaine, envisagera facilement une résolution correcte de l'anaphore, tandis qu'un non-spécialiste (élève, public ordinaire de la toile, etc.) trouvera des difficultés quant aux pistes d'analyse exactes à employer. La résolution de l'ambiguïté anaphorique étant essentielle pour la compréhension du texte, (pour des spécialistes et aussi pour toute autre personne moins informée), nous avons choisi de résoudre l'anaphore pronominale, une catégorie importante de l'anaphore dans le texte et le maintien de sa cohérence ou pour une compréhension correcte de son contenu, dans les résumés littéraires du site aLalettre.com. Les résumés à caractère de synthèse ont l'avantage d'être mieux assimilés par les spécialistes et les non-spécialistes. Notons cependant qu'ils pourraient dans certains cas ne pas satisfaire ceux qui privilégient le résumé au contenu informationnel dense.

Nous avons pris le temps de chercher des textes électroniques à la fois denses en anaphore pronominale et en cas d'ambiguïté anaphorique. Nous avons pris soin d'analyser de nombreux textes comme par exemple des manuels d'utilisation, des textes juridiques, ou encore des ordonnances médicales, mais nous nous sommes tantôt confrontée à l'absence d'ambigüité anaphorique, tantôt au vocabulaire spécifique du support textuel. La possibilité de trouver un texte électronique bref, cohérent et riche en anaphore s'est concrétisée après avoir examiné des résumés en ligne. Nous avons choisi, donc, les résumés du site aLaLettre.com, un site qui propose des résumés d'ouvrages de littérature

française. Nous avons obtenu l'autorisation de l'administratrice du site afin de les exploiter[68]. Il est l'un des sites de référence de la littérature française autant pour les auteurs classiques (évènements relatifs aux auteurs : conférences, exposition, vie des associations littéraires, films…) que contemporains (nouveautés, débats, interviews, forums, prix littéraires…). La gratuité de son accès favorise la diversité de ses visiteurs : élèves, étudiants, passionnés de littérature française ou encore internautes ordinaires. Tous représentent une cible potentielle pour notre étude sur l'anaphore dans les résumés de ce site.

L'hypothèse suivante constitue notre base de départ : s'agissant de textes narratifs, les anaphores pronominales seront les plus fréquentes suivies ensuite des syntagmes nominaux. Devraient être plus nombreux : les pronoms personnels de troisième personne, les pronoms et adjectifs possessifs de troisième personne et les relatifs, compte tenu de la dissociation par rapport au référent et du mode discursif autonome. Afin de confirmer le choix de ces textes comme corpus de notre thèse, et pour explorer les textes de $RESUMAN_c$, nous disposons d'un outil typique permettant l'analyse de corpus, à savoir le concordancier AntConc. Il s'agit d'un outil d'analyse statistique de texte qui cherche l'exploration d'un mot (ou d'une expression) et qui fournit la liste des occurrences. Il permet aussi d'obtenir l'environnement du mot (avec un contexte gauche et un contexte droite), qui éclaircit la sémantique du mot en observant aussi le mot qui « apparaît avec » (la co-occurrence). Comme AntConc accepte comme entrée des fichiers *.txt*, nous avons converti les fichiers *.doc* de notre corpus en des fichiers *.txt*. La capture d'écran suivante présente un extrait des concordances que nous avons obtenues avec le pronom *il* :

[68] **De :** Guy Jacquemelle <guy.jacquemelle@gmail.com>
Objet : Re: Contact aLaLettre.com
Bonjour
merci de votre message
Je vous donne mon accord bien volontiers. Merci de me tenir informé de l'avancée de votre recherche doctorale
bien cordialement
Guy jacquemelle

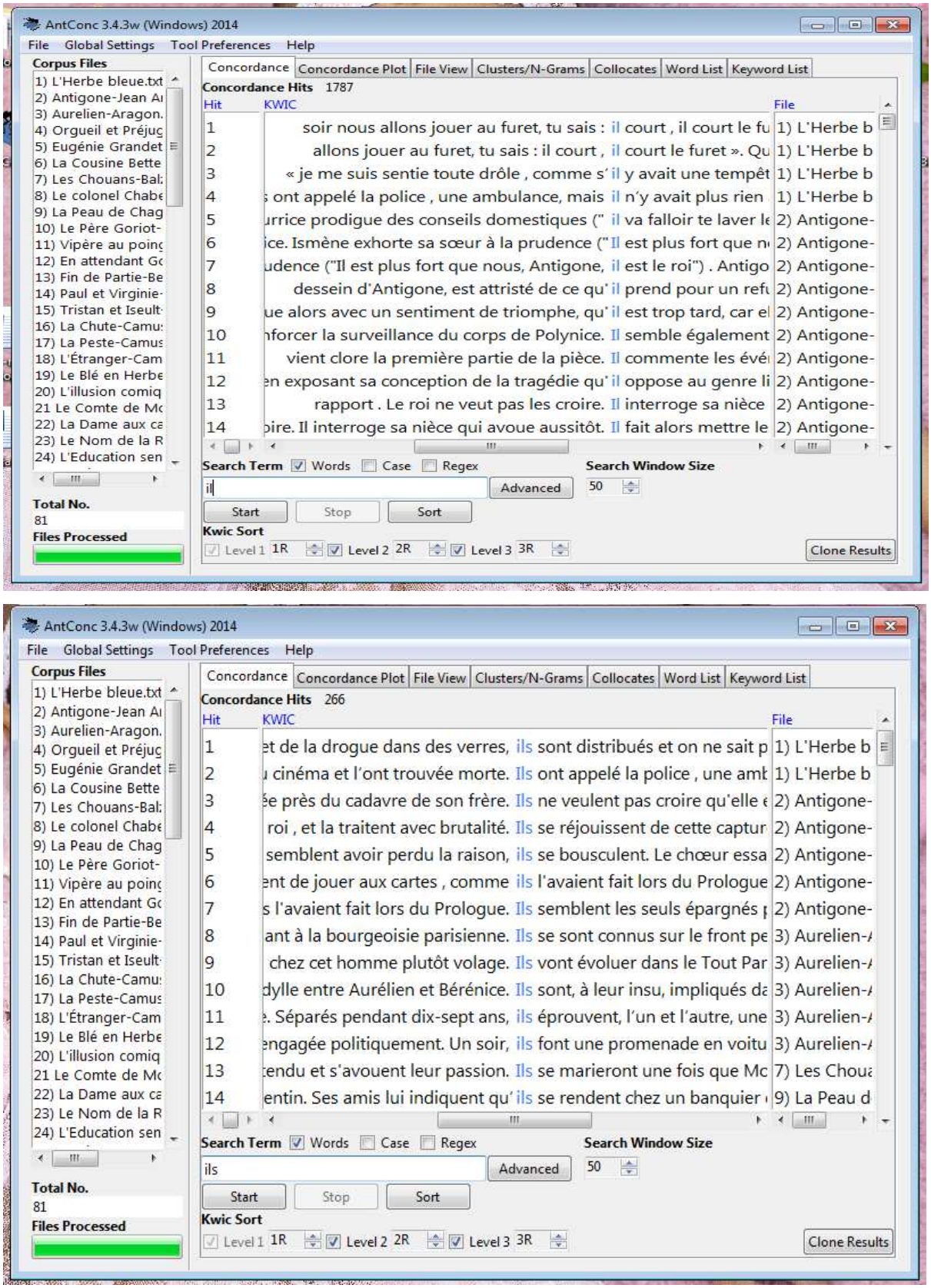
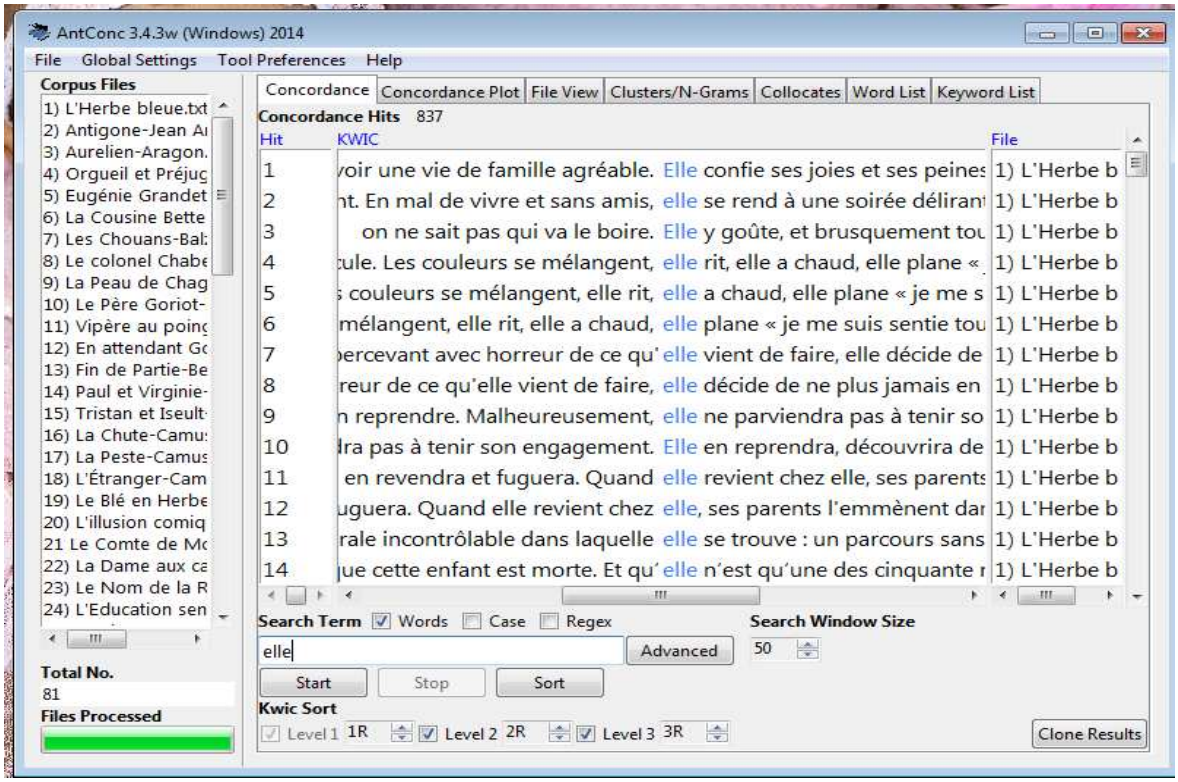


Tableau 5 : Extrait des concordances composées à partir des pronoms *il* et *ils*, obtenues à partir d'AntConc



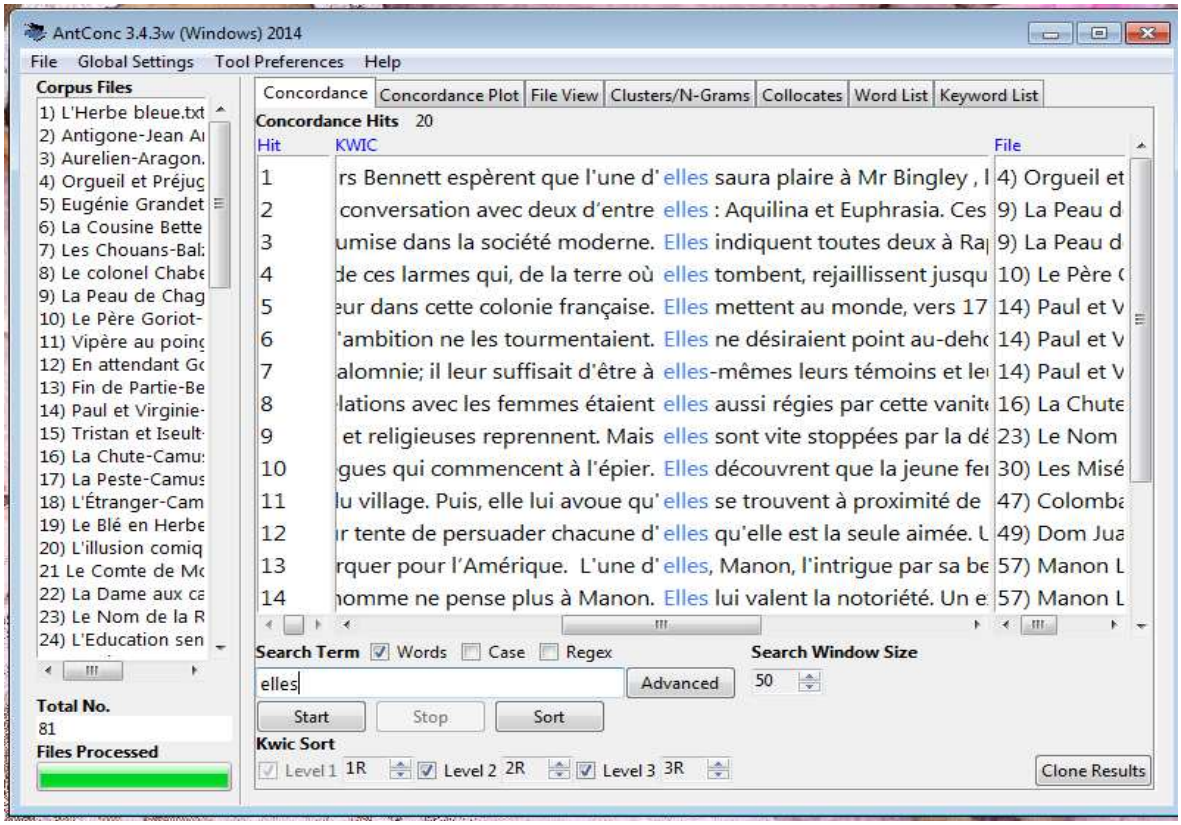
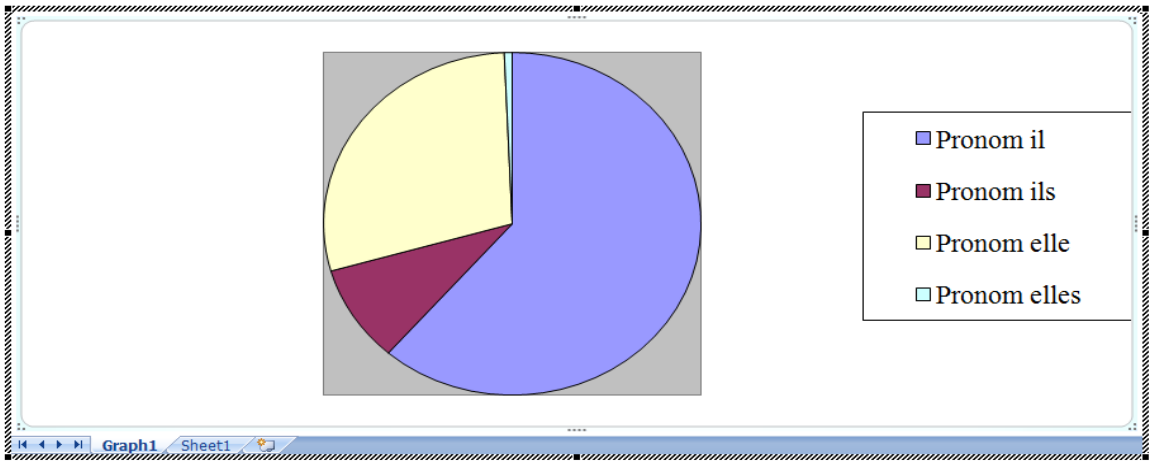


Tableau 6 : Extrait des concordances composées à partir des pronoms *elle* et *elles*, obtenues à partir d'AntConc

Après extraction, nous avons 1787 pronoms *il*, 266 pronoms *ils*, 837 pronoms *elle* et 20 pronoms *elles* sur 84 409 tokens. Ce premier repérage argumente en faveur de notre choix des résumés informatifs comme corpus de travail. Pour $RESUMAN_c$, la pluralité des personnages des résumés engendre le recours à l'anaphorisation.



4. Bilan

Tout d'abord, nous avons choisi une approche sur corpus, ce qui impliquait la constitution de RESUMAN, à partir de la base de données textuelle du site aLalettre.com pour le français. Nous avons, alors, été amenée à effectuer un travail de fouille centré sur les occurrences des pronoms dans notre corpus. Nous avons obtenu au total 2910 pronoms (*il, ils, elle* et *elles).* Aston (1997) note que « ces petits corpus existent entre 20 000 et 200 000 mots »[69]. De ce fait, le corpus RESUMAN, comptant 84 409 mots, est alors un petit corpus en termes de nombre de mots. En revanche, ce caractère limité du nombre de mots est expliqué par son utilisation et les objectifs assignés : il présente 2910 cas d'anaphore pronominale. Ce taux est bien plus élevé que dans les études antérieures effectuées dans ce domaine. La sélection des données est soumise à une orientation en fonction de l'objectif fixé en ce qui concerne le corpus. De notre côté, il ne s'agit pas d'un travail réalisé dans une optique représentative mais plutôt d'une étude d'une haute densité des phénomènes analysés menée dans une perspective informatique, partant d'une analyse sur corpus.

Le corpus RESUMAN vise à interroger, automatiquement, le fonctionnement de l'anaphore pronominale ambiguë dans les textes retenus en vue de mettre en évidence des caractéristiques syntaxico-sémantiques propres aux chaînes anaphoriques. Notre objectif

[69] Texte original: « that small corpora exist in the 20,000-200,000 word range ».

n’est pas de classer les antécédents potentiels (Kleiber, 1994), mais de chercher s’il y a des préférences entre eux, qui véhiculent des dimensions sémantiques spécifiques, et certains patrons syntaxiques dans des textes pouvant être qualifiés de brefs. Pour démontrer le degré d’apparition des anaphores pronominales dans les textes en français, nous avons utilisé le concordancier AntConc.

Deuxième partie
Anaphore : de la dimension textuelle à la dimension automatique

La quête du bon référent pour une anaphore et les mécanismes exploités pour l'atteindre n'ont pas cessé de faire couler beaucoup d'encre et créer un milieu effervescent de recherches et d'approches. Afin d'identifier le phénomène de continuité textuelle, il existe ainsi des fonctionnements caractéristiques qui aident à comprendre le sens du discours quelque soit la partie du texte. Un mécanisme essentiel à cette compréhension du sens est la résolution de l'anaphore. Car le processus de compréhension d'un texte se fait à travers l'analyse de la continuité référentielle, la cohérence et la cohésion, nous proposons dans notre étude une approche multimodale (linguistique et informatique) afin d'explorer les anaphores pronominales dans RESUMAN$_C$. Depuis la fin des années 70, les définitions opposent les notions de cohésion et de cohérence (Charolles 1988 : 53) :

> (…) tout le monde est à peu près d'accord pour opposer d'un côté la cohérence, qui a à voir avec l'interprétabilité des textes, et, de l'autre les marques de relation entre énoncés ou constituants d'énoncés. Concernant ces marques, depuis M.A.K. Halliday et R. Hasan (1976), on tend à les regrouper sous le nom générique de cohésion.

Dans la première, on retrouve des marques de relation entre constituants d'énoncés ou entre énoncés. La notion de cohérence, quant à elle, est liée à l'interprétabilité du texte dans un contexte de communication et demeure indépendante des marques linguistiques. Pour qu'une séquence soit perçue comme cohérente, la présence de marques de continuité n'est pas toujours nécessaire (Charolles 1988). Ainsi, dans le premier chapitre de cette partie, la clarification de la terminologie nous amène à distinguer les notions de cohésion et de cohérence et examiner leurs impacts sur les chaînes anaphoriques. Le deuxième chapitre est dédié à l'étude de la théorie de l'Accessibilité et les mécanismes intervenant lors de la résolution de l'anaphore. Le dernier chapitre expose la dimension automatique de notre approche en détaillant les différentes approches pour aboutir à la compréhension du texte. Pour pouvoir identifier correctement les anaphores contenues dans un document textuel écrit en langage naturel, il importe de concevoir et de développer des méthodes et des outils capables de saisir la structure de ces documents. En d'autres termes, ces méthodes et ces outils doivent être en mesure de 'comprendre' les documents traités.

Chapitre1 : Dimension textuelle de l'anaphore

Contrairement à la *grammaire de texte*[70], la dimension textuelle témoigne de la cohérence et de la cohésion des textes et va au-delà de la frontière phrastique. Dans ce sens, ce n'est pas une théorie de la phrase étendue au texte, mais plutôt une théorie de la production co(n)textuelle de sens. Prandi (2007) propose de passer d'une dimension régie par la grammaire, la dimension phrastique, à une autre régie par la cohérence des concepts, la dimension textuelle. Des moyens cohésifs appropriés doivent supporter cette cohérence. Ainsi, l'auteur, dans sa conception, dépasse la dimension *phraséocentrique.* Par ailleurs, cette théorie doit être édifiée à partir de l'analyse de textes concrets (Adam 2008 : 1483). Dans l'étude linguistique des dispositifs anaphoriques, une place centrale est accordée aux deux notions. Adam (2005 : 95) affirme que :

> Les liens anaphoriques jouent un rôle capital non seulement dans la cohésion, mais dans la progression par modifications progressives d'un référent qu'ils ne se contentent généralement pas de simplement reprendre.

Les relations anaphoriques dépassent les limites de la proposition et de la phrase puisque antécédent et anaphorique peuvent appartenir à des phrases différentes. Ainsi, nous parlons de dimension textuelle de l'anaphore (Apothéloz 1995) et de son intérêt : en effet, les relations anaphoriques contribuent à l'élaboration de la textualité. Ils marquent la continuité référentielle et participent ainsi à la cohérence du discours. Selon Charolles/Ehrlich (1991 : 261), les anaphores sont « conçues pour faciliter l'interprétation et la construction de la cohérence » en tant que marqueurs de cohésion.

En analysant les paramètres qui les définissent en fonctions de la littérature linguistique, nous discuterons, dans ce chapitre, les notions de cohésion et de cohérence. L'objectif est de caractériser l'impact de l'anaphore sur la cohérence : en effet, cette analyse facilitera l'étude de sa dimension textuelle.

1. Cohésion et cohérence

L'existence d'une compétence textuelle a été proposée par Charolles en 1978. Celle-

[70] Les grammaires de texte ont souligné rapidement la problématique de la cohérence. Par ailleurs, certains auteurs ont indéniablement contribué à l'installation et à la dissémination de ces grammaires (exemple : Petöfi (1973), Van Dijk (1972, 1977)).

ci sous-tendrait l'appréciation de la cohérence des textes et l'élaboration de textes *bien formés*. Par la suite, l'auteur propose que l'approche, étudiant l'occurrence régulière en discours d'une forme particulière remplissant une fonction précise, est obligatoirement fonctionnelle et contextuelle. Il s'attache, ensuite, plutôt à la dimension pragmatique (1983 ; 1989; 1995 ; 2002 ; 2006, 2011). Dans l'interprétation, en donnant de l'importance aux apports externes au texte, les choix de formulation seraient réduits par la pragmatique, en conséquence, le recours à la pragmatique toute seule risque de conduire à un paradoxe. Ainsi, certains auteurs précurseurs comme Charolles (1978), Reinhart (1980), Cornish (1986), Mann/Thompson (1988) et Cornish (1996) ont posé un principe premier de cohérence : c'est ce qui oriente le lecteur vers une lecture cohésive des marques dans les textes. Ces auteurs s'opposent à l'idée que la cohésion soit à la base de la cohérence, entre autres sous l'influence directe ou indirecte de Grice (1975). Des travaux sur la cohésion et la cohérence s'enchainèrent après comme Charolles (2005, 2006, 2011), Fries 2004, Gille (2007), Rondelli (2008), Alkhatib (2012), etc.

Toutefois, le risque de détourner les recherches de l'examen précis des choix de réalisation linguistique est présent dans cette perspective pragmatique, même si elle est par ailleurs essentielle à la compréhension des fonctionnements discursifs. La littérature que nous présentons se concentre ici sur le détail de la surface des textes pour le revaloriser, en prenant en compte les acquis de la pragmatique linguistique. L'objectif de ces travaux est d'expliquer comment des facteurs contextuels contraignent des choix d'agencement de constituants non déterminés par la syntaxe et comment la construction de l'interprétation est différemment orientée par ces choix. La tendance à minimiser les choix de réalisation linguistique était déjà rejetée en 1981 par des auteurs comme De Beaugrande et Dressler :

> Nous devons nous garder de laisser le texte disparaître derrière les processus mentaux. Les débats récents sur le rôle du lecteur soulignent les dangers de supposer que les récepteurs des textes peuvent faire ce qu'ils veulent avec une présentation. Si cette notion était exacte, la communication textuelle serait peu fiable, peut-être même solipsiste. Il doit y avoir des contrôles définitifs, mais non absolus, sur les variations entre les modes d'utilisation d'un texte par différents récepteurs[71]. (de Beaugrande et Dressler 1981: 35 ; notre traduction)

Une série de concepts qui lui sont propres se retrouvent dans la linguistique

[71] Texte original : "We must guard against allowing the text to vanish away behind mental processes. Recent debates over the role of the reader point up to the dangers of assuming that text receivers can do whatever they like with a presentation. If that notion was accurate, textual communication would be quite unreliable, perhaps even solipsistic. There must be definitive, though not absolute, controls on the variations among modes of utilising a text by different receivers."

textuelle[72], qui se différencie catégoriquement de la *grammaire de texte*. Ainsi les conjonctions de coordination (*"mais", "ou", "et", "donc", "or", "ni", "car"*) sont opposées à la classe textuelle des connecteurs. Par ailleurs, dès que l'on est au niveau du texte, la classe morphologique des pronoms personnels n'est plus homogène. C'est alors dans le domaine des reprises, avec les démonstratifs, certains indéfinis et certains groupes nominaux définis, que les pronoms de troisième personne (*il(s)* et *elle(s))* doivent être (re)classés. Les pronoms des deux premières personnes, quant à eux, doivent être mis en relation avec les possessifs et les modalisateurs, la classe des déictiques et l'ensemble du domaine énonciatif. Quel que soit le texte, toutes les phrases qui le composent contiennent, d'une part, des éléments qui assurent la cohésion de l'ensemble _ ce sont des éléments récurrents de référence présumés connus (par le co(n)texte) _ d'autre part, des éléments inconnus (nouveaux) porteurs de l'expansion et de la dynamique de la progression informative.

Nous consacrons donc la section suivante à l'étude de la cohésion.

1.1. Cohésion

Halliday et Hassan (1976) proposent la distinction entre la texture (propriétés constitutives des textes) et la structure (propriété formelle des phrases) afin de définir les propriétés qui rendent un texte différent d'une simple suite de phrases. La texture, selon eux, est déterminée par deux types de systèmes de relations : un premier système qui repose sur les relations assurant l'ancrage du texte dans le contexte situationnel et un deuxième, dans lequel apparaissent les relations de cohésion et qui concerne les relations internes au texte entre ses différents éléments constitutifs :

> Un texte […] n'est pas un simple enchaînement de phrases [*string of sentences*]. En d'autres termes, il ne s'agit pas d'une grande unité grammaticale, de quelque chose de même nature qu'une phrase mais qui en différerait par la taille – une sorte de superphrase. Un texte ne doit pas du tout être vu comme une unité grammaticale, mais comme une unité d'une autre espèce : une unité sémantique. Son unité est une unité de sens en contexte, une texture qui exprime le fait que, formant un tout [*as a whole*], il est lié à l'environnement dans lequel il se trouve placé. (1976 : 293 ; traduction d'Adam (2008 : 1483))

Halliday et Hassan ont proposé une définition de la cohésion dont le fonctionnement repose sur un système de relations sémantiques. Néanmoins, ces relations s'établissent entre des entités discursives où les structures phrastiques n'interviennent pas : ainsi, les

[72] Pour des informations plus approfondies sur la linguistique textuelle, cf. Adam (2008).

procédures qui assurent la cohésion ne lieraient que certains éléments des phrases. Dans cette approche très restrictive de la cohésion, l'anaphore comme exemple de dispositif cohésif relie souvent des unités lexicales et cette liaison entraîne une relation entre les propositions contenant les unités liées. De plus, on note que la référence de toute une proposition peut être maintenue par certains pronoms anaphoriques. Dans l'exemple :

[1] Adraste est éconduit par Isabelle mais **cela** ne le décourage guère. Il s'en va demander sa main à Géronte, son père. (Résumé L'illusion comique)

Le pronom cela renvoie à l'ensemble de la proposition précédente et non à une unité lexicale particulière. Halliday et Hassan (1976) distinguent par la suite deux types de dispositifs cohésifs : la cohésion lexicale et la cohésion grammaticale.

La cohésion lexicale repose sur la mise en relation de lexèmes à partir d'une proximité sémantique induite par le discours lui-même et non pas sur des unités lexicales spécialisées dans l'établissement de la cohésion. On peut citer parmi ces liens cohésifs lexicaux :

1. *Les anaphores lexicales*. Ce sont les relations établies, sur la base d'un rapport de synonymie ou d'hyperonymie / hyponymie, à partir de la proximité de deux lexèmes dans l'organisation lexicale, par exemple la relation qui s'établit entre *tigre* et *bête* dans l'exemple :

[2] Le maire, mort de peur, demande à Antonio de finir de traquer le tigre tout seul et de le tuer. Antonio accepte. Il traque **la bête** pendant plusieurs heures. (Résumé Le Vieux qui lisait des romans d'amour)

2. *La procédure de nominalisation.* Même si c'est toute une phrase, ou une partie de phrase qui est reprise sous la forme d'un substantif, cette procédure est généralement présentée comme un cas d'anaphore lexicale. Dans l'exemple :

[3] Déçu, Zadig décide de se consacrer à **l'étude des sciences**. Mais **cette nouvelle activité** ne va lui causer que des ennuis. Ses connaissances et la pertinence de son analyse lui valent de se retrouver en prison. Libéré, il essaye de se consoler en s'adonnant à la philosophie. Il reçoit chez lui les savants de Babylone. Mais ce succès suscite la jalousie du courtisan Arimaze, qui le fait incarcérer. (Résumé Zadig)

ce succès renvoie à la proposition précédente *«Libéré, il essaye de se consoler en s'adonnant à la philosophie. Il reçoit chez lui les savants de Babylone.»* et maintient la cohésion dans ce passage en évitant la répétition.

3. *Les relations lexico-sémantiques paradigmatiques.* Ce type de relations est souvent qualifié *d'anaphore associative* : en effet, la cohésion repose sur des relations parties/tout ou de contraste comme dans l'exemple proposé par Kleiber/Vassiliadou (2007 : 155) :

[4] Nous entrâmes dans un village. **L'église** était située sur une butte.

On retrouve une anaphore associative fondée sur une relation parties/tout entre *église* et *village*.

Puisque le concept de cohésion repose sur des relations sémantiques internes au texte, il est avant tout un concept sémantique et se rapporterait aux phénomènes de continuité thématique. La continuité et l'agencement des thèmes d'un discours sont assurés par des opérations de cohésion. La cohésion est caractérisée par le fait que la progression du discours doit nécessairement être accompagnée par le phénomène de reprise répétition sémantique (Charolles 2011). Ainsi, deux énoncés seraient cohésifs lorsqu'un constituant du premier énoncé est repris dans le second, c'est-à-dire lorsque cette continuité thématique est matérialisée dans le discours. Un ensemble de moyens lexicaux identifiables, *la substitution* ou *l'ellipse* par exemple, assurent la *cohésion grammaticale.* Cela implique l'utilisation d'unités lexicales de classes fermées (pronom personnel, pronom indéfini ou encore adjectif possessif) pour la substitution. Ces unités lexicales posséderaient des références virtuelles minimales qui ne seraient pas suffisamment déterminées pour acquérir une référence actuelle propre et seraient ainsi caractérisées par un déficit d'autonomie référentielle[73]. Afin d'acquérir une référence actuelle, ces lexèmes doivent être mis en rapport avec une autre unité référentielle présente dans le discours. C'est la même chose pour l'ellipse (ou anaphore zéro) qui est définie par Milner (1982) comme un phénomène de saturation référentielle. L'établissement d'une relation de saturation référentielle entre deux unités lexicales qui partagent la même référence actuelle définit donc la cohésion grammaticale, comme nous l'avons montré dans l'exemple 2.

Pour certains auteurs (Reichler-Béguelin 1988), le faible contenu référentiel des unités lexicales impliquées dans la cohésion grammaticale correspondrait en fait à une hyperonynie maximale. Par exemple, le pronom personnel *il* rentrerait dans le cadre de l'anaphore lexicale, car il serait un hyperonyme pour tous les noms masculins singuliers.

[73] *Cf.* Chapitre 1 de la première partie.

Ainsi, ces auteurs considèrent la cohésion grammaticale comme un cas particulier de cohésion lexicale. Ce point de vue unifie la cohésion en éliminant la frontière entre cohésion grammaticale et lexicale. Ainsi la cohésion « joue sur des relations d'identité, d'inclusion ou d'association entre constituants d'énoncés » (Charolles 1988a : 53). Même si leur présence n'est pas obligatoire, le rôle des dispositifs de cohésion est de premier plan dans l'établissement de la cohérence. En effet, en indiquant explicitement les énoncés devant être reliés et la forme que doit prendre cette liaison, les dispositifs de cohésion participent à l'établissement de la représentation du discours. Ces dispositifs définissent un co-texte nécessaire au déroulement du discours et organisent le contenu.

Selon Charolles (1988a : 57), afin de « manipuler l'activité inférentielle du destinataire », le locuteur dépose des marques de cohésion comme par exemple le pronom. Celui-ci, afin de guider l'interprétant, indique d'une part qu'il réfère et qu'il faut donc rechercher sa référence ; d'autre part, il signale que l'entité à laquelle il réfère est aisément accessible. En ayant recours à des procédures mises en œuvre par les dispositifs de cohésion et en évitant les procédures inférentielles cognitivement plus coûteuses, ces dispositifs anaphoriques permettraient de guider la construction de la représentation du discours. Kintsch et van Dijk (1978) ont confirmé l'hypothèse selon laquelle la vérification de la cohésion est la première étape de la formation d'une base de texte cohérente. Ces auteurs affirment que la base de texte peut être utilisée pour des traitements ultérieurs s'il existe des arguments communs parmi les propositions. Dans le but d'améliorer la cohérence de la base de texte, des procédures inférentielles sont employées pour combler les lacunes trouvées.

Afin de dépasser les relations superficielles entre les énoncés, la recherche de la cohérence implique l'utilisation de procédures inférentielles chez l'interprétant. De la même façon, « un travail de résolution » (Charolles 1988a : 59), qui suppose la production d'inférences, est souvent nécessaire pour les marques de relations impliquées dans la cohésion. En ce sens, il n'y a pas autant de différences[74] entre cohérence et cohésion que l'analyse linguistique précédemment présentée le laissait supposer. Pour la cohésion, des indications explicites sont fournies par le locuteur sur la mise en relation des énoncés par

[74] Salles (2006) a analysé les notions de cohérence et cohésion sous une optique conditionnelle. Elle a essayé de réponde à ces deux questions : « la 1ère, extrêmement classique, pour ne pas dire rebattue, subordonne la cohérence à la cohésion et se demande ainsi dans quelle mesure la cohésion est un facteur de cohérence ; la 2nde, sans doute moins classique, même si elle est loin d'être inédite, va, au contraire, subordonner la cohésion à la cohérence et se demander alors dans quelle mesure la cohérence est un facteur de cohésion. »

l'interprétant, alors que l'établissement de la cohérence implique, en plus, la constitution de relations non explicitement marquées dans le texte.

1.2. Cohérence

> « Il n'est pas sûr que l'on puisse définir précisément en quoi consiste la cohérence…» (Charolles 2006 : 26)

Au début des années soixante, l'émergence des grammaires de texte, avec le constat qu'une suite de phrases ne constitue pas forcément un texte, a contribué à l'apparition de la notion de cohérence en linguistique. Entre les années 50 et 70, de nombreux auteurs ont tenté de définir les conditions nécessaires à l'obtention d'un texte acceptable à partir d'une suite de phrases (Charolles 1978). En effet, les grammaires qui dominaient la linguistique à cette époque étaient les grammaires génératives et transformationnelles et celles-ci ne rendaient pas compte de l'aspect spécifiquement textuel du discours. Un consensus existe autour de quatre conditions nécessaires pour qu'une suite de phrases forme un texte cohérent, même si l'aspect interprétatif de la cohérence est mis en avant. Charolles (1978) définit ces *métarègles* comme suit :

1. *La condition de répétition* (Charolles 1978 ; Reinhart 1980) : « Pour qu'un texte soit *(microstructurellement ou macrostructurellement)* cohérent, il faut qu'il comporte dans son développement linéaire des éléments à récurrence stricte. » (Charolles 1978 : 14) Le degré de cohérence atteint par le locuteur définit en grande partie la qualité d'un texte. Selon les données de la littérature, la cohérence comporterait un aspect externe concernant le contexte de communication : la cohérence appelée *contextuelle* (Lundquist 1980) (adéquation du contenu du texte avec la situation de communication) et deux aspects internes concernant la structure du texte : la cohérence *macrostructurelle* aussi appelée *globale* (Charolles 1978 ; Adam 1992) (découpage d'un texte formant un ensemble uni, en différentes parties en relation les unes avec les autres) et la cohérence *microstructurelle* aussi appelée *locale* (apporter de nouvelles informations sans digression et grâce à l'enchainement des phrases d'un texte en progression (Charolles 1978), emploi de connecteurs pour préciser les rapports de relation logique entre les idées et emploi d'anaphore pour l'établissement d'une continuité (Charolles 1978). Selon cette règle, certains éléments référentiels doivent être répétés dans le discours; néanmoins, ne sont précisés ni le taux de répétition nécessaire, ni si

les répétitions peuvent être seulement inférables du contexte sémantique ou alors explicitement marquées dans le discours. En effet, 1'exemple :

[5] Il fait beau, les oiseaux chantent[75].

alors qu'il ne comporte aucune répétition explicite et que le texte est minimal, peut être qualifié de cohérent. Ici, la condition de répétition est satisfaite par un effet de redondance de la seconde proposition entrainé par les connaissances pragmatiques sur le fait qu'habituellement quand il fait beau, les oiseaux chantent.

2. *La condition de progression :* Tout en permettant l'application de la première règle, cette condition indique que « Pour qu'un texte soit microstructurellement ou macrostructurellement cohérent, il faut que son développement s'accompagne d'un apport sémantique constamment renouvelé» (Charolles 1978 : 20). Ainsi la première règle ne se suffit pas à elle-même ; quand un texte contient un référent qui est maintenu, sans cesse répété, dans des phrases de même sens, il ne remplit pas les conditions d'un texte cohérent (6).

[6] **Candide** se rend à Constantinople. **Le jeune héros** visite l'ancienne ville Byzance. **Il** voyage à la ville romaine. (Résumé Candide)

Dans cet exemple, le référent *Candide* est introduit à trois reprises en gardant les informations le concernant. Produire un texte cohérent exige alors un équilibre entre « une continuité thématique et (une) progression sémantique (…) une telle performance exige donc que soient conjointement maitrisées les méta-règles de répétition et de progression » (Charolles 1978 : 21).

3. *La condition de non-contradiction :* « Pour qu'un texte soit microstructurellement ou macrostructurellement cohérent, il faut que son développement n'introduise aucun élément sémantique contredisant un contenu posé ou présupposé par une occurrence antérieure ou déductible de celle-ci par inférence. » (p. 22) Cette règle stipule, tout d'abord, que des rapports logiques doivent pouvoir être établis entre les faits dénotés par le texte. Cette règle rejette ensuite les contradictions qui induisent un contraste sémantique trop important entre ces différentes propositions. Elle précise qu'il ne doit pas exister de

[75] Exemple emprunté de Charolles (1988a : 48).

contradictions dans le contenu des séquences propositionnelles d'un même discours. Par exemple, il n'y a pas de cohérence dans une phrase qui ne respecte pas la règle de non-contradiction comme :

[7] « Un célibataire est marié ».

alors que cette cohérence existe dans la phrase :

[8] « Ton célibataire, eh bien il est marié[76] ».

En effet, à partir des informations de cette phrase, on peut élaborer une représentation unitaire qui peut être paraphrasée par :

[9] L'homme que tu m'as décrit comme étant célibataire est, en fait, un homme marié.

4. *La condition de congruence,* ou *règle d'isotopie* ou « Méta-règle de relation : Pour qu'une séquence ou un texte soient cohérents, il faut que les faits qu'ils dénotent dans le monde représenté soient directement reliés. » (p. 32) Selon cette règle, un champ sémantique unique pour l'ensemble des informations du texte doit pouvoir être cerné par l'interlocuteur. Dans un monde représenté, une séquence désigne des états, événements ou actions qui sont obligatoirement concordants (ou congruents). De ce fait pour l'interlocuteur, dans un discours ; il faut une transparence des relations entre les éléments d'information. Savoir si les informations véhiculées par le texte peuvent être rattachées à un univers de référence est donc bien déterminé par l'interlocuteur[77].

En ce sens et afin d'établir les implications des différentes formes de relations interpropositionnelles possibles sur la perception de la cohérence, Kinstch et van Dijk (1983) ont proposé un modèle[78] décrivant les traitements cognitifs essentiels aux processus de compréhension de texte :

1. *L'absence de relation :* Le discours est perçu comme incohérent car il n'existe aucune forme de relation entre les propositions.

[76] Exemples d'après Reichler-Béguelin (1988).
[77] Pour déterminer l'acceptabilité _ il existe sept facteurs essentiels à la prise en considération d'un texte : la cohésion, la cohérence, l'intentionnalité, l'acceptabilité, l'informativité, la situationnalité et l'intertextualité (De Beaugrande et Dressler 1981) _de cet univers de référence, des informations extradiscursives (connaissances culturelles et encyclopédiques, évidences perceptives et situationnelles...) entrent en jeu.
[78] Nous parlerons, avec plus de détails, des modèles de compréhension des textes dans le chapitre 3 de cette partie.

2. *La cohérence indirecte* : La cohérence, qualifiée de macrostructurelle, s'établit car les faits énoncés font partie d'« un monde possible ». Néanmoins, ces faits ne sont pas contigus dans le discours et doivent être mis en relation à un niveau plus élevé de représentation. Ainsi, la cohérence ne s'établit pas à partir d'informations présentes dans des propositions contiguës de la forme de surface du discours.[79]
3. *La cohérence directe* : De la même façon, la cohérence s'établit car les faits énoncés font partie d'« un monde possible[80] », à la différence que « les propositions sont reliées par un lien logique qui peut être marqué par un connecteur logique ». La cohérence est qualifiée de microstructurelle, car la relation entre les informations contenues dans le discours peut être établie à partir de la forme de surface. Ce type de cohérence se caractérise principalement par la présence d'un lien logique entre les propositions ; ce lien peut être explicitement marqué ou inférable. Par exemple, dans la phrase :

   [10] Il est jeune, riche et beau. (Résumé Zadig)

   les propositions sont liées par le contexte de description d'une même personne représentée par le pronom *il*.
4. *La connexion coordonnée* : C'est une forme dérivée de la cohérence directe à la différence que les propositions « sont ordonnées et sont l'expression d'un fait complexe ». Le lien logique pouvant être établi entre deux propositions est établi en partie par leur ordre, et celles-ci, une fois liées, peuvent être intégrées en une seule information complexe. Dans l'exemple :

   [11] Il porte secours **et** sauve un gentilhomme attaqué par trois voleurs. Il s'agit de Dom Carlos. (Résumé Dom Juan)

   la proposition *sauve un gentilhomme* est une conséquence de la première *porte secours* ; ainsi, le lien entre les propositions peut-être déterminé par leur ordre.
5. *La connexion subordonnée* : C'est une forme dérivée de la connexion coordonnée où « les faits sont ordonnés hiérarchiquement. Un fait est énoncé comme la spécification d'un autre fait. ». L'exemple :

---

[79] Il y a trois niveaux de représentation du texte dans le modèle de situation. Tout d'abord, le niveau qui correspond à la relation syntaxique des mots et à leur analyse ou *niveau de surface*, ensuite, le niveau qui correspondant à la micro- et macro-structure ou *niveau base du texte*, enfin, le niveau qui correspond à la projection intellectuelle des événements racontés ou niveau *modèle de situation* (Van Dijk/Kintsch 1983).
[80] La notion de « monde possible » apparaît aussi dans la métarègle de congruence de Charolles (1978).

[12] Dors en paix, Marguerite! Il te sera beaucoup pardonné, parce que tu as beaucoup aimé ! (Résumé La dame aux camélias)

montre que la conjonction « *parce que* » qui indique une relation de causalité marque la liaison entre les deux propositions et donc maintient la connexion entre les deux propositions.

Selon cette classification, la cohérence s'apparente à une propriété qui trouve sa source dans des niveaux différents de l'organisation textuelle. Les relations définies doivent être analysées soit au niveau microstructurel _ en d'autres termes, à l'intérieur d'une proposition, au sein d'une phrase ou entre deux phrases successives _ soit au niveau macrostructurel. Les dénominations des relations présentées ci-dessus montrent que le terme de cohérence est réservé par les auteurs aux relations analysables aux plus hauts niveaux de la structure textuelle, qui ne sont pas obligatoirement matérialisés par des marques linguistiques. Une interprétation sémantique du discours qui fait intervenir la notion d'appartenance à un même « monde possible » est nécessaire pour établir les relations de cohérence de cette classification. Par ailleurs, on retrouve dans cette classification le fait que la connexion, décrite comme une forme spécifique de la cohérence, est un des dispositifs pouvant intervenir dans la cohérence. La cohésion n'est pas prise en compte par les auteurs, les dispositifs de cohésion reliraient selon eux des unités lexicales et non pas des propositions.

On distingue donc deux formes de cohérence : la cohérence interne dépendant fortement des relations référentielles et logiques internes au texte et la cohérence externe qui dépend de la relation entre les connaissances générales nécessaires à l'interprétation du texte et les informations qui influencent la cohérence interne : la première serait assurée par des formes linguistiques spécialisées (les pronoms et les connecteurs) et reposerait sur les connaissances linguistiques des interlocuteurs ; la seconde reposerait en particulier sur les procédures d'inférence qui permettent d'avoir accès au contenu implicite du texte et serait la conséquence de l'utilisation des connaissances du monde partagées par les interlocuteurs. La cohérence d'un texte dépend ainsi de l'interprétant et n'est pas une propriété figée. En effet, selon leur niveau de compétence linguistique et les connaissances générales dont ils disposent, les interprétants évaluent le degré de cohérence du discours de façon différente.

Généralement, pour établir la cohérence, les auteurs recourent à des connaissances extérieures aux textes lorsqu'ils analysent des exemples de ce type, puisque, comme nous

l'avons dit précédemment, la cohérence dépend étroitement de l'interprétabilité du discours. Or le sentiment de cohérence d'un locuteur est conditionné par sa capacité à construire un rapport de plausibilité entre les faits dénotés par les énoncés. En effet, cette interprétabilité résulte principalement de son contenu sémantique et non pas uniquement des aspects formels du texte. La cohérence dépend fondamentalement des connaissances référentielles du locuteur. Le locuteur accepte ou non comme plausible la représentation référentielle construite à partir des énoncés. En effet, la cohérence repose sur la possibilité d'établir des relations entre l'information nouvelle et ancienne, sur l'équilibre entre ces informations et sur la possibilité de les intégrer dans une représentation unitaire représentant les visées communicationnelles du locuteur. Certains auteurs vont plus loin (Charolles 1988a, Charolles/ Ehrlich 1991) en affirmant que le besoin de cohérence est une forme *a priori* de la réception discursive pour le sujet. En d'autres termes, le locuteur produit toujours un discours cohérent pour l'interprétant, ainsi, quelle que soit la proximité sémantique entre les différentes informations, le discours sera traité afin d'en retrouver la cohérence. Cette fonction cognitive, appelée cohérence par ces auteurs, permet d'établir des relations entre différents événements indépendamment de leur modalité de communication et serait une propriété générale de l'esprit.

Les différentes analyses de la notion de cohérence font ressortir quelques points saillants:

- Ce ne sont pas les propriétés structurelles internes au texte, mais l'interprétant qui détermine, en dernier ressort, la cohérence du discours ;
- la mise en relation d'informations présentes dans le texte et les connaissances générales sont le fondement de cette activité interprétative ;
- elle a pour objectif de représenter de manière unifiée l'ensemble des informations véhiculées par le texte.

Le texte est dit, alors, cohérent quand il décrit une situation, un univers plausible et quand il existe des relations internes entre éléments du texte grâce à certaines unités lexicales qui le composent, il est dit cohésif.

2. Cohésion et anaphore en production[81]

Quand on s'intéresse aux éléments essentiels à la cohésion et à la cohérence d'un texte, alors il faut s'intéresser à la place de l'anaphore. Ce serait l'utilisation de procédés

[81] Il n'y a pas beaucoup d'études qui se sont intéressées à la caractérisation de la cohérence, de la cohésion et des anaphores dans des textes autres que de type narratifs comme ceux qui sont expositifs.

linguistiques qui détermine la cohésion textuelle. En effet, ceux-ci structurent le discours en y reliant des éléments successifs comme par exemple par le choix des anaphores ou des articles (Ducrot/Schaeffer 1995 : 503). Quand les évaluations de texte se font en fonction de la cohésion référentielle, alors, en général, c'est vis-à-vis du maillon[82] qui précède que l'on évalue si une expression référentielle est appropriée. Les termes de premier maillon des chaînes de référence sont le nom propre et le SN indéfini, quand on s'intéresse à l'introduction des référents dans $RESUMAN_c$. Selon Charolles (1978), il faut utiliser des pronoms et des anaphores pronominales avec un rappel strict afin de conserver le référent.

Nous avons pris en compte deux définitions d'Adam afin de caractériser le modèle génératif du texte. Tout d'abord, celle d'Adam (1994) où ce sont les notions de séquence qui définissent le texte ; ensuite, celle d'Adam (2001) où c'est une combinaison de macro-propositions, reliées par des processus maintenant la cohésion, qui structure le texte. Dans cette partie, nous étudions, les anaphores pronominales. En effet, elles permettent une interprétation sémantique du texte et assurent, dans $RESUMAN_c$, sa cohésion. Dans les textes de $RESUMAN_c$, l'anaphore pronominale existe sous les formes morphologiques suivantes :

Types de pronoms	Personnels il(s), elle(s), la, lui, en, etc.	Démonstratifs ça, cela, ce, c'	Relatifs qui, que, dont, où, lequel

Tableau 7 : Répartition des anaphores pronominales dans $RESUMAN_c$

En analysant, dans $RESUMAN_c$, la fréquence d'apparition des anaphores pronominales, on constate que, relativement aux pronoms relatifs et démonstratifs, ce sont les pronoms personnels qui sont les plus fréquents (taux le plus manipulé).

L'optique substitutionnelle, thèse critiquée[83], place le substitut au même niveau que le substitué. De plus, elle considère le pronom personnel comme un simple substitut de son antécédent. Ainsi, nous ne considérerons pas cette optique, mais plutôt celle qui soutient que remplacer un pronom par un SN ne conduit pas toujours à des propositions acceptables

[82] Si l'interprétation référentielle d'une expression est sous la dépendance d'une ou autre(s) présente(s) dans le discours, alors c'est une expression anaphorique (Kleiber 1994). La notion de *chaîne de référence* a été caractérisée par Schnedecker, qui en 1997 compare ces expressions à des maillons d'une chaîne constituant l'entité concernée.

[83] La critique a été faite par Brown/Yule et reprise par Kleiber (1994) ainsi que Charolles et Schnedecker (1993).

ou qui ont le même sens. Par ailleurs, nous sommes consciente que la substitution est associée à un phénomène de cohésion ou un procédé de coréférence et que nous considérons que dans un but économique, dans le développement linéaire du discours, la reprise pronominale est une substitution obligatoire.

Afin d'éviter la répétition et de garantir la structuration textuelle et sa cohésion dans $RESUMAN_c$, c'est la reprise pronominale de la 3ème personne (rappels et répétitions du référent au personnage principal) qui est employée par l'auteur (scripteur/narrateur). Dans notre corpus, c'est cette stratégie, parmi d'autres, d'utilisation des pronoms personnels qui permet à l'auteur de reprendre l'information de l'antécédent de manière continue et de construire un enchainement d'idées. Dès lors, la question de la cohésion textuelle se pose, de même que celle de la grande fréquence d'apparition, dans le corpus, des pronoms personnels en tant que formes anaphoriques. En effet, le narrateur-scripteur, pour garantir le rappel et le maintien du référent qu'il a introduit, utilise prioritairement le pronom *il*. Cette fréquence d'utilisation importante conforte l'hypothèse qui privilégie toujours l'anaphore pronominale dans les narrations.

Ainsi, c'est à un actant principal (sujet ou complément), c'est-à-dire un référent que l'on peut identifier dans le contexte linguistique, que le pronom personnel à fonction anaphorique fait référence. Ceci est mis en évidence dans l'exemple [13]. En effet, ici c'est à l'actant principal que fait référence le pronom et pour anaphoriser les autres actants, l'auteur utilise des noms propres ou, s'il y a un terme compétiteur, celui-ci utilise une reprise-répétition :

> [13] **Sganarelle**, désinvolte, répond aux interrogations de **Gusman**. **Il** lui enlève ses illusions et esquisse un portrait de son maître, libre penseur, "grand seigneur méchant homme" et " épouseur à toutes mains». Arrive **Dom Juan** : **il** confie à **Sganarelle** que seule la conquête l'intéresse. **Il** évoque l'inconstance de l'amour et dévoile à son valet le secret de son propre caractère : **il** ne peut s'attacher à aucune femme, et rêve, tels les grands conquérants, de succès sans cesse recommencés. (Résumé Dom Juan)

En [13], le pronom personnel a une fonction thématique. En effet, il réfère au personnage principal désigné par un nom propre et participe au maintien de sa première mention. Au départ, le pronom personnel anaphorique *il* fait référence au SN *Sganarelle*. Mais l'apparition de *Dom Juan* comme personnage principal peut créer une ambigüité. Afin de lever cette l'ambiguïté, c'est au deuxième sujet que l'auteur dédie le nom propre répété plus tard. Ceci désigne, par ailleurs, le premier personnage comme élément participant à la progression hiérarchique des évènements.

Afin de conserver, dans l'enchainement des événements, une chaîne de référence relative à un actant principal, on a recours souvent aux pronoms personnels compléments qui sont importants pour la cohésion textuelle. Néanmoins, l'intérêt est porté sur les actions de l'actant principal réalisées dans la trame évènementielle et non sur l'actant lui-même, même si l'anaphore garantit la chaîne de référence qui lui est relative. Selon Fayol (1997), dans la stratégie de pronominalisation, on sélectionne d'abord un ou plusieurs personnages principaux et, ensuite, on les place dans une position initiale dans laquelle ils seront maintenus. Comme illustré dans les exemples [14] à [17], afin de déterminer la cohésion textuelle, la forme pronominale personnelle est la plus utilisée dans $RESUMAN_C$.

[14] Le voyage a lieu sous la pluie. Arrivée aux Peuples. **Jeanne** goûte avec **son père** la joie de redécouvrir le château de **son enfance**. **Elle** passe sa première nuit, à la fenêtre de sa chambre, à rêver au clair de lune. **Elle** attend un prince charmant dont « **elle** savait seulement qu'**elle** l'adorerait de toute son âme et qu'il **la** chérirait de toutes ses forces ». (Résumé Une vie)

[15] **Fils de Gervaise Macquart et de son amant Lantier**, **le jeune Etienne Lantier** s'est fait renvoyer de **son travail** pour avoir donné une gifle à **son employeur**. Chômeur, **il** part, en pleine crise industrielle, dans le Nord de la France, à la recherche d'un nouvel emploi. **Il** se fait embaucher aux mines de Montsou et connaît des conditions de travail effroyables. (Résumé Germinal)

[16] **Mrs C.** avait alors quarante deux ans et avait perdu **son mari** deux ans auparavant. **Elle** décide de se rendre à Monte Carlo et fréquente alors les casinos. **Elle** aime à examiner les mains des joueurs. Ces gestes qu'**elle** observe lui permettent de comprendre leur personnalité sans même avoir à regarder leur visage. Un jour, **elle** est fascinée par des mains magnifiques. **Elle** ne peut résister. **Elle** regarde alors **ce joueur** et découvre un beau jeune homme d'environ vingt quatre ans. **Il** semble totalement anéanti car **il** vient de perdre tout son argent. (Résumé Vingt-quatre heures de la vie d'Une femme)

[17] **M. et Mme Smith** ont fini de dîner. **Ils** bavardent au coin du feu. M. Smith parcourt son journal. **Le couple** se répand en propos futiles, souvent saugrenus, voire incohérents. **Leurs raisonnements** sont surprenants et **ils** passent sans transition d'un sujet à un autre. (Résumé La Cantatrice Chauve)

Quelle que soit la situation, l'actant sujet ou le personnage principal a toujours pour référent *il*. Par exemple, il n'y a qu'un seul actant dans [18] qui est repris par un pronom. Les deux personnages principaux *M. et Mme Smith* dans [21], sont repris d'abord par le pronom *ils*, ensuite par un défini *le couple* et enfin, encore par *ils*. C'est par des anaphores nominales, ou définies (*le père*,..) que le reste des personnages est repris. Ci-dessous nous représentons la chaîne relative à l'actant :

[18] Jeanne ….son père …..son enfance….. Elle. ….Elle ….elle ……elle
[19] Fils de Gervaise Macquart et de son amant Lantier, le jeune Etienne Lantier …son travail ….son employeur. …il. …..Il.
[20] Mrs C. ….son mari …..Elle …..elle ……elle ……. Elle…….. Elle …….ce joueur ….. Il ….il.
[21] M. et Mme Smith……. Ils ….Le couple …..Leurs raisonnements ……ils.

Nous remarquons que les actants sont très souvent repris par des pronoms. Cela peut se faire sans aucune ambiguïté comme dans l'exemple (18) où l'antécédent est définit par un syntagme nominal et ensuite repris par une répétition successive du pronom personnel *elle* (quatre occurrences). Dans d'autres cas, il peut y avoir une ambigüité. Dans l'exemple (19), le résumeur dédie une chaîne anaphorique au personnage principal pour appuyer son rôle prépondérant dans le déroulement des événements. Ici, l'antécédent est introduit par un syntagme nominal qui est lui-même introduit par un nom indéfini, puis il est repris par un syntagme nominal et enfin, par deux occurrences de *il*. Néanmoins, l'introduction de *son employeur* peut créer une ambigüité relative au référent du pronom *il* et afin d'y remédier, l'auteur introduit l'adjectif *chômeur*. Dans l'exemple (20), l'introduction d'un deuxième personnage n'induit pas d'ambigüité. En effet, le nom propre (antécédent) est repris par le pronom personnel *elle* (quatre occurrences) et on ne peut pas le confondre avec le deuxième actant *ce joueur* qui a pour référent *il*. Ainsi, en présence de termes compétiteurs, le résumeur utilise le genre pour différencier les personnages. Donc dans cet exemple, il n'y a pas de compétiteurs qui peuvent introduire une ambigüité dans l'interprétation (Fayol 1997 : 193). La stratégie de reprise pronominale dominante dans $RESUMAN_c$ est l'alternance de reprise nominale (pronominalisation de l'actant principal et reprise nominale pour les autres), ceci évitant la confusion entre les antécédents.

C'est principalement au personnage principal que font référence les séquences pronominales, ce qui le rend saillant[84] et donc plus accessible (Fayol 1997 : 192). Rappelons que la pronominalisation des SN[85] pour introduire les actants assure la continuité voire la cohésion textuelle. Selon Reichler-Beguelin (1989), le bon usage des expressions référentielles implique donc, avant tout, une capacité de décentration de la part de celui qui écrit, incité à adopter par anticipation le point de vue du décodeur et à estimer convenablement les connaissances dont celui-ci dispose. Elle observe que, du point de vue de l'encodeur, une aptitude spécifique est nécessaire afin d'utiliser des procédés anaphoriques. Celle-ci permet de maintenir dans la mémoire discursive un thème privilégié ou saillant comme contrôleur des anaphores mais aussi à changer de thème sans négliger les interférences possibles de certains facteurs sémantiques.

[84] Nous reviendrons sur la notion de la saillance, avec plus de détails, dans le chapitre suivant.
[85] Nous citons aussi de la variété des thèmes et des formes textuelles qu'il est possible d'anaphoriser.

Dans les phrases produites, des personnages secondaires peuvent apparaître en position thématique. Une reprise nominale, généralement définie, du nouveau référent thématique marque ce changement thématique. Il semble qu'en production comme en compréhension, la stratégie de maintien du thème soit le reflet de ce que Kinstch et van Dijk (1983 : 167) qualifient de « position privilégiée en mémoire ou dans l'attention »[86] du topique. Les structures thématiques produites et la diversité des dispositifs cohésifs utilisés sont en relation.

3. L'impact de l'anaphore sur la cohérence

L'anaphore et la cohérence sont considérées comme deux éléments fréquemment liés : l'anaphore souligne la cohérence d'un discours et la cohérence permet d'attribuer des antécédents aux termes anaphoriques. L'influence bilatérale entre le choix référentiel et la relation de cohérence a plusieurs fois été caractérisée. Le but du lecteur est d'arriver à une structure interprétative plus globale et doit intégrer le contenu et la valeur discursive des énoncés. Ce sont les relations anaphoriques interphrastiques mais aussi les relations de cohérence qui peuvent l'aider, puisqu'elles permettent l'établissement de la continuité référentielle. En effet, une séquence de phrases et de propositions ne peuvent former un texte sans cette continuité.[87]

Notre but est, d'abord, d'évaluer l'importance de l'anaphore, plus spécifiquement, pronominale sur la cohérence du texte et ainsi sur sa compréhension. Nous prenons un texte de RESUMAN[88] et le présentons en deux versions différentes dans le tableau ci-dessous. Tout d'abord, la version originale supposée être fortement cohérente, ensuite, une version où les phrases sont disposées dans un ordre aléatoire.

Version originale	[22] La première fois qu'Aurélien vit Bérénice, **il** la trouva franchement laide. **Elle** lui déplut, enfin. **Il** n'aima pas comment **elle** était habillée. Une étoffe qu'**il** n'aurait pas choisie. **Il** avait des idées sur les étoffes. Une étoffe qu'**il** avait vue sur plusieurs femmes.
Version aléatoire	[23] La première fois qu'Aurélien vit Bérénice, **il** la trouva franchement laide. Une étoffe qu'**il** avait vue sur plusieurs femmes. **Elle** lui déplut, enfin. **Il** avait des idées sur les étoffes. **Il** n'aima pas comment **elle** était habillée. Une étoffe qu'**il** n'aurait pas choisie.

[86] "Privileged position in memory or consciousness"
[87] Si la présence de l'anaphore est nécessaire à la cohérence (ce que nous démontrerons dans cette section), elle est certainement non suffisante. Voir, à cet égard par exemple, Reboul (1997) et Charolles (2005).
[88] Nous avons utilisé antconc pour repérer un passage plus au moins dense en occurrences anaphoriques.

Version modifiée	[24] La première fois qu'il vit elle, **il** la trouva franchement laide. **Elle** lui déplut, enfin. **Il** n'aima pas comment **elle** était habillée. Elle qu'**il** n'aurait pas choisie. **Il** avait des idées sur **eux**. Elle qu'**il** avait vue sur plusieurs.

Tableau 8 : Deux versions d'un extrait du résumé d'Aurelien.

On peut appliquer pour l'exemple [22], où le pronom *il* désigne *Aurélien* et *elle, Bérénice,* une interprétation traditionnelle de l'anaphore. En effet, dans cet exemple, il y a possibilité de substitution entre l'anaphorique et l'antécédent. Nous remarquons, qu'en [23], malgré le déplacement des phrases, le texte a gardé sa cohérence mais une autre lecture ou compréhension est admissible : le sens du texte a changé dès que l'attribution de l'antécédent *étoffe* à *elle* s'est opérée. Notre hypothèse est que le pronom en présence d'un antécédent de même genre et nombre, devient significatif malgré la fausse inférence de compréhension. Nous avons modifié les antécédents par leurs pronoms adéquats en [24] et le texte est devenu incohérent. Nous ne pouvons plus identifier l'identité des pronoms au travers de ses modifications. La suppression de la relation anaphorique a brisé la cohérence textuelle. Il serait très difficile de construire un discours cohérent sans utiliser d'anaphores qui renvoient aux phrases antérieures. Ainsi, la cohérence textuelle n'est possible que par l'existence d'anaphores. Ce test montre que les versions cohérentes peuvent être mieux comprises que les versions aléatoires. En effet, une partie de la cohérence peut être rétablie grâce à la cohésion référentielle. Néanmoins, la difficulté du texte modifié [24] pourrait masquer un éventuel effet de la cohérence : les descriptions sont composées d'un ensemble de termes appartenant au même cadre conceptuel définissant un thème précis, ce qui permettrait le rétablissement de la cohérence à partir des informations cohésives reposant sur la mise en œuvre de capacités inférentielles,

Nous nous sommes donnée pour but dans la section qui suit d'évaluer l'interaction entre anaphore et cohérence textuelle. Ainsi, nous présenterons tout d'abord la théorie de Hobbs (1979) qui porte sur la relation entre l'établissement des relations de cohérence et la résolution des anaphores. Ensuite, nous présenterons trois textes et, grâce à la relation entre l'anaphore pronominale et la relation *Assertion-Indice* (Cornish 2009a et b), nous les analyserons en fonction de quelques marques linguistiques pour l'interprétation des relations de cohérence. Nous formulerons ces analyses en supposant que le lecteur lit les parties du texte linéairement en temps réel. Le rôle des anaphores dans les phrases consécutives des textes proposés sera étudié dans ce cadre.

3.1. L'hypothèse de Hobbs[89] (1979)

L'hypothèse avancée par Hobbs (1979) est la suivante : il faut choisir une relation de cohérence pour intégrer les relations logiques dans un texte pour réaliser l'interprétation de(s) l'anaphore(s). Ainsi, l'interprétation des anaphores interphrastiques[90] (Cornish 2009b : 159) pourrait être une conséquence de l'usage d'une relation de cohérence pour assimiler deux entités du texte. Seuls les principes utilisés pour établir cette relation de cohérence seront nécessaires pour l'interprétation de l'anaphore. En effet, selon Hobbs les solutions à de nombreux problèmes de référence et de coréférence sont simplement engendrés au cours de la reconnaissance des relations de cohérence (1979 : 68).

En guise d'illustration, Cornish (2009b) a ré-analysé l'exemple phare de Hobbs (1979 : 78, ex. (3)) :

[25] John can open Bill's safe. He knows the combination.

Nous allons suivre la même démarche, selon Hobbs (1979) et d'après Cornish (2009b), en analysant un exemple[91] de notre corpus relevant un cas d'ambiguïté anaphorique :

[26] Charles Bovary épouse sous l'influence de **sa mère** une veuve de quarante-cinq ans, riche, laide et tyrannique, **Mme Dubuc**. **Elle** aime Charles avec passion. (Résumé Madame Bovary)

Dans l'exemple [26], dans la première phrase il y a deux référents qui pourront l'un ou l'autre être repris dans la phrase suivante par *elle*. En outre, l'état psychologique « être aimer avec passion » est une propriété sémantique qui se rapporte autant à *Mme Dubuc*, étant l'épouse de *Charles* qu'à *la mère de Charles Bovary*. Néanmoins, la proposition logique exprimée dans la deuxième phrase ne peut fournir un Indice (ou une *Preuve* ('Evidence')), sur l'*Assertion* ('Claim') faite à propos de *Mme Dubuc*, que quand on s'aperçoit que le pronom *elle* renvoie à *Mme Dubuc*. On peut déduire que, la première phrase, traitée comme un énoncé, exprime « une structure informationnelle thétique » (où l'information portée est présentée comme « toute nouvelle »). Quant à la deuxième phrase, elle reprend *Mme Dubuc* comme le référent le plus saillant évoqué comme topique dans la

[89] Suite à la publication du modèle de Kintsch/Van Dijk (Kintsch/Van Dijk 1978 ;Van Dijk/Kintsch 1983), de nombreux travaux sur le traitement automatique et de psycholinguistiques (comme Mann/Thompson 1988) se sont ajoutés aux travaux de linguistique descriptive et théorique sur la pertinence ou cohérence des discours. Nous n'avons pas étudié la *Rhetorical Structure Theory* (RST) de Mann et Thompson (1988) car nous utiliserons l'approche de Hobbs ultérieurement dans la troisième partie de notre travail.

[90] Le cas majeur des anaphores de notre corpus.

[91] Nous signalons que notre exemple est contextualisé à la différence de celui de Hobbs (1979).

première et elle est plus traitée comme articulation « catégorique » (à topique-commentaire). Si on avait conclu que le pronom *elle* renvoyait plutôt à *la mère de Charles*, alors la proposition logique créée n'aurait pas donné d'information additionnelle concernant *Mme Dubuc.* De plus, grâce à l'affirmation de la phrase initiale, la continuité de la situation évoquée n'est pas garantie. En effet, de toutes les façons il est difficile d'interpréter *elle* car ce type de pronoms est focalisé sur la reprise de référents très topicaux et présupposés.

Selon Kleiber (1994), le pronom *elle*, au vu de ces propriétés sémantiques et référentielles, a tendance à convoquer « la connexion la plus forte » entre les phrases en saisissant le référent en continuité avec ce qui l'a rendu proéminent précédemment. Ainsi, l'interprétation du pronom *elle* n'est pas seulement l'installation d'une relation de cohérence intégrant deux unités (Hobbs, 1979), mais cette relation est un nécessaire pour l'emploi d'une relation adéquate. Si A épouse B et si B est un jeune homme, alors c'est A qui aime avec passion B (A est *une veuve de quarante-cinq ans, riche, laide et tyrannique*). Ainsi, selon Hobbs, nous pouvons analyser la seconde phrase de l'exemple [16] par rapport à la première comme étant dans une relation d'*Élaboration*[92] (une proposition identique est inférée dans les deux cas en employant différents mots). Néanmoins, si la phrase initiale effectuait une Assertion particulière qui serait confirmée dans la seconde, alors cette relation entre les deux propositions serait renforcée (ici *épouser* motive cette interprétation). Selon Cornish (2009b : 161), un principe fondamental de la communication est qu'il faut une bonne raison pour révéler au lecteur ce qu'il sait déjà : ainsi, la compréhension du texte [16] ne doit pas dépendre d'une simple paraphrase par la seconde phrase de la première (cela aurait pu être le cas si le locuteur/scripteur voulait persuader l'allocutaire/lecteur que son assertion via la première était crédible ; dans ce cas, la deuxième phrase fournirait une preuve de l'assertion effectuée via la première).

3.2. Relations de cohérence et résolution de l'anaphore

Dans cette section, nous allons utiliser comme base les conceptions de Hobbs (1979) sur les relations de cohérence et résolution des anaphores.

3.2.1. Relations de cohérence

La Causalité, le *Fond-Figure* et l'*Expansion* sont les trois types principaux fondant les principes des relations de cohérence selon Hobbs. Cornish (2009b) a ajouté des

[92] Nous définissons la relation *Elaboration* dans la sous-section suivante.

modifications résumées dans le tableau 7[93]. Tout d'abord, il a classé sous trois rubriques les sous-catégories de relations de la théorie de Hobbs (en plus de trois autres). Ensuite, il a modifié (*Élaboration*, *Explication*) et ajouté (*Cause-Conséquence/Résultat*, *Circonstance*, *Assertion-Indice*) à sa propre définition. Enfin il aussi modifié la classification de Hobbs (*Explication*, transférée de « *Fond-Figure* » à « *Causales* »).

Relations Causales	**Relations Fond-Figure**	**Relations d'Expansion**
Occasion : 1) Inférer un changement d'état de l'assertion de $S^{\mathbf{0}}$, dont l'état final peut être inféré de S^1. 2) Inférer un changement d'état de l'assertion de S^1, dont l'état initial peut être inféré de $S^{\mathbf{0}}$. (Hobbs, 1990 : 87)	*Fond-Figure* : Inférer de $S^{\mathbf{0}}$ une description d'un système d'entités et de relations, et inférer de S^1 qu'une certaine entité est située ou bien se déplace dans ce système en tant qu'arrière-fond. (Hobbs, 1990 : 91)	*Parallèle* : Inférer $p(a_1, a_2,\ldots)$ de l'assertion de $S^{\mathbf{0}}$ et $p(b_1, b_2,\ldots)$ de l'assertion de S^1, où a_i et b_i sont semblables, pour tout *i*. (Hobbs, 1990 : 93)
Cause-Conséquence [« *Résultat* »] : cas particulier de la relation Occasion (sa version *forte*). Inférer que l'état ou l'événement asserté via $S^{\mathbf{0}}$ cause ou pourrait causer l'état ou l'événement asserté via S^1. (Cornish 2009b)	*Circonstance* : Une proposition P^0 exprimant un état, processus ou événement dans S^0 sera comprise comme fournissant le cadre temporel, spatial ou cognitif à l'intérieur duquel l'événement dénoté par S^1 est à situer. L'événement principal (exprimé par S^1) devra soit être totalement inclus dans l'événement ou l'état circonstanciel ($S^{\mathbf{0}}$) ou le chevaucher. Définition basée sur celle de Mann/Thompson, (1988 : 272). (Cornish 2009b)	*Élaboration* : Inférer la même proposition P de l'assertion de $S^{\mathbf{0}}$ comme de S^1. Ceci correspond en fait à la relation Parallèle lorsque les entités semblables a_i et b_i sont identiques, pour tout *i* (Hobbs 1990 : 95). De plus, S^1 doit permettre d'ajouter d'autres détails à la proposition commune inférable de chaque assertion, et $e^1 \leq e^{\mathbf{0}}$ (l'événement principal évoqué par S^1 est une partie de celui dénoté par $S^{\mathbf{0}}$). (Cornish 2009b)
Explication : Inférer que l'état ou l'événement asserté via S^1 cause ou pourrait causer l'état ou l'événement asserté via $S^{\mathbf{0}}$ (Hobbs, 1990 : 91). En outre, $e^1 > e^{\mathbf{0}}$ (l'événement principal évoqué par S^1 précède celui désigné par $S^{\mathbf{0}}$) (Asher/Lascarides 2003 : 160). Le locuteur/scripteur a l'intention que l'auditeur/le lecteur soit persuadé de la relation causale existant entre les deux éventualités. (Cornish 2009b)		*Contraste* : 1) Inférer *p(a)* de l'assertion de $S^{\mathbf{0}}$ et ¬ *p(b)* de l'assertion de S^1, où *a* et *b* sont semblables. 2) Inférer *p(a)* de l'assertion de $S^{\mathbf{0}}$ et *p(b)* de l'assertion de S^1, où il existe quelque propriété *q* telle que *q(a)* et ¬ *q(b)*. (Hobbs 1990 : 99)
		Assertion-Indice : 1) Inférer P de l'assertion de $S^{\mathbf{0}}$ ainsi que de S^1, où S^1 ajoute d'autres détails à P et $e^1 \leq e^{\mathbf{0}}$ (= la relation *Élaboration*) et 2) Interpréter S1 comme rendant plus convaincante l'hypothèse du locuteur qui correspond à l'assertion de $S^{\mathbf{0}}$. (Cornish, 2009b)

[93] Les symboles ' S^{0}', 'S^{1}', Etc., indiquent '*unité initiale de la proposition*', '*seconde unité de la proposition*', etc., dans un texte donné. les autres formules à l'intérieur de chaque cellule sont commentées au fur et à mesure de leur apparition dans notre rédaction.

Tableau 9 : Définitions d'un sous-ensemble de Relations de Cohérence selon Cornish (2009a et b), et d'après, Hobbs (1990, ch. 5)

Selon Hobbs (ch. 5, 1990) c'est sur la *causalité* que se baseraient les relations *Occasion* et *Cause-Conséquence* dans la colonne « relations causales ». La relation où l'événement évoqué dans la proposition initiale (ou la deuxième) prépare celui évoqué dans celle qui suit (ou la précédente) correspond à la relation *Occasion* (Tableau 8). Nous illustrons cette relation dans l'exemple suivant :

> [27] [S^0_{EV}[94] **Les gardes** font **leur** rapport]. [S^1_{ET} **Le roi** ne veut pas **les** croire]. [S^2_{EV} **Il** interroge **sa** nièce [S^3_{EV} **qui** avoue aussitôt]]. [S^4_{EV} **Il** fait alors mettre **les gardes** au secret, [S^5_{ET} avant que le scandale ne s'ébruite]]. (Résumé Antigone)

Tout d'abord, la paire S^1 et S^2 serait reliée en fonction de la relation *Causale*. La volonté de ne pas croire quelqu'un est préalable à l'action d'interroger d'autres personnes. En fonction de la relation *Cause-Conséquence*, S^2, *Il interroge sa* nièce, rapporte l'événement *Conséquence* à S^1, *Le roi ne veut pas les croire*. Comme exemple de *Résultat*, nous citons S^4 qui suit l'unité composite [S^1]+ [S^2+ [S^3]].

> [28] [S^1_{ET} **Le roi** ne veut pas **les** croire]. [S^2_{EV} **Il** interroge **sa** nièce [S^3_{EV} **qui** avoue aussitôt]]. [S^4_{EV} **Il** fait alors mettre **les gardes** au secret, [S^5_{ET} avant que le scandale ne s'ébruite]]. (Résumé Antigone)

La relation *Explication* « suppose que le locuteur/scripteur a l'intention que l'éventualité causatrice *explique* celle qui est causée pour l'allocutaire/le lecteur » (Cornish 2009b : 171) : à savoir, l'événement causateur précède normalement l'événement causé. Comme illustration, nous renvoyons à la relation reliant S^4 et S^5. Les relations causales ont pour arguments les propositions reliées qui expriment les propositions logiques.

Chez l'allocutaire, les relations *Fond-Figure* (Tableau 8) relieraient des propositions au savoir préexistant. Le lien entre S^0 et S^1 (exemple 27, ci-dessus) est un cas de relation *Fond-Figure*, tout comme le cas de relations de Circonstance[95]. Pour finir, les relations d'*Expansion* (Tableau 8) pourraient être des duplications du principe de *Ressemblance*. Ces relations, selon Hobbs, facilitent les processus inférentiels du lecteur et impliquent

---

[94] Chaque unité minimale du texte (une proposition qu'elle soit coordonnée ou subordonnée, finie ou non, principale ou indépendante, ou bien nominalisée —voire à un SN ou SP *cadratif*) est annotée soit pour son rôle sémantique soit pour sa structure événementielle (Localisation temporelle : *LOC-TPS*). « ET correspond à « état » et « EV » correspond à « événement ». Hobbs (1990) envisage les unités minimales du texte comme unités phrastiques, notées par les symboles préfixés S^0, S^1, etc. Quant à Cornish (2009 a et b), il emploie UD^0, UD^1, UD symbolise *Unité de Discours*. Nous adoptons l'annotation de Hobbs vu que nous y reviendrons lors de notre approche proposée dans la troisième partie.

[95] *Cf.* la 2e colonne du Tableau 2 (2e cellule) pour la définition à inflexion plus temporelle que Cornish propose à la Hobbs, en partie inspirée par celle de Mann et Thompson.

des relations inférentielles entre les segments du co-texte. Ainsi, ils ne font pas avancer et ne développent pas l'arrière-plan du discours mais au contraire, elles l'étendent *in situ*. Quant à la définition émise par Hobbs de la relation *Élaboration* (Tableau 8), selon Cornish (2009b : 169), celle-ci ne devient qu'une simple relation de *Paraphrase* (voir l'exemple de Hobbs, 1979 donné comme [25] ci-dessus). De ce fait, selon l'auteur, la proposition commune inférée de chaque phrase devra être enrichie par des informations de la proposition élaborante, S^1. Une définition fondamentalement « sémantique » de cette relation, basée essentiellement sur le contenu sémantique des unités, a été proposée par l'auteur (cf. Cornish, 2009b : 172, note 19) :

> Inférer la même proposition P à partir de l'assertion de S^0 comme de S^1. De plus S^1 doit permettre d'ajouter d'autres détails à la proposition commune inférable de chaque assertion, et $e^1 \subseteq e^0$ (l'événement principal évoqué par S^1 est une partie de celui dénoté par S^0),

Cornish (2009a et b) classe la relation *Assertion-Indice* avec les relations d'*Expansion*, elle intègre deux parties : une, relative aux contenus propositionnels des unités reliées, partie élaborante « sémantique » [*cf.* 1) ci-dessous], et une autre, relative aux relations entre les actes illocutoires associés, partie argumentative, « pragmatique », [*cf.* 2) ci-dessous] :

1) Inférer P de l'assertion de S^0 ainsi que de S^1, où S^1 ajoute d'autres détails à P et $e^1 \subseteq e^0$ [e^1 et e^0 sont les événements principaux évoqués par S^1 et S^0] (= la relation *Élaboration*)

2) Interpréter S^1 comme rendant plus convaincante l'hypothèse du locuteur qui correspond à l'assertion de S^0. (Cornish, 2009b : 169)

La valeur argumentative de la deuxième unité est soulignée dans la deuxième partie de sa définition. Cette fonction argumentative est comparable à celle de la relation de présentation, qui augmente la croyance du lecteur, ou relation *Démonstration* (*Evidence*) (Mann/Thompson 1988)). Le but de cette deuxième unité est donc de convaincre le lecteur : « *la seconde des deux unités impliquées,* écrit Cornish (2009b : 175), *devra être interprétable comme rendant l'assertion de la première plus convaincante pour l'allocutaire ou le lecteur* ».

Selon Hobbs, les relations de cohérence influencent l'interprétation des anaphoriques et un exemple d'*Élaboration* est représenté par [25][96] :

[96] *John can open Bill's safe. He knows the combination.*

> By assuming that 'he' refers to John and that the combination is the combination of Bill's safe, we have the same proposition *P* and have thus established the *elaboration* relation (and solved some coreference problems as a by-product […]).

John ne peut ouvrir le *coffre-fort de Bill* que s'il connaît la combinaison, ici la deuxième phrase (S^1) implique la première (S^0). On comprend dans chaque proposition que *John* peut ouvrir le *coffre*-fort de Bill (ici l'assertion de S^0 est la proposition inférable *P* ; *cf.* Hobbs, 1990 : 96). Néanmoins, c'est aussi parce qu'il connaît la combinaison du *coffre-fort de Bill* que *John* est capable de l'ouvrir. Hobbs conçoit qu'ici, *Élaboration* peut se mélanger avec *Explication.* En effet, il existe une proximité entre cette inférence et une relation causale (S^1 cause S^0).

Néanmoins, le fait que *John* connaisse la combinaison soutient l'hypothèse qu'il est capable d'ouvrir le *coffre-fort*. Ainsi, selon Cornish, la seconde phrase fournit une raison d'y croire, un argument de l'assertion effectuée en S^0, en plus de son rôle en tant qu'élaborante (2009a et b). Ici la relation de cohérence n'est pas strictement sémantique comme la relation d'*Explication* de Hobbs mais plutôt pragmatique et S^1 a une dimension démonstrative. Pour l'accentuer, on pourrait ajouter un connecteur comme *en effet* ou *après tout* entre les deux phrases (Cornish, 2009b) (*John peut ouvrir le coffre-fort de Bill. Après tout / En effet, il connaît la combinaison*). Le fait que l'insertion du marqueur *après tout* ne modifie pas l'interprétation d'origine démontre la pertinence de la relation *Assertion-Indice* qui dans le cas de notre corpus représente plus qu'une *Élaboration* [16].

3.2.2. Interaction anaphore et relations de cohérence

Commençons par analyser un texte de notre corpus, qui contient des anaphores pronominales ambiguës[97] :

> [29] [S^0_{TPS} Au milieu d'un souper de carnaval], [S^1_{EV} alors que **Candide** dîne avec six malheureux anciens rois qui ont perdu leur royaume, [S^2_{EV} **il** retrouve **Cacambo** qui est devenu esclave]]. [S^3_{EV} **Il lui** apprend que [S^4_{EV} **Cunégonde l'**attend sur les bords de la Propontide, près de Constantinople]]. (Résumé Candide)

Tout d'abord, nous avons annoté manuellement [29] en le divisant en des unités de discours S^0, S^1, S^3 et S^4 selon leur structure d'événement (EV), ou bien leur rôle sémantique (TPS). S^0, cadrative temporelle, qui modifie S^1, serait reliée à la proposition principale en fonction de *Circonstance*, car elle fournit un repère temporel pour l'action (*au milieu d'un souper de carnaval)*. S^1 présente la situation de base qui forme la trame du

[97] Selon notre propre intuition.

résumé : le temps est un présent de narration (cette valeur est en partie due au type du texte, un résumé de roman dans lequel on s'attend au récit d'une histoire). La paire d'unités S^1 et S^2 serait reliée en fonction de la relation *Explication* (initialement, en fonction de la relation *Occasion*, à cause de la présence du connecteur temporel *alors que* et de la succession de deux événements : le fait de dîner avec des rois, préalable à l'action de rencontrer des esclaves). Nous sommes, donc, en présence d'une unité [S^0 [+S^1 [+S^2]]] thétique qui se rapporte à l'événement « cadre » de ce qui constitue l'essentiel de l'information. Pour sa part, la proposition principale représentée par l'unité S^2 serait reliée à S^1 (par l'intermédiaire de la localisation temporelle *alors que*), car elle renvoie obligatoirement au référent du sujet de S1 grâce au fonctionnement du sujet pronominal référentiel *il*, cette phrase va naturellement revêtir un statut informationnel sur la catégorie (relation *Explication*).

Ensuite, l'unité composée [S^3 + S^4] sera reliée à l'unité [S^0 [+S^1 [+S^2]]] en fonction de la relation *Élaboration*, car cette unité donne des détails supplémentaires sur la situation évoquée dans l'unité initiale. Le référent des pronoms *il* et *lui*, sont inférés en vertu de la proposition exprimée dans S^1 *il retrouve Cacambo* dont *il = Candide*. L'antécédent de ces deux pronoms est flou : nous sommes devant deux personnages (*Candide* et *Cacambo*). La variable d'argument Agent est disponible lexicalement grâce à l'emploi du verbe transitif *apprendre* (prédicat à trois arguments sémantiques[98] : Agent, Patient et Bénéficiaire, respectivement). Le rôle des esclaves est de rendre service à leurs maîtres et comme Candide fréquente des rois, dans ce court passage, *Cucambo* doit lui rendre service en lui apportant des nouvelles de *Cunégonde*[99] (l'apparition du pronom *elle* dans S^4 apporte l'indication *femme*). L'argument Agent *il*, exprimé en tant que sujet syntaxique, vient alors à désigner *Cucambo, qui est devenu esclave* et l'argument Bénéficiaire désigne *Candide*.

D'autres informations sur les référents *il* et *lui* sont fournies par les unités de discours S^3 et S^4 ; ainsi, nous pouvons raccrocher l'unité [S^3 + S^4] à la macro-unité contextuelle [S^0

[98] Ou argument sémantique (Tesnière 1965) :
Agent ➔ être animé instigateur d'un procès, d'une action
Thème/Patient ➔ entité non animée/animée sur laquelle s'exerce directement le procès
Siège (état) ➔ entité où se manifeste un état physique/psychique
Bénéficiaire ➔ être animé affecté par les retombées du procès
Instrument ➔ entité non animée qui est à l'origine du procès
Locatif ➔ repère spatial du procès
But ➔ entité concrète ou abstraite vers laquelle est dirigé le procès
Source ➔ entité dont provient ou s'éloigne une autre entité
Résultatif ➔ objet, être ou état qui est la conséquence du procès

[99] Généralement, ce sont les serviteurs esclaves qui jouent le rôle de messager entre un maitre et une femme.

[+S^1 [+S^2]]] et les unités macro partagent un cadre contextuel identique[100]. Cet exemple [18] est représentatif de la fonction intégratrice de l'anaphore à la construction de la cohérence. En effet, S^3 et S4 sont reliés au niveau de la paire [S^3 + S^4]. Tout d'abord, le *But* de la situation énoncé en S^3 est mentionné en S^4 de manière implicite. De plus, ces deux unités son reliées par la présence du pronom élidé *l'* (***l'****attend sur les bords de la Propontide*) et qui fait référence à *Candide*.

Une représentation de la structure du discours est possible sous la forme d'un schéma arborescent où l'on intègre de manière successive les unités concernées (voir figure 10). Cette représentation se lit de la droite vers la gauche et de bas en haut, c'est-à-dire depuis les nœuds inférieurs jusqu'aux supérieurs. Néanmoins, sur le plan textuel, l'intégration d'une unité composite (accompagnée d'une autre qui en est dépendante) qui se trouve à droite de celle qui la précède, doit se faire en priorité. Ainsi, les processus d'intégration ne se font pas obligatoirement de façon verticale, c'est-à-dire d'une unité minimale jusqu'à celle traitée précédemment et qui se situe à sa gauche.

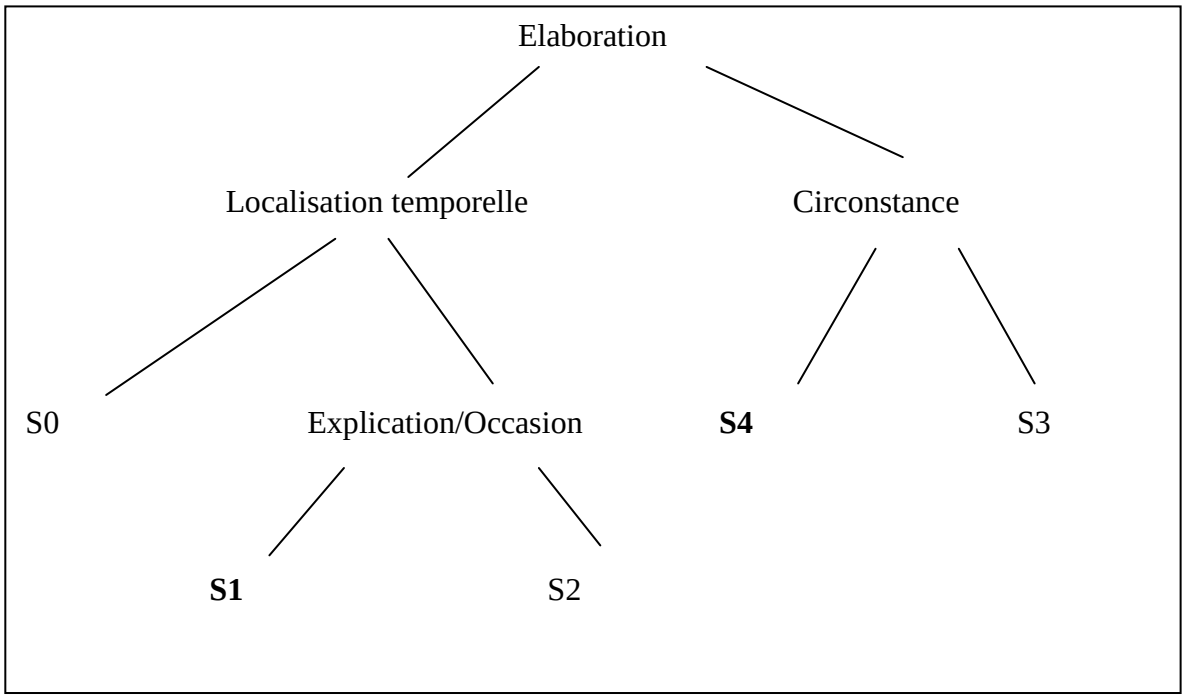


Figure 10 : Schéma, basé sur la modélisation Hobbs (1990), de de la structure[101] du discours [29]

La possibilité d'établir des relations entre les informations contenues dans le discours conditionne en grande partie la précision de la représentation mentale : en effet, les

[100] S^0 qui est *Au milieu d'un souper de carnaval.*
[101] Quand **S** est en gras dans une relation donnée alors elle représente l'unité dominante.

interprétants sont très sensibles à la cohérence du message. Nous avons observé que les moyens linguistiques assurant la cohésion et la connexité permettent d'éviter le recours à des procédures cognitivement coûteuses, comme certains traitements inférentiels. En effet, en indiquant les parties du discours à relier et le type de relation à établir, ils facilitent la détermination de la cohérence. En vu de permettre au récepteur d'établir les liens nécessaires pour élaborer une représentation mentale en accord avec celle du locuteur, les dispositifs de cohésion reposent sur l'utilisation par l'émetteur du discours de marques permettant d'organiser la structure informationnelle du texte. Pour cela et afin que son interlocuteur en établisse la cohérence à un moindre coût cognitif, sans surcharger le texte de marques inutiles ou redondantes, il faut que le locuteur possède la capacité de déterminer les liens qui doivent être explicitement marqués dans le discours. Pour identifier les informations dont le récepteur dispose, celles qui sont facilement inférables ou qui doivent être marquées, cette capacité exige d'examiner son point de vue.

Chapitre 2 : Facteurs de résolution de l'anaphore pronominale

Les deux dimensions textuelles de la relation anaphorique, étudiées dans le chapitre précédent, nous conduisent à étudier son traitement cognitif dans ce chapitre. Tout d'abord, l'anaphore joue un rôle crucial dans l'établissement de la cohésion. Ensuite, en tant que dispositif cohésif participant à la cohérence du discours, elle a un impact important sur la compréhension de ce dernier. L'objectif de ce chapitre consiste à vérifier dans quelle mesure la Théorie de l'Accessibilité[102] (Ariel 1990, 1996, 2001) permet de rendre compte du choix du bon antécédent d'une anaphore pronominale. Dans la première section, nous présenterons brièvement la Théorie de l'Accessibilité (Ariel 1990, 1996, 2001), les principaux avantages qu'elle présente ainsi qu'un certain nombre de critiques formulées à son encontre. La décision de l'appliquer sera justifiée. Dans la deuxième section, nous formulerons nos propres questions de recherche. Nous commenterons la méthodologie pour passer ensuite à la présentation et à la discussion des résultats. Les conclusions principales et les pistes de recherche qui en découlent seront exploitées dans les chapitres suivants.

1. Accessibilité et Saillance

1.1. La Théorie de l'Accessibilité

Pour faciliter l'identification des chaînes de référence, plusieurs modèles linguistiques (Givón 1983 ; Ariel 1990 ; Gundel *et al.* 1993) ont classé les expressions référentielles en fonction de leur forme et du degré d'accessibilité cognitive de leur référent. Afin de valider leur théorie de la *givenness hierarchy*[103], Gundel *et al.* (1993) ont utilisé un corpus anglais constitué de données orales et écrites présentant plusieurs degrés de formalité et de structuration. Ils distinguent six niveaux d'accessibilité correspondant à des expressions référentielles spécifiques en anglais, classés ci-dessous du plus accessible au moins accessible :

[102] Rappelons, avant d'aborder la théorie de l'accessibilité proposée par Ariel, que l'anaphore pronominale sera abordée comme une *marque linguistique de la cohérence d'un discours*. Dans ce sens, les pronoms anaphoriques, en tant que moyens cohésifs assurant la continuité référentielle, sont des marqueurs privilégiés permettant l'accès à l'entité référentielle la plus focalisée.

[103] Hiérarchie de donation, dans l'article en français de Gundel et al. (2000).

1. focus (« In focus ») : Le référent est représenté en mémoire de travail et l'attention est centrée sur cette entité : un pronom est le plus souvent utilisé (*it*). Pour un usage adéquat des pronoms personnels et des ellipses, ce statut est obligatoire.

2. activé (« Activated ») : Après avoir été extrait de la mémoire à long terme ou du contexte linguistique précédent, le référent est présent en mémoire à court terme. Ce statut permet l'utilisation du déterminant démonstratif (*that, this, this N*).

3. connu (« Familiar ») : Selon qu'il a ou non été mentionné récemment, le référent étant déjà représenté en mémoire à court terme ou à long terme, le destinataire est capable de l'identifier de façon unique. Ce statut permet l'utilisation du déterminant démonstratif (*that N*).

4. strictement identifiable (« Uniquely identifiable ») : A partir de la partie nominale de l'expression référentielle seule, le destinataire peut identifier le référent. Si le contenu référentiel véhiculé par le nom est par lui-même suffisant, cette identification peut être due, mais pas obligatoirement, à la présence d'une représentation déjà existante en mémoire. Ce statut permet l'utilisation de l'article défini (*the N*).

5. référentiel (« Referential ») : un ou des objets particuliers représentent la référence. Pour identifier la référence, le destinataire doit retrouver une représentation déjà existante de ce référent ou en construire une et ne peut pas se contenter d'accéder à une classe de référents. Ce statut permet l'utilisation du défini (*this N indéfini*) mais n'est toutefois pas identifiable dans certaines autres langues, on peut citer le chinois, l'espagnol et le russe. Les difficultés à traduire l'indéfini *this* s'explique par le fait qu'il ne semble pas non plus être marqué en français.

6. de type identifiable (« Type identifiable ») : l'expression référentielle décrit une représentation de classe d'objet accessible au destinataire. C'est un statut qui permet l'utilisation de l'article (*a N*).

L'exemple (1)[104] suivant illustre cette hiérarchie, où les différentes phrases (1a-f) représentent les différents statuts cognitifs exposés ci-dessus :

[1] *I couldn't sleep last night. (Je n'ai pas pu dormir hier soir)*

a. *It kept me awake. (Ça m'a gardé éveillé)*

b. *This train/this/that kept me awake. (Ce train/ceci/cela m'a gardé éveillé)*

[104] Emprunté à Gundel *et al.* (1993).

c. *That train kept me awake. (Ce train m'a gardé éveillé)*

d. *The train kept me awake. (Le train m'a gardé éveillé)*

e. *This train kept me awake. (Ce train m'a gardé éveillé)*

f. *A train kept me awake. (Un train m'a gardé éveillé)*

Si le statut de focus peut laisser place à toutes les autres formes référentielles, permettant l'apparition des formes pronominales comme *il* ou *elle* (« *it* » en anglais), c'est uniquement les formes de type *« un N »* qui peuvent marquer le statut de type identifiable. Ainsi, identifiés en anglais, ces statuts déterminent la forme des expressions référentielles qui peuvent apparaître dans le discours. Pour permettre toutefois au destinataire de restreindre le champ des référents envisageables et en indiquant le niveau d'accessibilité exact du référent, les locuteurs tendraient à employer la forme référentielle correspondant au statut le plus élevé. Cela correspond donc à la partie de la mémoire dans laquelle il faut effectuer la recherche. Indiquant dans quelle partie de la mémoire il est nécessaire de rechercher les référents correspondants, les différentes formes référentielles serviraient de signaux de traitement. Dans leur étude, Gundel *et al.* (1993) n'ont pas décrit spécifiquement leur méthodologie. En effet, ils n'ont pas donné d'explication quant à l'évaluation de la récence de la dernière mention et sur la spécificité des critères syntaxiques sélectionnés. D'autre part, il nous parait difficile de répliquer l'expérience sur un corpus de données en langue française. Par conséquent, dans notre étude quantitative, nous avons abandonné les hypothèses liées à la Hiérarchie du Donné au profit de celles formulées par la Théorie de l'Accessibilité.

Givon (1983, 1992), qui étudiait la relation entre niveau de topicalité et forme des occurrences référentielles en adoptant une démarche proche de celle de Gundel *et al.* (1993), a cherché des indicateurs de surface dans le but de pouvoir mesurer la topicalité des occurrences référentielles. Pour cela, à l'aide d'indicateurs quantifiables, il se propose de mesurer l'accessibilité référentielle et le statut thématique d'une occurrence référentielle. Trois indicateurs textuels servent à l'évaluation de l'accessibilité référentielle selon Givon :

1. La distance entre deux occurrences du même référent. Si le nombre de phrases séparant deux occurrences référentielles augmente, alors lors de la seconde occurrence, l'accessibilité du référent sera faible. C'est à partir de 20 phrases séparant deux occurrences que le niveau d'accessibilité atteindrait un minimum.

2. Les changements de référence. L'accessibilité d'un référent serait augmentée dans la proposition en cours de traitement par la présence d'un référent comme argument de la proposition précédente, son absence ferait diminuer cette accessibilité.

3. Les interférences référentielles potentielles. Lorsque les deux phrases précédentes contiennent des référents qui ont un sens proche, alors il y a interférence référentielle. C'est en fonction de la présence d'une interférence référentielle et du nombre de référents impliqués que l'accessibilité varierait. L'accessibilité des référents en concurrence serait diminuée par la présence d'une interférence et ce, d'autant plus qu'ils sont plus nombreux.

Givon fait une dissociation, qui selon nous est peu adaptée, entre les mesures de l'accessibilité référentielle et du statut thème. En effet, comme il le souligne, le statut thème est un des quatre facteurs qui conditionneraient l'accessibilité référentielle, une liaison existe donc entre les deux mesures. Ainsi, la mesure du statut thème devrait appartenir à l'ensemble des mesures qui permettent l'évaluation de l'accessibilité référentielle. On n'a pas directement accès à l'accessibilité puisqu'elle serait le corrélat cognitif de la continuité, on doit alors parler d'évaluation de l'accessibilité plus que de mesure (Givon 1992 : 7).

Au niveau local en particulier, l'importance thématique d'un référent semble plus difficile à évaluer. Selon Givon (1992), cette évaluation peut toutefois être basée sur deux indices textuels :

a. La fréquence globale : c'est la fréquence d'apparition du référent dans tout le discours.

b. La persistance topicale : qui correspond à la fréquence d'apparition du référent uniquement dans les dix dernières phrases.

L'importance thématique globale et locale, qui se combinent pour former l'importance thématique d'un référent à un moment précis du discours, sont évaluées par ces deux indicateurs. Combinant caractéristiques de l'accessibilité référentielle et importance thématique, ces mesures de la valeur de topicalité des référents ont été utilisées pour déterminer les relations entre le caractère plus ou moins topique des référents et leurs formes grammaticales d'apparition. La topicalité serait une propriété discontinue, ne possédant que quelques niveaux possibles (Givon, 1983) et non pas une propriété continue, possédant une infinité de valeurs comme ces mesures pourraient le laisser penser.

Ariel (1988, 1990), dans la même idée, émet l'hypothèse, connue sous le terme de théorie de l'accessibilité, qu'en dehors de leur signifié propre, les expressions référentielles

indiquent où et comment doit être recherché le référent correspondant. Selon cette théorie qui repose sur une conception coopérative de la production du discours, au moment de leur occurrence, les différents types d'expressions référentielles utilisés dans le discours seraient des marqueurs de l'accessibilité du référent. Pour qu'il puisse interpréter le discours dans les meilleures conditions, l'encodeur chercherait à fournir l'information nécessaire au décodeur. Ainsi, la facilité pour le décodeur à retrouver le référent correspondant serait fonction du contenu référentiel des expressions utilisées par l'encodeur. Plus le contenu référentiel propre à l'expression référentielle est faible et plus un référent serait accessible, donc facilement identifiable. Néanmoins, dans le choix des expressions, en plus de l'accessibilité référentielle, d'autres paramètres, comme la structure textuelle, peuvent intervenir. La théorie de l'accessibilité prévoit plutôt l'utilisation d'un pronom personnel, classiquement expression à fort contenu référentiel, pour renvoyer à un référent fortement accessible, tel que le personnage principal, en début de paragraphe par exemple.

La Théorie de l'Accessibilité, développée par Ariel (1988, 1990, 2001), a été inspirée de la notion d'accessibilité décrite par Givón (1983). Elle signale, au sein d'un texte, le degré d'accessibilité en mémoire des entités référentielles ; entités faisant référence à un élément précédemment cité dans le discours. Les critères proposés par Givón pour évaluer le degré d'accessibilité/de continuité topicale des référents correspondent aussi en partie aux facteurs commentés par Ariel (1988, 1990, 2001). Ils considèrent tous les deux le texte comme un produit de la communication, qui nous informe de façon indirecte sur les aspects cognitifs du processus communicatif. Plus concrètement, ces informations sont fournies par certains outils grammaticaux et par la position exacte des référents dans le discours (Givón, 1983 : 13). Givón (1983), puis Ariel (1990) ont ainsi mis au point une échelle des degrés d'accessibilité du référent qui définit son importance du référent (voir Figure 11).

Divers outils grammaticaux encodent, selon Givón, le degré d'accessibilité présenté par un référent. Ils nous informent donc (en tant qu'interlocuteurs et en tant que chercheurs) sur le statut cognitif du référent et par conséquent aussi sur la question de savoir si un référent est le topique du discours (c'est-à-dire le référent le plus accessible et le plus continu). Givón (1993) considère l'accessibilité des référents comme une notion graduée. Il propose plusieurs échelles de continuité topicale/d'accessibilité : à chaque fois les éléments grammaticaux sont rangés selon le degré d'accessibilité ou de continuité topicale qu'ils expriment (ou encodent) : les procédés grammaticaux codant l'accessibilité

référentielle peuvent être interprétés comme des instructions de traitement cognitif, cette hypothèse était formulée dans les travaux adoptant l'optique syntaxique (Givon 1983).

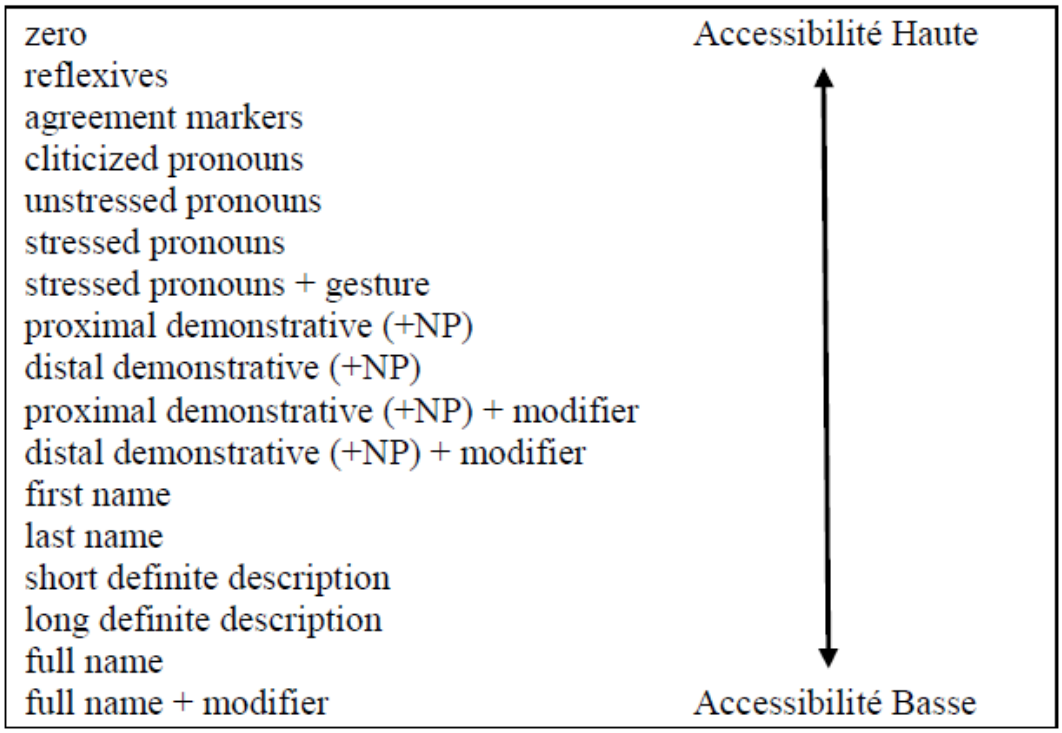


Figure 11 : Echelle d'accessibilité pour l'anglais selon Ariel (1990)

Un seul référent « topique[105] » pourrait être activé par proposition, cette limitation permettrait d'indiquer clairement à quelle partie de la représentation discursive doit être reliée l'information nouvelle. Pour identifier cette partie, des indices grammaticaux portés par ce référent permettraient de contraindre le type de recherche devant être opéré en mémoire. Ainsi, selon Ariel (2001 : 30), la théorie de l'accessibilité offre une analyse procédurale des expressions de référence, en marquant différents degrés d'accessibilité

[105] Le terme *topique* partage la même définition que *thème* et est souvent utilisé comme synonyme de thème. En effet, le topique a été défini dans un premier temps comme une information ancienne, présupposée, donnée dans une perspective propositionnelle. Puisqu'elle rend compte de la cohérence référentielle, la définition du topique a évolué avec l'apparition d'un intérêt pour les structures textuelles et le développement de la psycholinguistique (Givon 1992 : 6).
Selon l'auteur, la définition du topique doit prendre en compte le niveau *d'accessibilité référentielle* et *l'importance thématique.* En effet, dans cette acception du terme, un référent topique possèderait, par un nombre important d'apparition dans le discours de surface ou par son caractère central dans la représentation discursive, une action cohésive.
Ariel met l'accessibilité sur le même pied d'égalité que la topicalité « *Saliency: The antecedent being a salient referent, mainly whether it is a topic or a non-topic.* » (Ariel 1990 : 29).

mentale. L'idée de base est que les expressions de référence indiquent au lecteur de récupérer une certaine information de sa mémoire en indiquant à quel point cette information lui est accessible à lui à cette étape du discours.

Sur cette échelle, le système des pronoms varie selon la langue alors que l'ordre prédit est invariant : en ce sens, elle est qualifiée d'arbitraire par Ariel. Ce qui n'est pas arbitraire en revanche c'est la relation entre les différentes formes référentielles et le degré d'accessibilité auquel elles sont liées. Cette relation est basée sur trois principes d'iconicité qui se chevauchent partiellement :

- l'informativité : ce principe fait référence au contenu informationnel de l'expression référentielle par rapport à l'antécédent : une entité moins accessible aura un référent offrant beaucoup d'informations (lexicales et conceptuelles) ;

- l'atténuation : ce principe fait référence à la longueur de l'expression référentielle, plus celle- ci est brève et plus elle désignera une entité accessible ;

- la rigidité : dans un discours qui peut être ambigu, le principe de rigidité permet de souligner l'univocité de l'expression référentielle par l'emploi de référents tels que des noms propres.

La combinaison de ces trois paramètres peut ainsi nous aider à prédire l'accessibilité d'une forme. Un référent accessible aura une expression informative, longue et rigide et *vice et versa*. De plus, les formes les moins informatives, les plus ambiguës et les plus courtes (formes zéros, pronoms atones…) coderont le plus haut degré d'accessibilité et *vice et versa*. Par exemple, l'expression "*L'héroïne de la légende de Thèbes*" est longue et informative et fait référence à une entité de faible accessibilité. Une entité qui ferait référence à une accessibilité moyenne serait plutôt "*Antigone*". Concernant la rigidité, le prénom et le nom propre, prenant le rôle d'étiquettes, seraient plus rigides que le prénom seul par rapport à une entité supposée unique. En ce sens, les expressions référentielles utilisées dans un discours sont reflétées dans la théorie de l'accessibilité et ces trois principes donnent l'opportunité au locuteur de choisir l'expression référentielle appropriée. Néanmoins, ils ne sont pas assez informatifs concernant la forme à sélectionner dans un contexte précis puisque c'est le degré d'accessibilité des référents qui détermine ce choix.

A partir de cette hiérarchie, nous pouvons distinguer trois types de référents :

- Les référents de faible accessibilité : Ils sont dit inactifs, car même s'ils sont présents, une autre entité a rempli au préalable leur fonction.
  Exemple : *Ismène, la sœur d'Antigone.*

- Les référents d'accessibilité moyenne : Ce sont des référents activés mais pas forcément dans le focus. Exemple: *"Le père Grandet....cet avare"*.
- Les référents à forte accessibilité : Ce sont des référents saillants dans le discours, pour y référer, un pronom suffit. Exemple: "*Antigone.....elle"*.

Le degré d'activation du référent déterminera le choix de l'entité référentielle (Ariel 1990), ceci de façon bidirectionnelle. Ainsi, un marqueur à forte accessibilité sera choisi si la représentation mentale du référent est très active ; plus la représentation mentale du référent est faible et plus le référent sera de faible accessibilité (un nom propre par exemple). De ce fait, quand le locuteur sélectionne une expression référentielle, il le fait en fonction du niveau de la représentation mentale du référent chez l'interlocuteur. De la même façon, l'interlocuteur recherchera dans sa mémoire l'entité à laquelle l'expression réfère (Ariel 1996).

Dans la théorie de l'Accessibilité, quatre paramètres linguistiques conditionnent le choix du degré d'accessibilité du référent (Ariel 1990, 1996, 2001) :

1. La distance entre l'expression référentielle et son antécédent : le référent le plus accessible est celui qui est le plus proche de son antécédent. En effet, plus il est proche et plus il a de chances de se trouver dans le stock de mémoire à court terme. Asher (2011) décrivent cette distance en termes "*d'unités de discours élémentaires*". Dans leurs recherches, ils sont en accord avec la théorie d'Ariel sur la distribution des expressions référentielles : les expressions nominales sont plus éloignées de leurs antécédents que les expressions anaphoriques et démonstratives ; ceci est d'autant plus vrai si les expressions nominales sont réduites à un prénom ou un nom propre. L'auteur spécifie par ailleurs que pour mesurer la distance, le nombre de mots n'est pas suffisant et qu'il faut prendre en compte les paragraphes et les parties qui augmentent la complexité de l'accessibilité.
2. Le nombre de candidats au poste d'antécédent (s'il y a compétition ou non) : l'accessibilité est d'autant plus grande que le nombre de référents potentiels est réduit, c'est-à-dire que dans le cas d'une compétition référentielle le contexte du discours devient ambigu. Ainsi, un discours ne contenant que très peu de référents sera plus facile à interpréter par l'interlocuteur. Précisons que dans le cas d'une compétition référentielle, pour éviter l'ambigüité référentielle, il est souhaitable d'utiliser un nom propre plutôt qu'un simple pronom.

3. La saillance : La saillance est en corrélation directe avec l'accessibilité. Plus le référent est saillant et plus il est accessible. Ariel met l'accessibilité au même niveau que la topicalité[106]. Ainsi, plus le référent se retrouve en position de sujet, plus il est accessible. D'après Ariel (1990 : 28-29), dans les discours naturels, les thèmes principaux constituent le plus souvent les entités les plus marquantes. Il semble que les sujets occupent une position privilégiée dans la mémoire. Ainsi les entités les plus saillantes seraient d'après elle les thèmes principaux des discours. Par exemple, dans un récit, le lecteur retiendra surtout les personnages principaux et ceux qui sont introduits par leur noms/prénoms.
4. L'unité textuelle : Selon Ariel (1990 : 28-29), la notion d'unité textuelle se définit comme « l'antécédent étant dans le même cadre / point de vue / segment ou paragraphe que l'anaphore ». Si l'antécédent et le référent se retrouvent dans le même segment que l'anaphore (le même cadre spatial, temporel ou le même point de vue), alors le référent sera plus accessible. Ce sont les référents de type pronoms qui seront le choix préférentiel pour des éléments contenus dans le même segment discursif. Par exemple dans :

   [2] Le matin, Meursault se confie à un gendarme et lui avoue l'intérêt qu'**il** (Meursault) éprouve à assister à un procès. (Résumé L'étranger)

   l'antécédent *Meursault* est très accessible car d'une part, il est très proche de l'anaphore et d'autre part, il est dans le thème de la phrase. Dans cet exemple, le pronom *il* constitue la reprise la plus appropriée. Pour des propositions coordonnées, l'utilisation de noms propre serait plus appropriée comme dans l'exemple suivant :

   [3] **Meursault** est arrêté et subit plusieurs interrogatoires au commissariat, puis chez le juge d'instruction. Trouvant son affaire " très simple", **Meursault** ne juge pas utile de prendre un avocat. On lui en désigne un d'office. Il (un avocat) questionne Meursault sur sa mère et les sentiments qu'il (Meursault) avait pour elle (sa mère). Les propos à la fois sincères et naïfs de Meursault gênent son avocat. (Résumé L'étranger)

Dans d'autres théories, les facteurs distance, saillance, unité et compétition sont étudiés indépendamment les uns des autres et la théorie de l'accessibilité est originale en ce sens qu'elle combine tous ces facteurs pour fournir une représentation plus intuitive sur le plan cognitif. Cela étant dit, elle ne s'oppose pas à d'autres études antérieures en sciences cognitives, en linguistique ou en psycholinguistique qui démontrent que la saillance joue

[106] "Topical entities will be more salient than non-topics".

un rôle, ainsi que l'unité ou la compétition et constitue de ce fait un bon cadre pour la détermination de facteurs aidant au choix des expressions référentielles. Toutefois, elle peut engendrer des problèmes méthodologiques dans le sens où, afin de tester les hypothèses formulées, il est difficile d'opérationnaliser les facteurs d'accessibilité puisque certains, comme la distance, sont difficilement quantifiables (Ariel 1990 : 31). Ces notions montrent que l'accessibilité est affectée par la relation relativement étroite entre l'antécédent et le référent.

Malgré l'universalité de cette théorie, selon Kleiber (1990) elle ne prend pas assez en compte le contenu référentiel des expressions référentielles et pour se soustraire à ce type de raisonnement circulaire, l'auteur suggère de considérer à la fois le sens expressions référentielles et la manière dont le référent est exposé par le locuteur. Par exemple, dans :

> [4] **Meursault** est arrêté et subit plusieurs interrogatoires au commissariat, puis chez le juge d'instruction. Trouvant son affaire " très simple", **Meursault** ne juge pas utile de prendre un avocat. (Résumé L'étranger)

au lieu d'utiliser *celui-ci* ou encore *ce dernier*, l'auteur a délibérément utilisé le nom propre *Meursault*.

D'autres auteurs, comme Reboul et *al.* (1997), critiquent plus ouvertement cette théorie qu'ils qualifient même d'inutile, simpliste et réductrice et représentant une « réintroduction de la notion de cohésion ». Selon eux, la référence ne serait réduite qu'à un phénomène purement linguistique à travers l'échelle donnée. Enfin, en analysant le facteur distance, des auteurs tels que Maes/Noordman (1995) n'ont pas retrouvé les mêmes résultats qu'Ariel respectivement pour le néerlandais et le français. Malgré tout, afin d'évaluer l'impact des quatre facteurs, Toole a mis au point en 1996 des méthodes clairement exposées que nous utiliserons dans la présente étude.

En conclusion, cette théorie, malgré le fait qu'elle n'exprime qu'une propriété graduelle, sera utilisée pour notre calcul automatique et nous utiliserons comme référence le classement hiérarchique des expressions ainsi que le parallélisme syntaxique.

1.2. Saillance

Dans les recherches en linguistique, la notion de saillance, qui met en avant certains éléments du message linguistique, est de plus en plus prise en compte :

> La saillance, c'est avant tout l'émergence d'une forme sur un fond, la mise en avant d'une entité par rapport à d'autres entités. Etre saillant, c'est ressortir particulièrement, au point de capter l'attention et de donner une accroche, un point de départ à la compréhension. (Landragin 2012).

La saillance est appelée "*pop-up*" en anglais et ce terme rend compte de l'aspect communicatif caractéristique de tous les types de discours et ce, quelle que soit la langue. Ainsi, le phénomène de saillance s'ajoute aux notions linguistiques établies, comme le topique par exemple, pour interpréter le discours. Néanmoins, le phénomène n'est pas complètement déchiffré et il est essentiel de bien le définir, d'en établir la portée en linguistique et d'établir une méthodologie adaptée à son étude.

La saillance ne se définit pas comme un concept linguistique en soi et elle ferait partie des facteurs qui définiraient plutôt le topique (Ariel 1990). Ainsi, afin de mettre en saillance un élément du discours, il faut utiliser des indicateurs syntaxiques ou prosodiques. Par la suite, la saillance a été plutôt définie comme une notion incluant elle-même plusieurs facteurs (Schnedecker 2009, Landragin 2012). Nous en citerons quelques un dans ce qui suit.

Tout d'abord, ceux de types syntaxiques où l'on regroupe plusieurs entités afin de mettre en saillance un élément du discours (soit en tête, soit en fin de phrase par exemple). Ensuite, l'ordre (la position, la symétrie) et la fréquence (les répétitions) des mots représentent un deuxième facteur pouvant définir la saillance. Un troisième facteur serait la fonction grammaticale avec une hiérarchie où le sujet se trouve devant le complément d'objet direct, lui-même devant le complément d'objet indirect. Il y a d'autres facteurs, incluant la totalité des niveaux d'interprétation et d'analyse ainsi que les dimensions du langage, qui définissent la saillance comme les facteurs de types lexical et sémantique (par exemple : la catégorie d'un mot, la sémantique de la phrase et le thème ou le topique selon le contexte du discours.). Cette multiplicité fait alors de la saillance un phénomène multifactoriel.

Apothéloz (1995) soutient que le choix des formes anaphoriques dépend des propriétés de saillance locale et cognitive de leurs référents. L'auteur affirme que la saillance refléterait l'accessibilité des référents en mémoire. En effet, elle est une caractéristique des éléments « donnés » ou « introduits dans l'univers du discours » (p. 45). Ces élément sont soit introduits explicitement, soit par leur caractère saillant dans la situation d'énonciation, ils sont ainsi partagés par les interlocuteurs. La saillance locale est :

> un paramètre dont la valeur est purement contingente, non dépendante du sens qui se construit : un objet est saillant localement en raison d'un fait accidentel, soit qu'il vient d'être évoqué par des moyens verbaux ou non verbaux ; soit que, dans la situation, il se signale à l'attention des interlocuteurs par ses propriétés perceptives. (Apothéloz 1995 : 315).

La saillance cognitive est au contraire étroitement dépendante de la représentation textuelle en construction. Une propriété qui rend ces référents saillants même dans des segments du discours où ils ne sont pas explicitement évoqués est le fait que les éléments centraux de cette représentation seront des référents cognitivement saillants tant que leur statut d'élément central de la représentation se maintient, citons par exemple le personnage principal dans un résumé. Dans ses travaux, Apothéloz va aussi dans le sens d'une accessibilité avec un faible nombre d'états possibles. Il postule, au vu de l'indépendance entre la saillance locale et cognitive, que trois degrés d'accessibilité contrastés pour les référents sont le résultat de la combinaison entre les propriétés de saillance locale et cognitive (voir Tableau 9).

	Degré d'accessibilité			
	Degré I	Degré II		Degré III
Saillance locale	+	+	-	-
Saillance cognitive	+	-	+	-

Tableau 10 : Degrés d'accessibilité des objets en fonction de leur saillance locale et cognitive (d'après Apothéloz 1995 : 316)

Selon Apothéloz (1995), ce sont les référents qui cumulent à la fois saillance locale et cognitive qui sont les plus accessibles. On attribue un niveau d'accessibilité intermédiaire aux référents saillants uniquement au niveau local ou au niveau cognitif. Selon l'auteur, une des formes de saillance pourrait avoir un impact plus important que l'autre sur le niveau d'accessibilité puisque les trois degrés proposés découlent uniquement d'une combinaison logique des deux facteurs. Dans ce cas, il serait nécessaire d'envisager deux niveaux d'accessibilité intermédiaire. Pour finir, les référents qui ne sont saillants ni localement, ni cognitivement, sont les moins accessibles. Toutefois, le fait que les degrés I et II pourraient conduire à un même niveau d'accessibilité ne semble pas être envisagé par l'auteur. En effet, l'augmentation de l'accessibilité référentielle n'est pas apportée par le cumul de deux facteurs et les degrés d'accessibilité I et II seraient confondus, même si un seul facteur de saillance suffit pour atteindre l'accessibilité maximale. Un référent s'il présente au moins un facteur de saillance est accessible, ou alors, non-accessible, s'il n'en présente aucun et l'accessibilité apparaitrait comme une propriété dichotomique.

Une position plus restrictive est adoptée par Kleiber afin de définir les facteurs qui rendent une entité discursive saillante. Selon l'auteur :

> la saillance n'est accordée à une entité que par une situation, un événement, un état dans lequel se trouve impliquée cette entité. (Kleiber 1994 : 122)

et un référent ne peut pas être saillant s'il est simplement introduit dans le focus par l'intermédiaire du texte ou de la situation d'énonciation. Ainsi pour être saillante, une entité doit être impliquée de manière active dans la situation, en plus d'être introduite implicitement ou explicitement dans la situation décrite par le texte. Par exemple, dans le résumé de *L'étranger*, aussi longtemps qu'ils n'interviendront pas dans le récit, *les amis de la mère de Meursault*, ne feront pas partie des référents saillants. Alors que leur présence, comme celle du *concierge*, est un élément pouvant être activé par l'évocation de la mort de la mère.

> [5] Meursault, le narrateur, est un jeune et modeste employé de bureau habitant Alger. Le récit commence le jour de la mort de sa mère. Au petit matin, il (Meursault, le narrateur) reçoit un télégramme de l'asile de vieillards de Marengo, situé à quatre-vingt kilomètres d'Alger lui annonçant son décès. Elle (sa mère) y séjournait depuis trois ans. Meursault demande et obtient un congé de quarante huit heures et va déjeuner chez Céleste, un restaurant où il (Meursault) a l'habitude d'aller. Vers deux heures de l'après-midi, il (Meursault) prend l'autobus. Il fait chaud, Meursault dort pendant presque tout le voyage. L'asile étant à deux kilomètres du village, Meursault termine le trajet à pied. Après les formalités, il (Meursault) a une entrevue avec le directeur de l'asile, qu'**il** (**Meursault mais ça pourrait être aussi le directeur de l'asile**) écoute d'une oreille distraite. Ce dernier **(ce pronom vient pour désambiguïser le précédent pronom personnel il)** lui indique que sa mère n'était pas malheureuse à l'asile. Il (le directeur de l'asile) lui annonce également que l'enterrement religieux est fixé au lendemain matin. Puis Meursault se rend dans une salle blanchie à la chaux où se trouve entreposé le corps de sa mère mais il (Meursault) refuse de voir le corps. Il (Meursault) a une conversation avec le concierge. Cet homme bavard lui raconte sa vie et lui propose de dîner au réfectoire. Meursault, décline l'invitation. Le concierge lui offre alors un café au lait que Meursault accepte. Puis a lieu la veillée, interminable : **les amis de sa mère**, tous semblables, y assistent. **Ils (les amis de sa mère)** s'installent autour du cercueil et laissent échapper des bruits bizarres de leurs bouches édentées. Une vieille femme pleure sans cesse. Meursault a la désagréable impression que ces vieillards sont là pour le juger. Le jour se lève. Meursault admire la beauté de ce nouveau matin. Après une toilette rapide et un nouveau café au lait que lui a préparé le concierge, le narrateur se rend chez le directeur où il(le narrateur/Meursault) accomplit de nouvelles formalités administratives. Puis le cortège funèbre se rend vers l'église du village, située à trois quarts d'heure de marche. Un vieillard suit péniblement le cortège, il s'agit de Thomas Pérez, un compagnon d'asile de la mère de Meursault. les voisins se moquaient d'eux en les appelants "les fiancés". La chaleur est insoutenable. L'enterrement défile comme un songe dans l'esprit de Meursault : l'église, le cimetière, l'évanouissement du vieux Pérez, l'attente, puis la joie quand l'autobus le ramène enfin à Alger. Meursault a enterré sa mère sans larmes et n'a pas voulu simuler un chagrin qu'il (Meursault) n'éprouvait pas. (Résumé L'étranger)

Selon Kleiber, la saillance serait fonction du degré d'implication des entités référentielles dans cette situation ; en effet, c'est la situation décrite qui conditionne l'état de saillance des référents. Kleiber mettrait en relation la saillance d'un référent avec son accessibilité, son caractère central dans la représentation du discours.

La question de la relation entre saillance et accessibilité se pose alors. Le facteur thématique d'accessibilité évoqué par Ariel semble correspondre à la saillance cognitive telle qu'elle est définie par Apothéloz (1995). En effet, quand on examine les facteurs expliquant les différents niveaux d'accessibilité ou de saillance, la proximité entre ces deux notions apparaît clairement. Ces deux notions linguistiques semblent rendre compte d'un même phénomène cognitif : l'accessibilité référentielle, elles ont néanmoins été développées en parallèle et par deux champs de recherche distincts. La différence est que la notion de saillance est plus utilisée lorsque les choix lexicaux induits par l'accessibilité des référents sont mis en avant, alors que la notion d'accessibilité est plutôt employée dans l'étude de l'impact de l'accessibilité référentielle sur la structure syntaxique.

La saillance n'étant pas le seul facteur de résolution de l'anaphore, nous allons détailler dans la section suivante les autres facteurs intervenants lors de ce processus.

2. Comment résoudre une relation anaphorique ?

Les recherches consacrées à l'étude de l'acquisition des possibilités de traitement du système anaphorique ont le plus souvent utilisé des anaphores de type pronominal, en particulier des pronoms personnels *(il, elle, le, la)*. Parmi les procédures les plus souvent citées, nous pouvons retenir les procédures *d'analyse morphologique*, la procédure *des fonctions syntaxiques parallèles*, la procédure *de distance minimale*, la procédure *du sujet* ou la procédure *thématique*, ainsi que *pragmatique et sémantique.*

2.1. La procédure d'analyse morphologique

Cette procédure fait l'unanimité des auteurs, comme Kleiber (1994a), Amsili *et al.* (2002), Gasperin *et al.* (2007), Schenedecker (2015) et autres, qui abordent le sujet de la résolution de l'anaphore. Avec des anaphores morphologiquement non-ambigües, les lecteurs sont capables de fonder leurs interprétations sur l'analyse des indices morphologiques de genre, puis de nombre pour déterminer l'antécédent d'un pronom. En lecture, le genre et le nombre sont utilisés pour déterminer le référent des pronoms sujets, mais l'analyse qui en découle n'est pas toujours parfaite. Ces procédures utilisent l'analyse des données sémantiques du texte pour rétablir la référence du pronom lorsque les procédures d'analyse morphosyntaxique ne permettent pas de déterminer l'antécédent correct. Examinons l'exemple suivant :

> [6] L'asile étant à deux kilomètres du village, Meursault termine le trajet à pied. Après les formalités, il (Meursault) a une entrevue avec le directeur de l'asile, qu'**il** (**Meursault mais ça pourrait être aussi le directeur de l'asile**) écoute d'une oreille distraite. Ce dernier **(ce**

pronom vient pour désambiguïser le précédent pronom personnel il) lui indique que sa mère n'était pas malheureuse à l'asile. (Résumé L'étranger)

La référence du pronom *il* ne peut être attribuée qu'après avoir interprété le sens de la proposition suivante. A partir des indices sémantiques de la proposition *" Ce dernier lui indique que sa mère n'était pas malheureuse à l'asile"*, on peut attribuer le pronom *il* à *Meursault.* Nous remarquons que dans un contexte référentiel entraînant une indétermination morphologique, la possibilité d'utiliser les informations sémantiques pour interpréter les pronoms apparaît comme évidente. Cela dit, une prise en compte correcte des informations sémantiques pourrait permettre de déterminer le bon antécédent d'un pronom ambigu.

2.2. La procédure des fonctions syntaxiques parallèles

La procédure des fonctions parallèles ou de parallélisme syntaxique consiste à attribuer au SN relativisé la même fonction que celle de son antécédent. Pour Masseron/Schnedecker (1988) cette procédure est basée sur le fait que *"il"* est un pronom de reprise neutre, ce qui signifie qu'il anaphorise un groupe qui occupait la fonction syntaxique qui est la sienne. Toutefois, à l'utilisation des critères syntaxiques de sujet et d'objet, certains préfèrent substituer un critère de non-changement de rôle qui fonctionnerait selon le principe d'identification de la coréférence entre un SN et un pronom sur la base de leur identité de rôle dans l'action. En effet, ce critère permet de rendre compte du fait que le coréférent du pronom est préférentiellement le sujet syntaxique de la proposition précédente, quelle que soit la fonction du pronom.

2.3. La procédure thématique

Pour Reichler-Beguelin (1989), les procédures de distance minimale et du sujet sont des procédures d'attribution de la référence au thème. La procédure du sujet peut être interprétée comme une procédure de maintien du thème, ce qui correspond à l'utilisation d'un schéma à sujet unique pour piloter l'analyse de la structure référentielle du discours. Pour Reichler-Beguelin, la procédure thématique repose donc sur l'attribution de la référence des pronoms au référent thématique précédent. C'est donc le thème local qui intervient dans l'interprétation du pronom. La procédure de proximité peut être analysée comme une procédure de changement de thème, ce qui correspond à l'utilisation d'un schéma à thème linéaire.

Kinstch/van Dijk (1983 : 170) parlent quant à eux de procédure de cotopicalité pour l'interprétation des pronoms en première position ("*first position pronoun*"), en général les

pronoms sujets. Selon eux, les interprétants ont tendance à faire coréférer les pronoms de première position avec un référent topique « car la continuité topicale serait un comportement stéréotypique en production du discours » (Kinstch/van Dijk 1983 : 170). En fait, pour ces auteurs, la procédure de cotopicalité repose, dans un premier temps, sur la recherche d'un référent topique dans la représentation sémantique du discours ; l'interprétation découlant du choix du référent topique serait confirmée si ce référent cumule sa topicalité avec d'autres propriétés sémantiques et / ou syntaxiques. Les auteurs proposent un classement (dans l'ordre décroissant) des propriétés favorisant l'attribution de la référence des pronoms de première position à un référent qui :

- est un syntagme nominal topique ;
- a la propriété sémantique d'agent ;
- occupe la fonction de sujet syntaxique ;
- apparaît en première position dans la proposition ;
- appartient à une proposition principale.

Pour Kinstch/van Dijk (1983), ce sont donc des informations sur la structure globale du discours (le référent est un référent topique dans la représentation du discours) et sur sa structure locale qui interviennent dans l'interprétation des pronoms sujets. Les auteurs (1983 : 171) précisent que cette procédure ne fait qu'établir une cohérence partielle, c'est-à-dire, un lien cohésif provisoire. L'interprétation définitive du pronom étant déterminée à partir de l'interprétation de l'ensemble de la proposition ou de la phrase et de sa cohérence avec les parties déjà traitées du discours. Lorsqu'ils acquièrent la capacité de réviser une interprétation provisoire, les lecteurs se fondent sur de nouvelles propriétés pour établir l'interprétation définitive. Dans un premier temps, les propriétés morphologiques associées aux référents sont utilisées pour modifier l'interprétation des pronoms. C'est plus tard que les propriétés sémantiques et pragmatiques du discours pourront intervenir dans la détermination de l'interprétation définitive en cas d’ambiguïté.

2.4. Procédure de distance minimale

Proposée par Rosenbaum (1965), Ariel (1990 : 29), Mitkov (2002) et expérimentée par Dimitrov et al. (2005), Demol (2007), Pironneau *et al.* (2014) etc, cette procédure est basée sur deux facteurs :

- un facteur spatial qui consiste à prendre comme référent du pronom le syntagme le plus proche. Quand l'interprétation doit se fonder sur les informations sémantiques présentes dans l'énoncé et que les référents

potentiels sont introduits sous la forme de noms propres, ces paramètres pourraient, lorsqu'ils sont réunis, renforcer le recours à l'interprétation par proximité spatiale.

- facteur cognitif : quand la procédure de proximité est dénommée procédure de récence. Dans ce cas ce n'est pas le facteur spatial qui est évoqué pour justifier le choix du référent du pronom, mais, sa présence en mémoire à court terme due à sa récence de mention.

Ariel (1990 : 20) a opté pour mesurer la distance en se basant sur les limites de la phrase et du paragraphe parce que :

> les limites des unités textuelles sont des facteurs importants en eux-mêmes, en plus de leur indication de distance. La portée de la mémoire est étroitement liée aux unités textuelles. Évidemment, au moins pour certaines unités, le matériel est libéré de la mémoire à court terme[107].

2.5. La procédure pragmatique

Cette procédure utilise les connaissances extralinguistiques du sujet qui sont mobilisées pendant la compréhension, pour déterminer l'antécédent adéquat du pronom. L'attribution de l'antécédent se fait sur la base de certaines caractéristiques des situations décrites par le texte. Elles sont utilisées lorsque les indices morphosyntaxiques et sémantiques ne suffisent pas à déterminer un antécédent unique. Reboul (1991) cite un exemple de Mehler/Dupoux (1987) qui illustre le phénomène :

> [7] Après avoir considéré son dossier, le directeur limogea l'ouvrier, parce qu'**il** était un communiste convaincu.

On a affaire à un pronom morphosyntaxiquement ambigu, dont l'attribution référentielle est différente suivant que l'on se trouve dans un pays communiste ou capitaliste. Ici ce n'est pas le sens du texte lui-même qui est pris en compte, mais une interprétation dépendant des connaissances générales propres à un individu. Les procédures sémantiques et pragmatiques sont souvent regroupées pour rendre compte de la coopération entre les données sémantiques et pragmatiques dans la détermination de la cohérence, et de la difficulté à les différencier nettement dans le discours :

[107] Texte d'origine "[…] textual unit boundaries are important factors in themselves, in addition to their indication of Distance. Memory scope is crucially interrelated with textual units. Obviously, at least at some unit closures material is released from short-term memory."

[8] Antigone rentre chez **elle**, à l'aube, après une escapade nocturne. **Elle** est surprise par sa nourrice qui **lui** adresse des reproches. (Résumé Antigone)

Dans l'exemple [8], la référence du pronom *lui* s'effectue aussi bien à partir du contenu sémantique des phrases que de nos connaissances générales sur les motivations des individus. Le sens des phrases nous permet de déterminer que Antigone (A) rentre tard, surprise par la nourrice (B), que (B) adresse des reproches à (A). Nos connaissances nous permettent de comprendre qu'un individu qui rentre tard peut être réprimandé par sa nourrice et non le contraire. Ce n'est que lorsque l'on cherche à établir la cohérence de l'ensemble des propositions que l'on va utiliser ces différentes informations pour établir des relations logiques entre les propositions qui vont permettre d'identifier le bon antécédent. Les stratégies sémantiques et pragmatiques peuvent donc être unifiées au sein d'une stratégie d'extraction de la cohérence textuelle.

L'hypothèse d'un contrôle final de l'interprétation à l'aide des procédures pragmatiques et sémantiques proposée par Kinstch/van Dijk s'articule parfaitement avec le modèle minimaliste de l'interprétation. En effet, dans ce modèle, lorsque le processus d'interprétation automatique n'aboutit pas à la sélection d'un référent unique, le choix du référent peut, en dernier ressort, s'effectuer sur la base de la cohérence de la représentation textuelle. Ce sont donc essentiellement des informations sémantiques et pragmatiques qui conditionnent l'interprétation finale lorsque le processus automatique n'aboutit pas et que le lecteur a pour but d'identifier précisément le référent. De même, on peut envisager les procédures de récence, du sujet ou de coréférence au thème comme des interprétations du processus automatique d'appariement entre l'anaphorique et sa référence, sur la base de l'accessibilité relative des référents. En effet, la récence de mention, la fonction syntaxique et le statut thématique conditionnent le niveau d'accessibilité des référents (*Cf.* section précédente). L'utilisation de ces facteurs, lors de l'interprétation des anaphores, semble être la manifestation d'un processus d'appariement avec le référent le plus accessible. Le passage d'une procédure à l'autre pourrait correspondre à une modification des critères pris en compte pour déterminer l'accessibilité référentielle. Ces critères pourraient évoluer en fonction des capacités de la mémoire de travail. Il semble en fait que l'on assiste à l'émergence d'une véritable stratégie d'interprétation associant des procédures automatiques d'extraction de la référence à des procédures de confirmation et de rectification dont l'emploi est modulé en fonction du contexte linguistique.

2.6. La procédure métacognitive

Le terme de métacognition est apparu dans les années 70 avec les travaux de Flavell (1976) à propos de la *cognition sur la cognition.* L'intérêt pour cette notion s'est considérablement développé depuis. Selon la définition originelle proposée par Flavell (1976 : 232), la métacognition renvoie à « la connaissance qu'un individu a de ses propres processus et productions cognitives et de tout ce qui peut être en relation avec ses processus et productions cognitives ». Flavell précise à ce propos que le contrôle, la régulation et la planification de ces processus rentrent aussi dans le cadre de la métacognition.

La métacognition renvoie à deux axes de recherche distincts mais complémentaires (Ehrlich 1994). Il s'agit pour le premier de l'étude des connaissances métacognitives et pour le second de la recherche sur le contrôle exercé par les sujets sur leurs propres processus cognitifs. Le premier vise à étudier ce que les sujets savent sur les structures et les processus cognitifs impliqués dans une activité particulière. A partir de questionnaires ou en faisant commenter l'exécution d'une tâche par le sujet qui l'effectue, les expérimentateurs cherchent à déterminer les connaissances métacognitives que possède un sujet et qu'il est capable de verbaliser. Toutefois, toutes les connaissances métacognitives ne seraient pas ou pas clairement conscientes, ni verbalisables. C'est du moins ce que semble affirmer Flavell (1985 : 35) lorsqu'elle dit que certaines expériences métacognitives sont « moins nettement conscientes et moins bien verbalisables ». Cependant, la nature consciente et verbalisable d'un processus cognitif est souvent conçue comme l'indicateur que l'on a bien affaire à un processus contrôlé de façon métacognitive. Le deuxième axe de recherche renvoie au contrôle que les sujets exercent sur leur propre fonctionnement cognitif. Selon Ehrlich (1994), si de nombreuses études « rangent sous le terme métacognition tout ce qui concerne les mécanismes de contrôle et de régulation », ces études ne relèvent que rarement d'une approche métacognitive. Pour l'auteur, il faudrait réserver ce qualificatif aux seules études portant sur les mécanismes de gestion « dépendant des connaissances métacognitives et des expériences métacognitives ». C'est-à-dire que, pour qu'il y ait contrôle métacognitif, il faut non seulement que le sujet exerce un contrôle sur son fonctionnement cognitif, mais qu'en plus, le processus contrôlé implique des connaissances ou des expériences métacognitives.

La métacognition est une question qui ne trouve pas de réponse explicite dans les travaux portant sur les stratégies d'interprétation des pronoms anaphoriques. On peut

s'interroger sur le processus de passage d'une procédure à l'autre : pour le passage de la procédure de proximité, ou de récence, à la procédure du sujet, certaines hypothèses ont été évoquées. Il s'agit du changement de structure thématique adoptée pour l'analyse du texte ou de l'augmentation des ressources mnémoniques attribuées au maintien des référents grâce, notamment, à la baisse des ressources devant être allouées aux procédures de traitement de 1'écrit. En revanche, le passage de ces procédures aux procédures sémantiques et pragmatiques n'est que rarement justifié. Cette évolution est parfois mise en perspective avec l'apparition de capacités métalinguistiques.

En effet, pour éprouver la nécessité d'utiliser les informations sémantiques en vue d'interpréter un pronom au lieu de se contenter de l'application de la stratégie thématique, il faut se rendre compte que l'on fait une interprétation erronée. Cela est possible uniquement si on peut contrôler la cohérence de l'interprétation obtenue avec la représentation du texte déjà construite. Or, cette capacité à contrôler le résultat de ses propres processus cognitifs fait partie des capacités d'ordre métacognitif, de même que la capacité à modifier le résultat d'un processus cognitif comme, par exemple, rectifier l'interprétation erronée d'un pronom.

Synthèse

Dans certains cas, l'identification du référent ciblé par une expression anaphorique, plus précisément un pronom, est assez difficile. Une anaphore pourrait être ambiguë au point que même un humain ne puisse la résoudre. Il faudrait alors pouvoir différencier ce type de cas de celui où l'ordinateur n'a pas réussi à résoudre l'anaphore[108].

Il semblerait que le point du traitement de l'ambiguïté dans l'analyse de textes soit le principal vecteur de l'Intelligence Artificielle. Certes, lever des ambiguïtés est une tâche tout à fait naturelle pour l'interlocuteur humain, du fait qu'elle est le principal objectif du processus de compréhension. Elle est, pour les machines, une opération « artificielle », étant donné qu'il s'agit d'une série d'artifices qui ont pour finalité l'élimination d'un phénomène parasite qui perturbe le fonctionnement « normal » du système. Par là, les ambiguïtés considérées comme simples à résoudre pour le sujet humain et qui en l'occurrence sont rarement perçues intentionnellement impliquent pour leur traitement automatique le recours à des moyens informatiques importants mais qui restent, malgré les grandes avancées enregistrées ces dernières années, très peu satisfaisants. L'on peut se demander alors si l'on peut s'inspirer de la manière dont fonctionne le système de

[108] Dans un cas optimal, le système automatique devrait réussir à identifier les cas d'anaphores ambiguës.

compréhension humain pour affiner de nouveaux mécanismes automatiques[109]. Il ne s'agit nullement de vouloir modéliser le système psychique : en effet notre connaissance en est encore assez modeste, sans oublier que le but est plus d'obtenir des machines efficaces que de simuler les comportements humains. Cependant, on ne peut écarter l'hypothèse que l'étude des mécanismes exploités dans la levée des ambiguïtés par le sujet humain peut offrir des indications très pertinentes quant au type de techniques informatiques qui pourraient faciliter l'intégration de ce phénomène dans le processus de compréhension automatique.

[109] Çe sera notre objectif dans le chapitre suivant.

Chapitre 3 : Dimension automatique de l'anaphore

Le développement des systèmes de communication électroniques, de plus en plus performants, est accompagné d'une augmentation incessante du nombre de documents textuels électroniques disponibles tels que les résumés de notre corpus RESUMAN. Cette évolution nécessite la mise au point d'outils informatiques efficaces capables de sélectionner, de structurer et d'extraire les informations pertinentes contenues dans ces documents. Plusieurs communautés de recherche spécialisées dans le traitement de l'information, notamment celles de l'Extraction d'Information (EI) et de la Recherche d'Information (RI) ont proposé, à travers divers travaux, des applications d'accès à l'information textuelle comme la résolution automatique de l'anaphore pronominale. Pour pouvoir faire de l'extraction d'information à partir de textes en langage naturel, il importe de comprendre ces textes. Cela pose donc la question de la compréhension automatique des textes écrits en langage naturel. Dans ce chapitre, nous examinerons, tout d'abord, les procédures de l'extraction d'information et la recherche d'information et par la suite, la compréhension automatique des textes en langage naturel. Le choix de ces procédures sera justifié tout au long notre analyse.

1. Applications nécessitant la résolution de l'anaphore pronominale

1.1. Extraction d'information

L'extraction d'information consiste, dans un domaine restreint, à extraire d'un texte en langage naturel les informations pertinentes dans le but de satisfaire un besoin défini à l'avance. Ce besoin correspondant, dans le cadre de notre travail, aux antécédents recherchés par l'utilisateur, doit donc être spécifié au préalable par ce dernier sous forme d'une requête. Les informations extraites doivent ensuite être représentées de façon synthétique résumant les différents liens sémantiques existant entre les entités concernées. L'extraction d'information nécessite cependant une compréhension des textes qui doit être suffisante pour pouvoir répondre aux besoins de l'utilisateur. En effet, la tâche d'extraction du bon antécédent se concentre la plupart du temps sur la récupération à partir des textes des informations sous forme d'un ensemble de critères destinés à remplir des structures prédéfinies.

Le développement des techniques d'extraction d'information a été profondément influencé par une série de conférences connues sous le nom de MUC (*cf.* chapitre 3 de la

première partie). Les conférences MUC sont basées sur le principe de la compétition. Il s'agit de soumettre aux différentes équipes participant à ces conférences un ensemble de textes à analyser pour ensuite remplir des formulaires correspondant à une demande d'information spécifique. Typiquement, un formulaire correspondra à un ensemble de champs qui vont être remplis à partir des informations figurant dans le corpus de textes. Un exemple de RESUMAN est présenté en figure 12. L'ensemble des informations à extraire sont aussi appelées *patron* (ou *template).* La spécification d'un événement particulier et les relations à extraire sont désignées par un scénario.

Patron :

Pronom :

Antécédent :

Phrase :

Antigone souhaite également s'expliquer avec son fiancé Hémon. Elle lui demande de la pardonner pour leur dispute de la veille.

Résultat :

Pronom : Elle

Antécédent : Antigone

Figure 12 : Extraction d'informations par remplissage de formulaires

Nous soulignons que préalablement, pour toute tâche de résolution de l'anaphore, les textes à partir desquels les informations sont extraites sont d'abord divisés en phrases et en mots. Par la suite une analyse lexicale est appliquée et à chaque mot est associée une étiquette morpho-syntaxique. Pour cela l'analyseur syntaxique Fips[110] est utilisé. Cette phase précède nécessairement celle de l'extraction des antécédents. La détection des relations référentielles ainsi que leur résolution nécessitent des traitements allant au delà d'une simple analyse basée sur une approche statistique. Ce processus d'inférence concerne essentiellement deux aspects. Tout d'abord, le document peut contenir des informations sur un événement particulier dispersées tout au long du texte. Toutes ces informations ont besoin d'être combinées avant la génération du patron. Ensuite, il peut

[110] Nous utiliserons Fips lors de la création de l'outil RESUMAN.

exister des informations qui sont implicites. Ces informations doivent être expliquées et par conséquent explicitées. Les tâches de combinaison d'informations éparpillées et d'explicitation d'informations implicites sont réalisées au moyen des procédures d'inférence[111].

La résolution de l'anaphore pronominale est une tâche dont dépend l'extraction d'information avec certaines autres sous-tâches présentées ci-après[112] :

1. La reconnaissance des entités nommées :

En tout premier lieu, l'extraction d'information requiert la reconnaissance des entités nommées dans le texte. Il s'agit des entités telles que des noms de personnes, des noms d'organisations, des noms de lieux, des dates, etc. Les noms propres sont très importants dans notre corpus. Les méthodes utilisées pour extraire les entités nommées sont variées (Schnedecker 2015a, 2015b). Elles peuvent être basées sur le repérage d'indices comme les titres honorifiques (M., Mme, Mlle, Dr, etc) pour l'extraction des noms propres de personne. Ci-dessous, un exemple de $RESUMAN_c$ qui illustre ces usages :

> [1] Trois sœurs, Lady Bertram, Mrs Norris et Mrs Price ont épousé la première un Lord, la deuxième un révérend, la troisième un lieutenant de marine sans éducation. Mrs Norris qui n'a pas d'enfant et se croit charitable, fait venir de Portsmouth, sa nièce défavorisée Fanny Price âgée de dix ans pour l'élever à Mansfield Park, propriété familiale commune. Mais alors que Mrs Norris avait indiqué qu'elle était prête à héberger la jeune Fanny, celle-ci devra finalement habiter chez les Bertram, et à leurs frais, Mrs Norris étant trop pingre pour la prendre elle-même en charge. (Résumé Mansfield)

Ici, nous avons utilisé Fips pour repérer tous les noms propres présents dans RESUMAN et les extraire dans une base de donnée à part[113]. La figure ci-dessous montre la reconnaissance d'un échantillon des entités nommées dans $RESUMAN_c$:

[111] Nous reviendrons sur ces procédures dans la 3ème partie du travail.
[112] Il faut noter qu'un système d'extraction d'information ne contient pas forcément toutes les tâches énumérées ci-dessous. Tout dépend du niveau d'analyse attendu.
[113] Nous reviendrons sur cette démarche dans la 3ème partie.

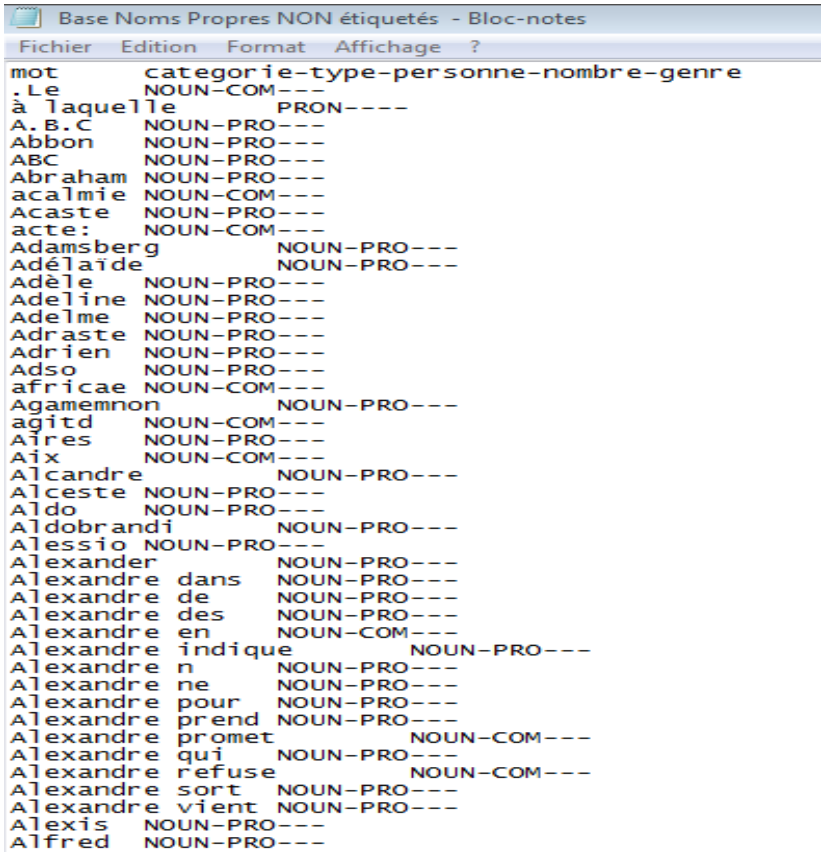


Figure 13 : Reconnaissance des noms propres dans RESUMAN par Fips

2. La découverte de relations :

Une fois les entités nommées identifiées, il s'agit de mettre au jour les liens sémantiques existant entre elles. Notamment il faut arriver à déterminer *qui fait quoi et à qui*. Cette tâche consiste à repérer dans un texte quand il est fait référence plusieurs fois à une même entité, même si cette entité est mentionnée de façon différente ou si un pronom personnel est utilisé en référence à cette entité. C'est essentiellement l'objet de ce travail. Une entité nommée présente dans un texte peut être présentée sous différentes dénominations. Il s'agit alors de pouvoir regrouper sous un même patron l'ensemble de ces dénominations. L'exemple suivant illustre ce que nous venons de dire :

> [2] Pourvus de cinq filles à marier, Mr et Mrs Bennett espèrent que l'une d'elles saura plaire à Mr Bingley, leur riche nouveau voisin. Malheureusement l'orgueilleux Mr Darcy, ami influent de Bingley, voit d'un très mauvais oeil son ami s'éprendre de Jane Bennett. (Résumé Orgueil et Préjugés)

Mr Bingley et *Bingley* seront regroupées sous un même patron antécédent. Les deux entités nommées renvoient à la même personne qui relève du masculin singulier.

Lorsque nous interrogeons RESUMAN, nous attendons un nombre de réponses

supérieur ou égal à un. À partir de l'ensemble de réponses obtenues, nous pouvons mesurer les performances de notre algorithme de résolution mis en œuvre pour retrouver le bon antécédent. Les critères de mesure des performances sont le rappel et la précision[114]. Ces mesures sont faites en tenant compte des considérations suivantes : un champ d'un formulaire sera jugé correct s'il contient une valeur clé ; un champ est incorrect s'il ne contient pas une valeur clé. Ainsi pour un système d'extraction d'entités nommées dont on souhaite connaître la pertinence des informations extraites, on définit par[115] :

- N_{total} le nombre total des champs dans les patrons réponses établis dans le corpus, autrement dit le nombre total de réponses attendues du système ;
- $N_{correcte}$ le nombre de réponses correctes produites par le système ;
- et $N_{incorrecte}$ le nombre de réponses incorrectes produites par le système.

Nous adoptons la définition de Chaumartin (2013 : 214) du rappel et de la précision ci-dessous :

- Le rappel correspond au nombre de pronoms correctement résolus et annotés par rapport au total de pronoms réellement présents dans le texte. Il est donc sensible aux faux négatifs : si le système possède de nombreux éléments annotés différemment où qui n'apparaissent pas dans la liste des réponses, on parle de silence. Le rappel est calculé par la formule suivante :

$$\textbf{\textit{Rappel}} = \frac{N_{correcte}}{N_{total}}$$

- La précision est sensible aux faux positifs (éléments annotés par erreur). Toutes les réponses retournées superflues ou non pertinentes constituent du bruit. Elle est définie comme :

$$\textbf{\textit{Précision}} = \frac{N_{correcte}}{N_{correcte} + N_{incorrete}}$$

[114] Nous utiliserons ces deux mesures dans la troisième partie pour évaluer la performance de $RESUMAN_O$.
[115] Nous nous plaçons dans la perspective des travaux effectués dans les conférences MUC.

Différents niveaux d'extraction d'information peuvent être, alors, distingués selon que la recherche privilègie le rappel ou la précision. En effet, le fait de cibler la recherche sur une entité particulière réduit le nombre de résultats mais augmente la précision. Inversement, si on utilise une requête très ouverte en employant des termes génériques, on se trouve confronté au risque d'avoir beaucoup de données extraites peu pertinentes car liées à des informations non recherchées. Aussi, plus on augmente le niveau d'abstraction, plus le risque d'erreurs augmente et plus la qualité de la précision en sera affectée. La stratégie adoptée par le concepteur ou l'utilisateur d'un système d'extraction d'information doit donc être motivée par l'objectif à atteindre, à savoir si c'est la couverture ou la précision qui est privilégiée.

1.2. Recherche d'information

Une modélisation efficace des anaphores est pertinente pour le développement d'outils performants de recherche d'information dans les corpus. La recherche d'information a pour objet la collecte des informations pertinentes à l'utilisateur en fonction de critères spécifiés dans des requêtes. Les informations pertinentes s'obtiennent en appariant une requête à une collection d'informations. Le domaine de la recherche d'information peut se caractériser par un certain nombre de tâches classiques, relativement aux types de requêtes formulées. Initialement, il s'est concentré sur la problématique des requêtes *ad-hoc* (ou requêtes ouvertes). Dans ce type de recherche, l'utilisateur formule des requêtes sous forme de mots clés ou de texte libre et le système est censé répondre en retournant un ensemble de documents ordonnés selon une mesure de pertinence. Généralement, on considère que les corpus de textes sont fixes dans ce cas. Ce type de recherche est utilisé surtout pour les bases de données bibliographiques ou les moteurs de recherche sur Internet. À de la recherche basée sur les requêtes *ad-hoc*, s'ajoute la recherche basée sur les requêtes fermées ou fixes. Dans ce cas, les requêtes sont finies et connues à l'avance. Les corpus sont généralement dynamiques et changent au cours du temps (exemple : les corpus évolutifs comme les forums électroniques ou les messages électroniques). Ce type de recherche, appelée aussi routage, peut être vu comme un problème de classification pour lequel on désire ordonner les documents par rapport à leur pertinence pour une classe. La différence entre les requêtes *ad-hoc* et le routage vient du fait que dans le cas du routage les informations sur la pertinence sont disponibles.

On peut aussi distinguer les systèmes de recherche basés sur le filtrage. Comme le routage, le filtrage peut être vu comme un problème de classification à la différence que le

filtrage effectue un choix sur la pertinence d'un document par rapport à une classe. Il s'agit donc d'une décision binaire (le document est soit pertinent soit non pertinent), tandis que pour le routage, l'évaluation de la pertinence d'un document est relative (exemple : on dit que le document d_1 est plus pertinent que le document d_2). Pour cette raison, on peut avancer l'hypothèse selon laquelle le filtrage est plus difficile que le routage. En effet dans le cas d'un filtrage, une estimation de pertinence doit être faite pour chaque document au moyen d'une évaluation de probabilité (Manning/Schûtze 1999).

L'évaluation des systèmes de recherche d'informations est un problème complexe puisqu'elle doit prendre en considération l'utilisateur. Plusieurs mesures d'évaluation des systèmes de recherche d'informations ont été proposées[116]. Parmi ces mesures, on peut citer la précision et le rappel qui sont les plus employées.

Malgré la grande utilité des systèmes de RI et leur succès auprès d'un large public, les techniques de recherche d'information ne permettent pas de répondre à des besoins d'information précis dans la mesure où elles n'appréhendent que le niveau du document et ne donnent pas directement accès au contenu informationnel des documents eux-mêmes. Cela ne correspond donc pas à l'objectif de notre travail.

1.3. Compréhension automatique de textes

Nous avons distingué, dans le chapitre précédent, essentiellement trois types de textes : des textes argumentatifs, des textes descriptifs, et des textes narratifs. Le type de texte influence fortement son mode de compréhension. La majorité des travaux existants, jusqu'aux années quatre-vingt, en intelligence artificielle et en linguistique se sont limités aux récits. Nous reprenons les mots de Sabah/Grau (2000 : 293) pour justifier encore notre choix des résumés de textes narratifs comme corpus :

> La compréhension de récits présente l'avantage de ne pas être orientée par une tâche précise et permet ainsi d'étudier les problèmes réels de la compréhension. Les mécanismes mis en œuvre dans ce cadre sont donc représentatifs des processus cognitifs utilisés pour la compréhension en général et peuvent être utilisés dans des applications variées.

Certes, la compréhension automatique des textes reste l'objectif principal de tout système complet de traitement automatique du langage naturel. Il ne s'agit pas seulement de créer des documents ou des informations qui pourraient satisfaire des besoins spécifiques, mais aussi de viser à expliciter les structures intentionnelles et donner lieu à une sémantique globale du texte. Ceci nécessite donc « d'expliciter pourquoi le texte est

[116] Des exemples détaillés sur différentes mesures ont été proposés par Manning/Schûtze (2000).

cohérent vis-à-vis des connaissances générales que possède le système sur le monde de référence : reconnaissance du sujet du texte, construction des relations de cohérence[117]» (Sabah/Grau 2000 : 293) et procéder à une résolution correcte des anaphores identifiées dans un texte.

Les différents mécanismes exploités pour parvenir à une représentation de la sémantique globale du texte sont analogues à ceux mis en œuvre pour l'extraction d'informations, notamment une analyse linguistique comprenant une représentation sémantique des faits ainsi qu'une considération des tâches d'identification de relations entre les différentes entités et des résolutions d'anaphores. Par ailleurs, le contexte où se place le texte analysé doit être pris en compte. Les relations entre les différentes entités concernent à la fois les relations inter-phrastiques et intra-phrastiques. Le fait qu'il existe un lien direct entre les différentes entités du texte et ses anaphores s'explique par le fait que la sémantique d'un texte se construit progressivement. Prendre en compte le contexte où se situe le texte nécessite impérativement la disponibilité de certaines connaissances générales et spécifiques sur les conditions d'énonciation des faits évoquées dans ce texte. La compréhension se construit donc parallèlement à l'acquisition des connaissances et à leur exploitation. Minel (2003) a souligné que l'échec des systèmes de compréhension automatique de textes du point de vue de leur généricité et de leur efficacité a conduit différentes équipes de recherche à se tourner vers des systèmes d'extraction d'information même si la compréhension automatique reste un domaine où de nombreux travaux continuent à être menés.

1.3.1. Une définition de la compréhension d'un texte

Il est vrai que si la faculté de compréhension est indispensable, sa définition reste des plus délicates. Pour Nazarenko (2004), la compréhension n'est pas une faculté dichotomique. En effet, il est possible de *comprendre un texte* sans le comprendre *vraiment* et sans le comprendre *entièrement*. Le fait d'ignorer le sens de certains mots ou de ne pas comprendre certaines phrases n'est ainsi pas un obstacle insurmontable à la compréhension du texte dans sa globalité. Il est nécessaire de préciser que dans une approche cognitiviste du traitement de l'information, comprendre consiste à élaborer une représentation dans un but bien précis. Cela implique donc que la compréhension n'est pas une finalité en soi, mais une activité préliminaire à d'autres tâches, comme le résumé, la traduction, etc.

[117] *Cf.* premier chapitre de l'actuelle partie.

Depuis quatre décennies, quatre types d'approches sémantiques se sont développées (Rastier 2001) :

- La sémantique logique : Elle a pour tâche d'évaluer la vérité des énoncés. Elle présente ainsi la signification comme étant la relation entre un symbole et l'objet qu'il dénote, dans le monde de ce qui est, dans un monde possible ou dans un monde virtuel. La sémantique logique, désignée aussi par l'appellation sémantique vériconditionnelle ou sémantique formelle fait défaut de capacité descriptive par rapport à la complexité des formalisations qu'elle utilise.
- La sémantique psychologique : Elle définit la signification comme le rapport entre des signes et des représentations ou opérations mentales.
- La sémantique cognitive : Celle-ci s'inscrit dans le développement de la sémantique psychologique ; pour elle la signification n'est qu'une représentation mentale. Elle émane d'une linguistique mentaliste qui explique tous les phénomènes linguistiques par des opérations mentales.
- La sémantique linguistique autonome : Pour elle, la signification est semblable à un rapport linguistique entre les signes langagiers, plus précisément entre signifiés.

Bien que ces théories et approches de la sémantique soient différentes, elles ont toutes un seul objectif qui est la modélisation de la sémantique des langues naturelles.

Approches sémantiques	Logique	Psychologique	Cognitive	Linguistique
Valeur	Evaluation de la vérité de l'énoncé	Rapport entre signe et représentation	Représentation mentale	Rapport entre signe et signifié

Tableau 11 : Récapitulatif des approches sémantiques

Dans l'ouvrage *Sémantique pour l'analyse*, Rastier *et al.* (1994), soulignent le fait que la compréhension est une activité humaine qui produit la conscience d'un résultat sans pour autant appréhender la méthode suivie pour y aboutir. Il propose pour cela l'exemple qu'on ne saurait décrire pourquoi et comment on a procédé pour comprendre une phrase ou

un mot. Etant donné que le but visé par le processus de compréhension automatique des textes est l'extraction de leur sémantique, il convient de définir la sémantique.

Baylon/Mignot (2000) définissent la sémantique comme étant la science qui étudie le sens ou la signification. En d'autres termes, nous pouvons dire que la finalité de la sémantique est la mise en forme, par l'étude des structures de la langue, des règles qui déterminent la formation du sens pour véhiculer par ces règles une interprétation humaine sans pour autant chercher à la simuler. Nous pouvons rappeler ici la polémique autour de la notion de sens entre Kayser d'une part, Kleiber/Riegel d'autre part : Kayser (1987) critique la notion de sens et n'y voit qu'une « abstraction qui n'est pas fondée scientifiquement » ; Kleiber/Riegel (1989) constatent, quant à eux, que :

> la tendance actuelle, (...) serait plutôt à l'interprétation des phrases selon une géométrie variable qui articule leur sens littéral (compositionnel) avec des données contextuelles et illocutionnaires (ce qui est dit littéralement, qui le dit, à qui, comment, quand, où, pourquoi, etc.), sans oublier les connaissances générales et particulières que les interlocuteurs se prêtent réciproquement.

Nous partageons la description de Kleiber/Riegel (1989) qui est toujours d'actualité, car malgré l'évolution des techniques d'analyses, l'étude de la sémantique n'a pas trop changé en trente ans. Kayser (1987) affirme qu' « il est connu, que sauf entraînement spécial, un interlocuteur ne sait pas donner une représentation du sens d'un énoncé »[118]. En effet, un interlocuteur ne peut que produire une paraphrase ou faire des inférences à partir d'un énoncé. Ricœur (2006) définit comprendre par « dire la même chose autrement ». La question est en fait de mettre en relief des indicateurs sémantiques dans l'outil linguistique et à organiser les dépendances réciproques entre ces indicateurs.

Il est clair que les progrès de la syntaxe formelle de Chomsky (1969) ont été à l'origine des premières approches computationnelles de la sémantique. Celles-ci étaient donc basées sur le principe que l'analyse sémantique, et donc la sémantique, procédait forcément d'une analyse syntaxique. A la suite et avec le développement de la théorie des langages formels basés sur la logique, l'idée d'une analogie entre la sémantique formelle (i.e *la sémantique des langages formels*) et la sémantique des langues naturelles s'est ainsi confirmée. En effet, prenant en considération la différence qui existe entre les langages formels et les langues naturelles, ces études cherchent incontestablement à formuler la signification à l'aide de formules logiques. Cela explique en partie le progrès de la

[118] Pour plus de détails, Lyons (1995) analyse l'ensemble de la question de sens, quant à Rastier (2001) ou Récanati (2007), ils abordent deux points de vue engagés de la problématique de sens.

sémantique formelle des langues naturelles. Nous pouvons dans ce contexte évoquer l'une des théories sémantiques dans la sémantique formelle, celle des représentations du discours (DRT[119]). Elle s'inscrit dans la série des travaux de Montague (1970). En DRT, le discours est compris comme étant une suite de phrases conditionnées successivement par les règles de construction d'une représentation sémantique. Les représentations sémantiques sont appelées des DRS[120] (Représentations des Structures de Discours).

Certes, les règles de construction des DRS sont en fait des règles d'actualisation d'une DRS produite par le traitement du discours antérieur. Ce procédé conduit ainsi à la transformation de la DRS préalable en une nouvelle DRS, laquelle devient la DRS transformée par la phrase suivante et ainsi de suite. La DRT vise essentiellement à mettre en place une représentation propositionnelle canonique du sens qu'une analyse du discours perfectionnera par la suite en procédant au calcul d'expressions référentielles, de relations temporelles et de résolutions d'ellipses. Le résultat auquel on aboutit est une formule proposée comme une représentation d'un sens littéral, mais qui ne peut véhiculer les divers sens possibles d'un même énoncé.

Les théories de sémantique formelle comme la DRT n'évoquent pas la possibilité de changement d'information du discours. L'idée principale des sémantiques fondées sur l'approche dynamique est que l'interprétation d'un discours est un processus incrémental. L'interprétation de chaque phrase actualise un état d'information préliminaire sur le monde pour arriver à un nouvel état d'information.

La sémantique se répartit ainsi en deux classes : l'une relative à l'étude du sens des mots pris individuellement (la sémantique lexicale) et l'autre l'étude de la manière dont les sens des mots sont agencés pour former la signification des phrases ou des textes. Les mots peuvent alors être agencés selon une structure lexicale bien déterminée comme à la manière de WordNet[121]. Dans cette structure, il est possible d'identifier un certain nombre de relations particulières (relations d'hyponymie et d'hyperonymie, d'antonymie, de synonymie, etc.)

1.3.2. Modèles de compréhension automatique des textes

Le champ expérimental de la compréhension de textes a constitué un terrain d'investigation pour plusieurs chercheurs à des époques différentes et dans des domaines

[119] *DRT :* Discourse Représentation Theory
[120] DRS : *Discourse Représentation Structures.* Ce sont des notations logiques qui permettent de formuler clairement des problèmes de référence indéterminée et d'anaphore (Rastier et *al.* 1994).
[121] http://wordnet.princeton.edu/wordnet/

aussi variés que la linguistique : l'étude de la structure syntaxique de surface, les limites de la mémoire de travail (Ehrlich 1994 ; Blanc/Brouillet 2003 ; Ericsson/Kintsch 1995) en rapport avec les connaissances de monde du lecteur/auditeur (Kintsch, Patel/Ericsson 1999 ; McNamara/Kintsch 1996) et le type de structure textuelle (Caillies/Denhière 2001), etc. En TALN, l'intérêt pour ce domaine se signale par de nombreux travaux dans des perspectives diverses comme l'automatisation du lexique (Kintsch 1988 ; Kintsch 1993 ; Mullet/Denhière 1997) et la sémantique latente que nous aborderons dans cette section. Des études récentes, dans d'autres domaines que la linguistique, sont apparues comme le projet *CABeRneT*[122] en 2013.

1.3.2.1. Le modèle de construction-intégration de Kintsch

Le modèle de Construction-Intégration (CI) (Kintsch 1988), s'applique au traitement du discours parlé ou écrit, et décrit les processus de niveau supérieur par lesquels le lecteur comprend ce qui est relaté dans le texte. A ces processus de niveau supérieur, correspond une représentation macro-propositionnelle construite par un sujet lors de la compréhension d'un texte. Cette représentation, selon Kintsch, constitue le résumé du texte[123] qui apparaît alors comme le produit automatique de l'activité de compréhension. Ce modèle se caractérise par un certain nombre d'idées directrices. Pour Kintsch (1988), la structure de surface d'un texte est avant tout un ensemble de propositions organisées par des relations sémantiques. Cet ensemble, appelé la base de texte, représente ce qui est dit dans le texte. Il correspond à la microstructure du texte. Ensuite, le modèle décrit ce que pense le lecteur de ce que dit le texte. C'est le modèle de situation qui constitue la macrostructure du texte. La macrostructure se justifie par le fait que les propositions de la base de texte doivent être reliées au thème du texte. La microstructure est le niveau local du texte et la macrostructure son niveau global.

De manière plus détaillée ce modèle se décompose en deux étapes :

- L'étape de construction de la représentation

A ce niveau, les propositions d'une phrase sous forme prédicative sont traitées de manière cyclique les unes à la suite des autres. A chaque cycle de compréhension, un réseau est construit dynamiquement, dont les nœuds correspondent à des entrées sous une forme propositionnelle. Dans l'exemple suivant, deux propositions partagent en commun

[122]Compréhension Automatique de Textes Biomédicaux pour la Recherche Translationnelle. Pour plus d'informations sur ce projet, *Cf.* http://www.agence-nationale-recherche.fr/?Projet=ANR-13-JS02-0009

[123]Ceci rejoint la définition donnée ci-dessus à la compréhension comme activité préalable à d'autres tâches telles que le résumé, la base de notre corpus.

le même antécédent :

> [3] **Cyrano** rencontre Roxane chez **son** ami, le restaurateur Ragueneau. Roxane et Cyrano évoquent leur enfance heureuse. Puis Roxane révèle à **son cousin** qu'elle est amoureuse non de **lui**, mais d'un beau jeune homme qu'elle **lui** demande de protéger. (Résumé de Cyrano de Bergerac)

Les propositions *rencontre (Cyrano, Roxane)* et *révèle (Roxane, son cousin)* sont connectées par un lien positif. Par ailleurs les nœuds les plus activés du cycle précédent (la partie du résumé avant cet exemple) sont également intégrés à la représentation en cours de construction. Il peut arriver que le réseau construise aussi des liaisons entre des nœuds qui s'avèrent inappropriés. Dans ce cas, les nœuds sont connectés par des liens contraignants ce qui déclenche une ambiguïté référentielle dans la majorité des cas :

> [4] **Candide$_i$** est ébloui par la puissance de son$_?$ oncle$_?$, et par les sophismes lénifiants du docteur Pangloss$_l$, le précepteur$_l$. Il$_?$ admire également Cunégonde$_m$, la fille du baron$_j$. Tout bascule le jour des premiers ébats de **Candide$_i$** et de Cunégonde$_m$. La réaction du baron$_j$ est brutale, **Candide$_i$** est banni et chassé de cet$_?$ Eden. Il$_i$ se retrouve dans "le vaste monde".: **Candide$_i$** envoie Cacambo$_n$ racheter Cunégonde$_m$ au gouverneur$_o$ de Buenos Aires , tandis qu'il$_?$ ira l$_?$'attendre à Venise. (Résumé Candide)

Nous pouvons constater après plusieurs lectures et un certain temps de réflexion que le possessif de '*son*' peut référer aussi bien à *Candide* qu'au *baron* (il faut quasiment lire le roman pour vérifier qu'il s'agit de *Candide*, ce qui correspond généralement à la première intuition de lecture). Même remarque pour le personnel de *tandis qu'il ira l'attendre à Venise*, complètement ambigu entre *Candide, Cacambo, Cunégonde et le gouverneur de Buenos Aires*. On constate qu'il est difficile de délimiter l'antécédent du pronom personnel '*il*' et '*l*' : ils incluent forcément *Candide et Cacambo*, mais peuvent aussi inclure *Cunégonde* et *le gouverneur de Buenos Aires*.

- La phase d'intégration

C'est une phase qui correspond à la diffusion de l'activation. Elle réalise une intégration des contraintes représentées par des liens entre les nœuds-propositions, ce qui fait converger le réseau vers un état stable. Les nœuds qui sont contextuellement non pertinents sont au passage désactivés. Le pattern final d'activation est assumé en tant que représentation de la mémoire du lecteur du texte : chaque personnage évoqué plusieurs fois dans un texte se voit attribué une série de coréférences. Pour spécifier sa chaîne référentielle, il est nécessaire de la faire correspondre à la série adéquate de coréférence. Ces séries s'élaborent ainsi, par une succession d'appartenances. Or nous pouvons noter

qu'une expression référentielle peut référer non à un référent unique et précis mais d'une part à une alternative entre plusieurs référents possibles (ambiguïté), d'autre part à un ensemble non délimité qui intègre un ensemble de référents clairement identifiés.

On constate que ce modèle a essentiellement pour objectif de décrire les opérations sous-jacentes à l'activité de compréhension de textes sans pour autant chercher à automatiser ces opérations.

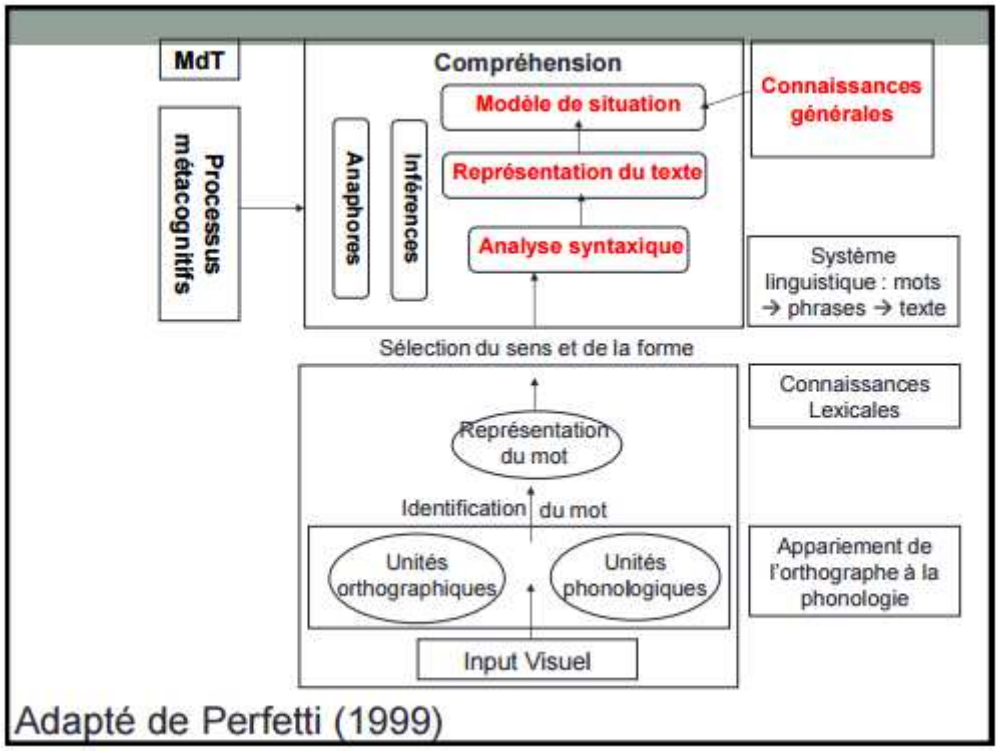


Figure 14 : Théorie de la compréhension de texte selon Kintsch (1992)

1.3.2.2. Le modèle de compréhension fondé sur l'analyse de la sémantique latente (LSA[124])

Pour comprendre un texte, une représentation des connaissances du sujet est fondamentale. Divers modèles de représentation ont été proposés pour en rendre compte. Certains de ces modèles ont donné lieu à des simulations informatiques. Notre interrogation est la suivante : comment présenter les connaissances humaines à une machine de façon à ce qu'elle parvienne à les comprendre correctement ? Malgré les tentatives de modélisation des connaissances d'un sujet humain, la plupart des heuristiques sont partielles et ne représentent qu'une petite partie des connaissances de départ. L'évaluation de la performance d'un modèle de compréhension automatique de textes est de mise puisqu'il est difficile de distinguer « entre ce qui provient des processus modélisés

[124] LSA : Latent Semantic Analysis.

et ce qui provient du caractère ad hoc des représentations » (Lemaire 2001 : 2). L'analyse de la sémantique latente a résolu une partie de ces limites :

> Le modèle LSA (…) permet de représenter sur une vaste échelle des connaissances correspondant approximativement à celles de sujets humains. Celles-ci sont représentées sous la forme de vecteurs de très grandes dimensions correspondant chacun à un mot ou à un ensemble de mots. Elles sont produites à partir de l'analyse automatique de grands corpus de textes. Ces représentations multidimensionnelles rendent possibles la simulation de modèles de compréhension de textes sur des données réelles. (Lemaire 2001 : 2)

LSA est un modèle d'acquisition de connaissances à partir de textes. Il permet de disposer d'une représentation des connaissances initiales et du texte. Parmi les travaux sur la modélisation de la compréhension basée sur l'analyse de la sémantique latente, on peut citer à titre d'exemple Kintsch (2000). Plus précisément, Kintsch (2000) proposait un modèle de compréhension des métaphores isolées basé sur l'analyse de la sémantique latente. Pour cela, il représente la signification de la métaphore comme un vecteur en fonction de la signification de deux termes impliqués appelés la topique et le véhicule. Pour simuler la compréhension d'une métaphore, ce modèle s'appuie sur l'identification de la topique et du véhicule ou en d'autres termes du prédicat P et de l'argument A. Le modèle consiste alors à rechercher parmi les termes sémantiquement voisins de P, ceux qui sont également proches de A. Dans ce cas, comprendre une métaphore de type "A est un P", c'est trouver parmi les différents sens de P ceux qui correspondent à A. La signification de la métaphore, représentée par un vecteur, est ainsi vue comme la somme des vecteurs de P, de A et de chacun des voisins communs.

Le modèle de Kintsch (2000) a inspiré les travaux de Lemaire et Dessus (2003) : ils ont présenté des modèles cognitifs fondés sur LSA et des expériences pour les valider. Ils ont pu démontrer que LSA peut modéliser quelques systèmes cognitifs. Nous partageons la conclusion suivante de Lemaire et Dessus :

> Qu'il soit clair que l'équivalence des performances humains et modèle ne présage en rien une identité des mécanismes sous-jacents : il serait ainsi surprenant que notre cerveau réalise une réduction de matrice. Par ailleurs, des pans entiers de la linguistique sont négligés par LSA, comme la syntaxe ou la pragmatique. Or, c'est aussi l'intérêt de ce modèle que d'être épuré, puisqu'il permet de cerner précisément les limites de cette sémantique conceptuellement pauvre mais cognitivement plausible. (Lemaire/Dessus 2003 : 74)

Nous remarquons, depuis la dernière décennie, que les chercheurs utilisent la LSA pour des finalités didactiques comme Zampa (2005) qui l'a utilisée dans un système

d'acquisition de langue étrangère de spécialité[125] et Dessus *et al.* (2011) qui a développé, Pensum[126] un système de compréhension automatique de textes qui est fondé sur LSA.

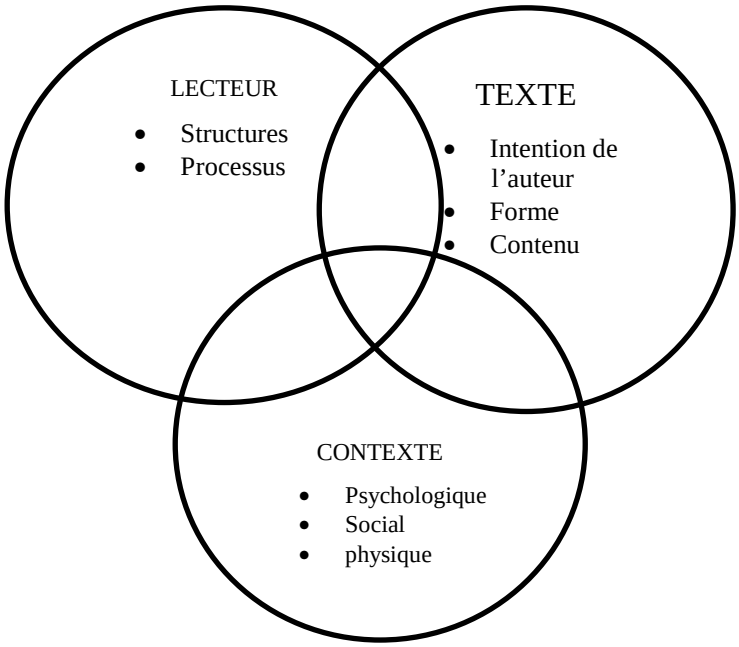


Figure 15 : Modèle contemporain de compréhension de lecture

La problématique au centre de cette thèse peut s'inscrire au carrefour de la problématique de l'extraction d'informations et celle de la compréhension automatique des textes. En effet, sans toutefois prétendre appliquer exactement et uniquement ces méthodes, des points communs peuvent être notés. Notre objectif est de pouvoir dégager une structuration référentielle dans RESUMAN. Cela va consister d'abord à identifier et à extraire les différents antécédents potentiels à analyser et à identifier les liens entre le pronom et le bon antécédent. Cette façon de procéder nous a conduit en effet à affronter la problématique de l'acquisition de connaissances et de ressources sémantiques à partir du corpus. Nous partons ainsi de l'idée que la compréhension est un processus qui s'appuie sur une représentation des connaissances (Lemaire/Dessus 2003). Pour pouvoir identifier et extraire le bon antécédent, nous avons construit des règles basées sur des indices linguistiques mais aussi sur les différentes connaissances acquises à partir de corpus de

[125] Ce système est nommé Rafales (Recueil Automatique Favorisant l'Acquisition d'une Langue Étrangère de Spécialité).

[126] Pensum est un système didactique qui « permet à des étudiants à distance de produire des synthèses de cours dans le but de les comprendre, et d'en avoir des retours automatiques fondés sur une analyse sémantique ». (Dessus *et al.* 2011)

textes. Nous nous sommes interrogée sur l'adaptation à RESUMAN des techniques de résolutions de l'anaphore, développées par le TALN. C'est pourquoi, nous nous intéresserons, dans la section suivante, à présenter les différents niveaux de traitement automatique des énoncés linguistiques.

2. Niveaux de traitement automatique nécessaires à la création de RESUMAN$_O$[127]

Le traitement automatique du langage naturel (TALN) est l'ensemble des recherches et développements mécaniques qui visent à modéliser la capacité humaine à comprendre des énoncés linguistiques. Notre volonté est de modéliser une approche de résolution des anaphores pronominales présentes dans RESUMAN. Nous introduisons, dans cette section, les différents niveaux de traitement[128] mettant en œuvre les connaissances linguistiques nécessaires pour la compréhension automatique d'un texte. Nous présenterons le fonctionnement automatique d'une analyse linguistique tout en mettant un accent particulier sur les analyses syntaxique et sémantique. Prenons l'exemple :

> [5] Le duc Alexandre prend la défense de son cousin. C'est à ce moment qu'apparaît Lorenzo. *Il* se moque du chancelier, qui le provoque en duel. (Résumé Lorenzaccio)

Pour réussir à comprendre automatiquement cet énoncé et à résoudre l'anaphore pronominale *il*, nous envisageons des traitements successifs à appliquer. Nous nous appuyons pour cela sur l'analyse proposée par Yvon (2007 : 6) : il faudra successivement :

– segmentation de texte (mots)[129] ;

– identification des composants lexicaux : traitement lexical ;

– identification des constituants de l'énoncé et des relations qu'ils entretiennent entre eux : traitement syntaxique ;

– construction d'une représentation du sens, en attribuant à chaque concept évoqué un référent dans un monde de référence : traitement sémantique ;

– identification enfin de la relation entre un énoncé (ou une phrase) avec les

---

[127] Nous parlons de l'outil RESUMAN.

[128] On distingue plusieurs étapes lors du traitement du langage naturel correspondant aux niveaux : phonétique, phonologique, morphologique, lexical, syntaxique, sémantique et pragmatique. Le niveau phonétique et phonologique est relatif à la composition des sons. C'est un domaine qui est étroitement lié à la physique et à l'acoustique. Ce niveau est rarement intégré dans les systèmes de TALN qui ne portent le plus souvent que sur des entrées écrites. A ce niveau, il existe un ensemble de règles précises décrivant comment prononcer un texte écrit (synthèse de la parole) ou comment transcrire une séquence de sons (reconnaissance de la parole). Précision importante, nous nous limiterons, dans notre travail, au traitement du langage sous forme écrite.

[129] Nous reviendrons sur ce niveau, en détail, au chapitre 3 de la troisième partie.

conditions situationnelles et contextuelles dans lesquelles les mots constituants sont utilisés: traitement pragmatique.

Les différents niveaux de traitement ou d'analyse du langage naturel écrit, décrits ci-dessus, sont rarement tous intégrés dans les systèmes de traitement du langage naturel. En effet, lors de toute activité d'analyse linguistique, on propose de suivre généralement le processus à trois étapes tel que le montre la figure 16 (Sabah 1988) :

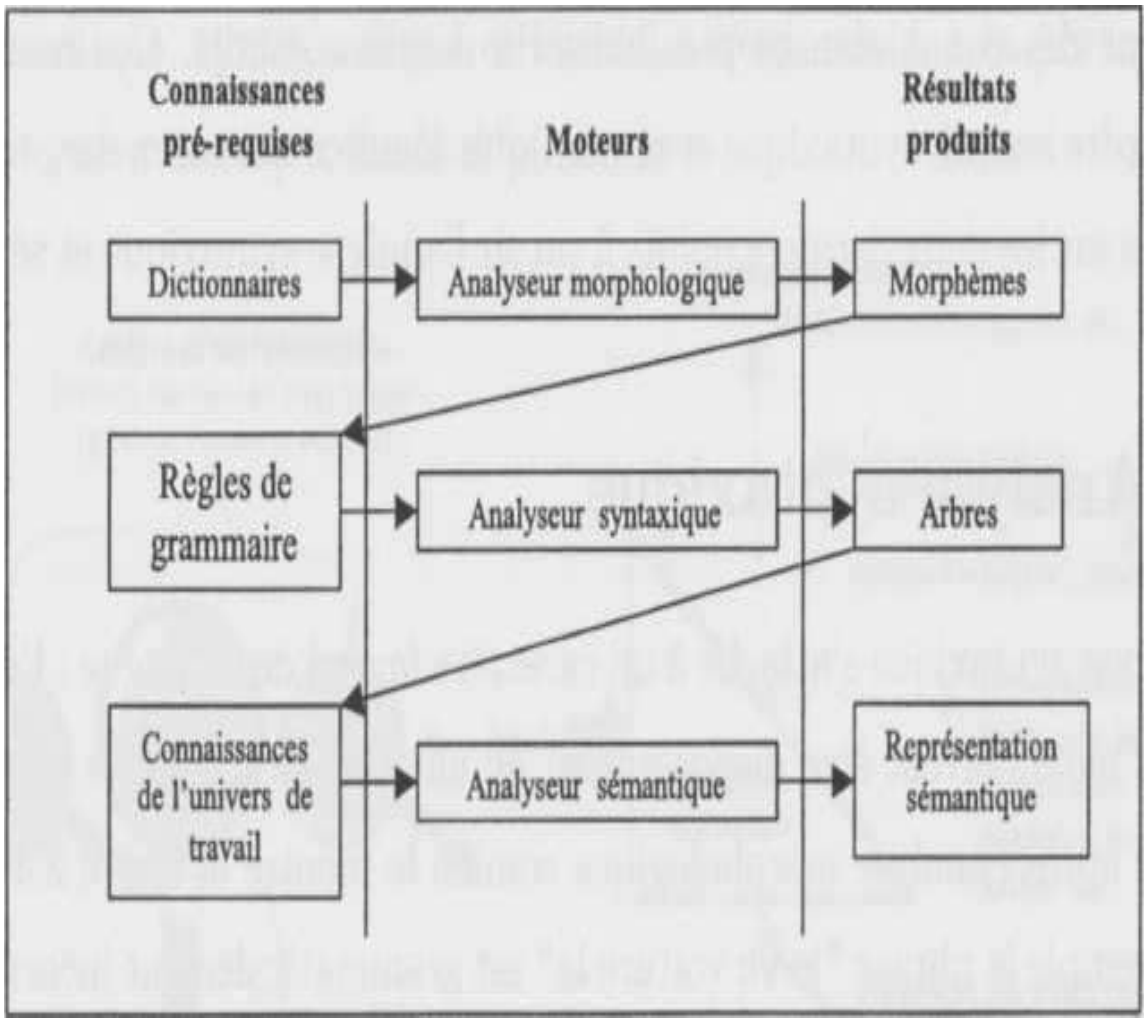


Figure 16 : Les étapes classiques d'une analyse linguistique

2.1. Niveau lexical

Le but de cette étape de traitement est d'identifier et reconnaitre les *mots* : ce niveau regroupe l'ensemble des mots d'une langue. Il porte sur la forme des mots dans une phrase. Ces mots peuvent être classés en catégories : généralement on distingue les noms, les adjectifs, les verbes, les adverbes, les conjonctions, les prépositions, les pronoms, les interjections. Cependant, on peut utiliser d'autres catégorisations propres suivant les

besoins. Certaines catégories telles que les noms, les adjectifs, les verbes et adverbes correspondent à des ensembles presque infinis de mots : ce sont des classes ouvertes. Les ensembles de conjonctions, de prépositions et de pronoms sont en revanche finis et sont des classes fermées.

L'ensemble des mots d'une langue est stocké dans le lexique. Les différentes formes morphologiques d'un mot sont accessibles via les règles morphologiques, c'est à dire que les formes fléchies ou dérivées peuvent être analysées ou générées par l'application de ces règles. Ce niveau est lié aussi à d'autres aspects complexes tels que la façon dont les mots sont formés à partir d'autres mots, par exemple la formation des adjectifs à partir des noms (*courage -> courageux*), nominalisation des verbes (*peindre -> peintre*), etc. : il s'agit de la morphologie lexicale. Généralement les dérivations sont obtenues par l'ajout des affixes. On inclut parfois à ce niveau le traitement des *mots composés* (combinaison des mots pour former un nouveau sens). Il est bien délimité et peut être défini avec un ensemble de règles et une liste plus ou moins courte d'exceptions.

En poursuivant sur l'exemple [4], l'étape d'identification lexicale devrait conduire à un résultat voisin de celui donné ci-dessous[130] :

Le	DET-DEF-SIN-MAS
duc	NOUN-COM-SIN-MAS
Alexandre	NOUN-PRO
prend	
la	PRON-CLI-ENCL-SIN-FEM
défense	NOUN-COM-SIN-FEM
de	PREP
son	DET-POS-SIN-MAS
cousin	NOUN-COM-SIN-MAS
.	PUNC
C'	PRON-DEM-SIN-MAS
est	VERB-IND-PRE-3-SIN
à ce moment	ADV-STD
qu'	CONJ-SUB
apparaît	VERB-IND-PRE-3-SIN
Lorenzo	NOUN-PRO
.	PUNC
Il	PRON-PER-3-SIN-MAS
se	PRON-CLI-PROCL-INN-ING
moque	VERB-IND-PRE-3-SIN
du	PREP-CON-de
chancelier	NOUN-COM-SIN-MAS
,	PUNC
qui	PRON-REL-SIN-MAS

[130] Résultat obtenu avec Fips.

le	PRON-CLI-PROCL-SIN-MAS
provoque	VERB-IND-SUB-PRE-3-SIN
en	PREP
duel	NOUN-COM-SIN-MAS
.	PUNC

Nous pouvons constater, en particulier, l'ambiguïté de la forme *NOUN-PRO* (nom propre) : ce traitement lexical n'a pas identifié le genre des noms propres (féminin ou masculin). Une telle précision est indispensable pour la résolution de l'anaphore pronominale dans RESUMAN[131].

2.2. Niveau syntaxique

La deuxième étape de l'analyse, qui est syntaxique, fonctionne au niveau de la phrase. Elle vise à fournir la liste des liens syntaxiques reliant les mots ou les groupes de mots entre eux. Elle se fait grâce à l'utilisation des règles de grammaire qui servent à caractériser la langue. Le résultat de cette analyse correspond au squelette syntaxique de la phrase. Il existe plusieurs façons formelles[132] d'écriture des règles à partir desquelles les phrases sont formées. Elle s'effectue après l'analyse morphologique.

2.2.1. Grammaire et syntaxe

A la lumière des travaux de Chomsky (1965) sur la grammaire générative (GG)[133], une grammaire peut être vue comme l'ensemble de règles d'une langue qui permet de dire si une phrase est syntaxiquement correcte ou non. La forme et la fonction des règles de définition diffèrent selon les types de grammaires considérés. La grammaire *context-free* (appelée aussi grammaire indépendante du contexte) est l'un des fondements de l'informatique linguistique. L'ensemble des phrases syntaxiquement correctes par rapport à une grammaire déterminée est obtenu par application d'une série de règles de réécriture de la forme X->Y où X et Y sont des constituants et la flèche (->) signifie *doit être réécrit* : Par exemple la phrase *je vu voiture la* est grammaticalement incorrecte selon une règle de français qui spécifie qu'un participe passé d'un verbe *voir* doit être précédé d'un auxiliaire *être* ou *avoir*. On dit alors qu'il y a une erreur de syntaxe. Des telles règles ne spécifient rien sur le sens des mots ou des phrases. Ces règles sont exprimées à l'aide de différentes grammaires que nous analyserons dans les sous-sections qui suivent. La connaissance de

[131] Nous reviendrons sur ce traitement des noms propres dans le chapitre 3 de la troisième partie, consacré à notre approche proposée.
[132] Nous y reviendrons plus longuement dans la troisième partie de notre thèse.
[133] La GG a pour objet l'établissement d'une théorie des structures linguistiques à travers une grammaire. Cette théorie doit être centrée autour de la syntaxe.

ces grammaires nous est primordiale afin de choisir celle qui sera adoptée lors de la création de RESUMAN.

2.2.2. Analyse en constituants immédiats

L'analyse en constituants immédiats (ACI), qui résulte pour l'essentiel des travaux des distributionnalistes américains Bloomfield (1887-1949), Harris (1909-1992) et Hockett (1916-2000), consiste à décrire la structure syntaxique d'une phrase sous la forme d'une construction hiérarchisée présentée le plus souvent comme un arbre. La figure 6 présente un exemple d'ensemble de règles pouvant être utilisées lors d'une analyse syntaxique. Cette grammaire pourra générer par exemple la phrase *Le duc prend la défense de son cousin,* dont la représentation arborescente est donnée à la figure 17 avec les trois niveaux correspondants : phrastique, syntagmatique et lexical :

P	—>	GNGV
GN	—>	Art Nom
GN	—>	Art Adj Nom
GV	—>	Verbe GN
Art	—>	le, la, un, une, les, des,...
Nom	—>	duc, cousin, garçon, fille, Alexandre,...
Verbe	—>	prendre, défendre,...
Adj	—>	brillant, belle, vieux, amoureux,...

Figure 17 : Un exemple de grammaire en constituants immédiats

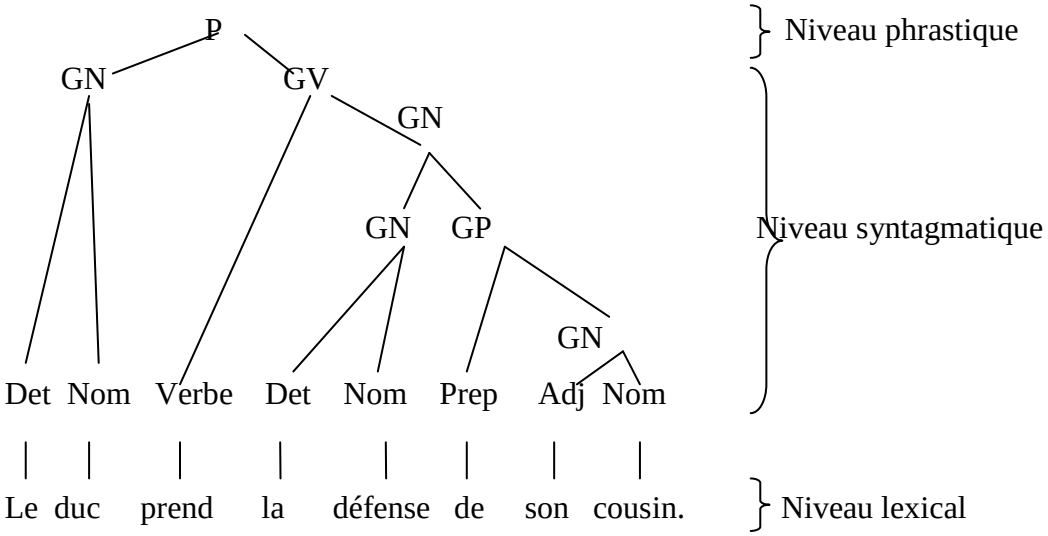


Figure 18 : Représentation arborescente de la structure grammaticale d'une phrase

Pour générer une phrase correcte on part du symbole initial P (Phrase) puis on

sélectionne la ou l'une des règles qui possède ce symbole dans la partie gauche et on lui substitue la partie droite et ainsi de suite jusqu'à l'obtention des symboles terminaux (*cf.* figure 18). On remarque que d'autres phrases peuvent être générées par la même grammaire.

Réécritures

P
GN GV
Det Nom GV
le Nom GV
le duc GV
le duc Verbe GN
le duc prend GN
le duc prend GN GP
le duc prend Det Nom GP
le duc prend la Nom GP
le duc prend la défense GP
le duc prend la défense Prep GN
le duc prend la défense de GN
le duc prend la défense de Adj Nom
le duc prend la défense de son Nom
le duc prend la défense de son cousin.

Règles utilisées

P → GN GV
GN → Det Nom
Art → le
Nom → duc
GV → Verbe GN
GN → GN GP
GN → D Nom
GP → Prep GN
GN → Adj Nom
Verbe → prend
Art → la
Nom → défense
Prep → de
Adj → son
Nom → cousin

Figure 19 : Réécritures nécessaires à la génération de la phrase *le duc prend la défense de son cousin.*

De la même manière qu'une grammaire en constituants immédiats peut générer une phrase correcte, elle peut aussi analyser une phrase donnée en entrée. L'analyse d'une phrase à partir d'une grammaire en constituants immédiats est basée sur la stratégie du *backtracking* (analyse descendante avec retour en arrière) comme le montre la figure 19. Généralement, les analyseurs sont constitués de deux parties essentielles : une partie constituée de l'ensemble de connaissances adaptées à une grammaire particulière et une autre qui est l'interpréteur utilisant ces données pour construire l'arbre syntaxique. Dans le cas des analyseurs en constituants immédiats, les connaissances linguistiques sont données directement sous la forme de règles de réécriture : Lorsqu'une phrase est soumise en entrée, l'analyseur déroule la grammaire depuis le constituant initial (P : phrase) jusqu'aux feuilles tout en attachant de nouveaux nœuds à l'arbre syntaxique. Lorsqu'une feuille de la grammaire (ou un mot) correspond à un terme de la phrase entrée, elle est alors placée dans la structure engendrée et le mot suivant est soumis à l'analyse. Puisqu'on part du symbole

initial P pour aller jusqu'aux mots de la phrase, on dit que l'approche est descendante. Cependant, on peut aussi procéder par approche ascendante, en partant des mots de la phrase jusqu'au symbole initial P. Les règles déjà utilisées sont stockées dans une pile[134] et les constituants qui ne sont pas encore parcourus y sont marqués, ce qui permet au système de retrouver ses points de décision.

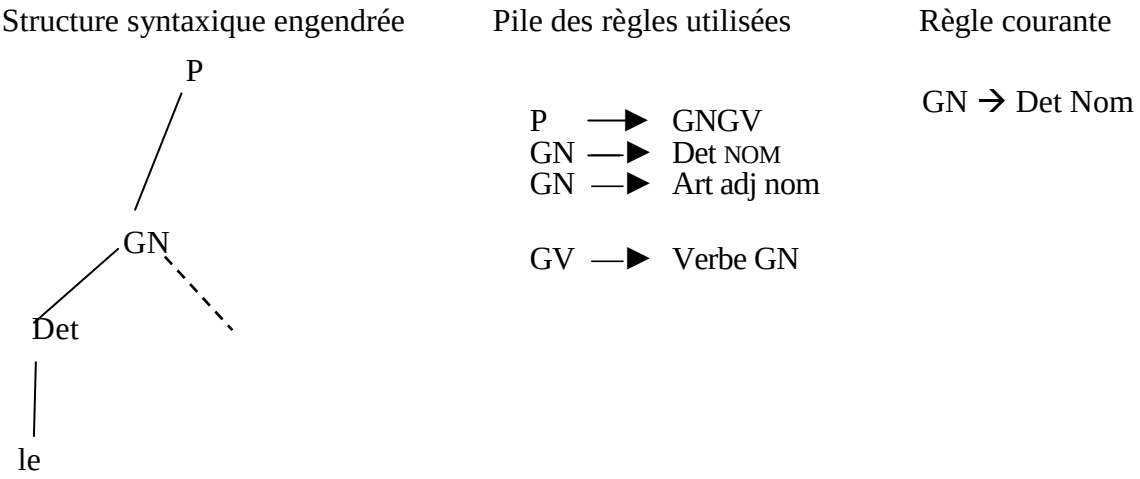


Figure 20 : Analyse descendante avec backtracking

Il peut arriver que deux règles différentes correspondent par exemple à un groupe nominal donné. Dans ce cas, le système doit faire un choix qui peut mener à une impasse : l'ambiguïté ! Chomsky a fourni un cas extrême d'ambiguïté syntaxique (plusieurs structures possibles pour une même séquence de mots) : selon lui, il y a quatre façons différentes d'analyser *Time flies like an arrow*, données par les quatres arbres de la figure 21 :

[134] Ensemble des nœuds d'un arbre syntaxique.

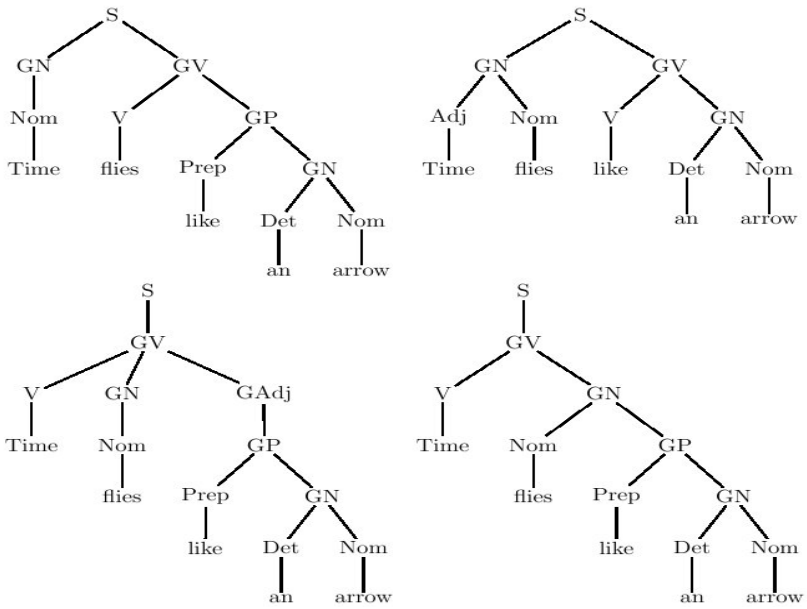


Figure 21 : Quatre analyses syntaxiques de la même phrase

Il existe aussi un autre moyen de spécification des grammaires permettant d'analyser les phrases. Il s'agit des réseaux de transition.

2.2.3. Grammaires à réseau de transition

Contrairement aux grammaires en constituants immédiats, les grammaires à réseau de transition (GRT) ne sont pas composées de règles. Elles sont constituées d'un ensemble de graphes étiquetés et identifiés par un nom. Les arcs de ces réseaux désignent des mots, des classes lexicales (article, nom, verbe, etc.) ou des catégories syntaxiques qui correspondent à d'autres réseaux (*cf.* figure 10). Les nœuds servent à indiquer des étapes dans l'analyse d'une phrase. Selon Bègue (1982), chaque réseau de transition est spécialisé dans l'analyse d'une catégorie grammaticale. Ainsi le réseau P est chargé de l'analyse des phrases, G s'occupe des groupes nominaux, GV analyse les groupes verbaux et PP les groupes prépositionnels. Pour qu'une catégorie grammaticale soit analysée favorablement, il faut que tous ses arcs soient traversés.

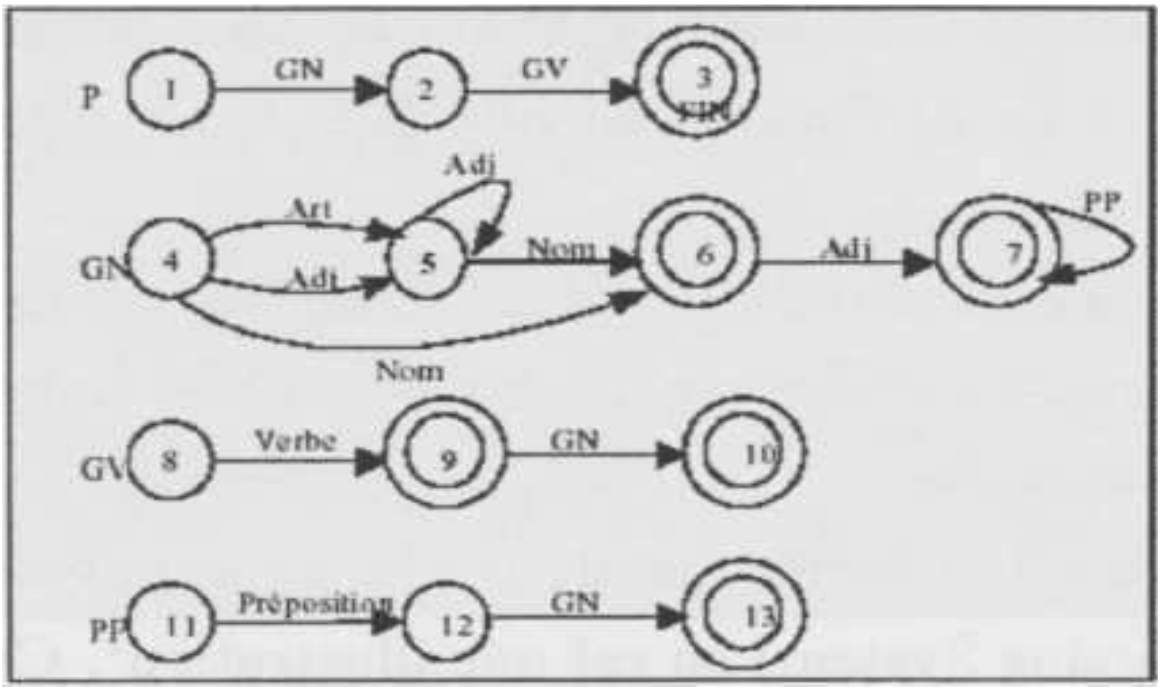


Figure 22 : Une grammaire à réseau de transition

Pour qu'une phrase soit analysée correctement, il est nécessaire que le réseau P ait recours aux autres réseaux (tels que GN, GV ou PP, etc.) pour l'analyse des sous-catégories grammaticales composant la phrase. La traversée d'un arc ne peut s'effectuer que sous trois conditions :

1. si l'arc est désigné par un mot identique au terme courant,
2. ou si l'arc se réfère à la même catégorie lexicale que le terme courant,
3. ou si l'arc renvoie à une autre catégorie semblable.

L'analyse d'une phrase par GRT consiste donc, au préalable, à s'adresser au réseau de plus haut niveau (P) qui doit vérifier la phrase et sa grammaticalité. Lorsque l'état final est atteint par P, on dit que la phrase est analysée correctement. Là aussi, on peut être amené à faire plusieurs choix et, par conséquent, à être confronté à une impasse, et dans ce cas, un *backtrack* s'avère nécessaire.

Cependant, les GRT, tout comme la grammaire en constituants immédiats, ont des limites s'il s'agit d'analyser complètement une phrase. Par exemple, il est difficile de prendre en compte l'accord en genre et en nombre entre un nom et ses adjectifs, ou entre un verbe, son sujet et ses compléments puisqu'elles n'analysent que les classes grammaticales sans prendre en considération l'aspect morphologique du lexique. Une amélioration des GRT a été proposée : il s'agit des grammaires ATN (Augmented Transition Networks) (Woods 1970, Pereira *et al.* 1980) voire les grammaires à réseaux

de transition augmentées (GRTA). Leur différence par rapport aux réseaux de transition réside dans le fait que deux types d'éléments sont associés à leurs arcs :

- des conditions augmentant les critères de sélection pour le choix d'un arc,
- des actions servant à conserver des données utiles dans la construction des structures syntaxiques.

Il s'ensuit des observations et des analyses proposées ci-dessus que l'analyse en constituants immédiats sera la grammaire à suivre lors de la création de RESUMAN. Quelques produits intéressants d'analyse syntaxique ont été réalisés grâce à ces grammaires formelles : Le produit Fips[135] en est une illustration. Ce produit a été développé par le Laboratoire d'Analyse et de Technologie du Langage (LATL[136]) du département de linguistique de l'université de Genève. Fips est un analyseur syntaxique multilingue qui peut analyser syntaxiquement des textes en français, anglais, allemand, etc). Wehrli[137] (2009) présente son outil ainsi :

> Au plan linguistique, l'approche sous-jacente repose sur une adaptation du modèle chomskyen "minimaliste" (Chomsky, 1995), avec de nombreux emprunts à d'autres modèles génératifs, tels que la grammaire lexicale fonctionnelle (LFG) (Bresnan, 2001) ou le modèle de Simple Syntax de Culicover/Jackendoff (2005).

Nous nous sommes servie de Fips pour extraire la structure syntaxique des phrases de RESUMAN en XML-TEI[138]. La figure 23 illustre la segmentation syntaxique de l'exemple :

> [6] Isabelle paraît dans des habits de princesse. Elle se plaint de Clindor, qu'elle a depuis, épousé. (Résumé l'illusion comique)

[135] En plus d'être un outil gratuit en ligne, il répond à nos attentes comme Parser morphosyntaxique. Pour y accéder : http://latlapps.unige.ch/Parser

[136] L'outil *FipsColor* développé par le même laboratoire permet d'avoir le résultat sous forme coloriée.

[137] Version en ligne, consultée le 11/05/2016 : http://alpage.inria.fr/iwpt09/atala/fips.pdf

[138] Nous reviendrons, avec plus de détails, sur cette extraction dans le chapitre 2 de la troisième partie.

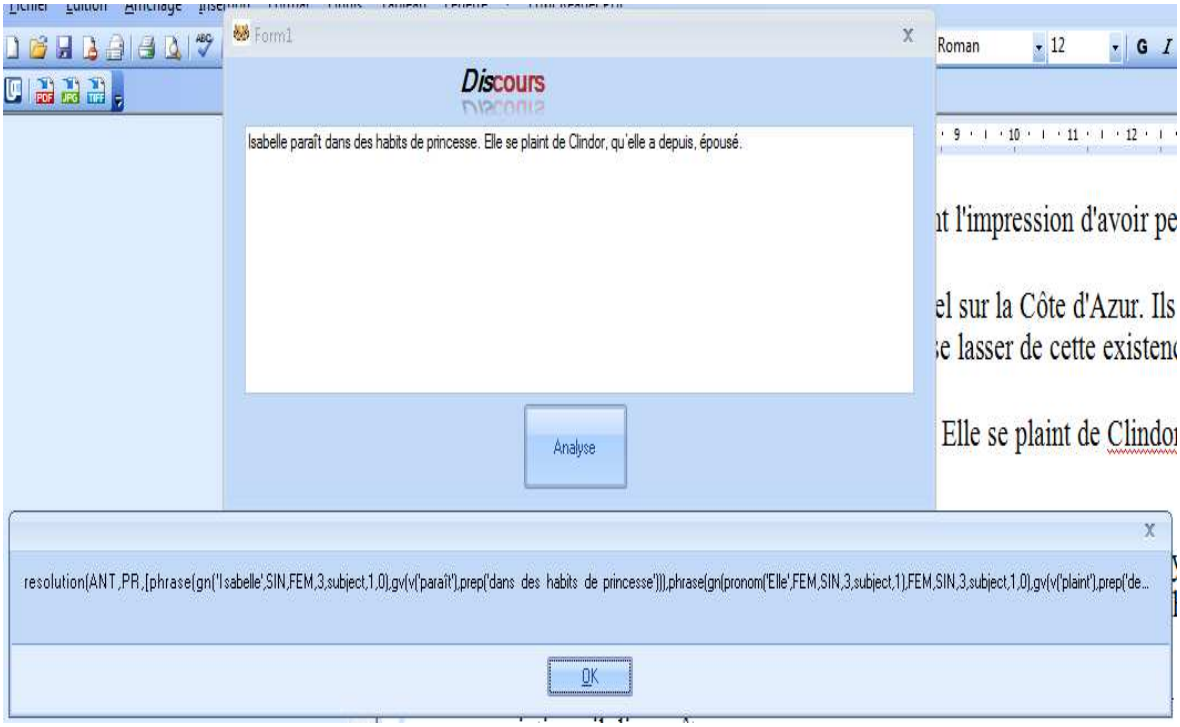


Figure 23 : Résultat de l'analyse de la phrase *Isabelle paraît dans des habits de princesse. Elle se plaint de Clindor, qu'elle a depuis, épousé.*
(Résumé l'illusion comique)

2.3. Le niveau sémantique

Le niveau sémantique a trait au sens des mots, des groupes de mots et de phrases. Chaque mot, au niveau lexical, est associé à une ou plusieurs représentations internes ou sémantiques qui peuvent varier d'une théorie à une autre et dépendent étroitement du contexte. Même si ce niveau est beaucoup plus complexe à décrire et à formaliser que les niveaux de traitement précédents car les énoncés, en général, apportent du sens qui ne peut pas être exprimé en logique formelle (Jackendoff 2002), des méthodes inspirées de l'analyse des programmes ont été considérées pour représenter la sémantique du langage naturel dans plusieurs domaines d'applications[139]. Pour le moment, on est incapable de construire des analyseurs sémantiques qui couvriraient la totalité d'une langue donnée et qui seraient indépendants d'un domaine d'application particulier : il semble impossible de représenter automatiquement les nuances *il mit furtivement le vin sur la table*[140].

L'analyse sémantique consiste à prendre comme entrée un mot, une expression, une phrase ou de manière générale un texte et à représenter sa partie significative. Sur la base

[139] Ces approches et limitations ont été discutées, entre autres, par Jackendoff (2002).
[140] Exemple emprunté à Yvon (2007).

des résultats de cette tâche, une structuration sémantique est générée. Cette structuration est représentée souvent au moyen des graphes de relations (Le Priol 2000) ou de graphes conceptuels (Muninn 2001, de Chalendar/Grau 2000).

La grammaire de cas[141] et la théorie des dépendances conceptuelles constituent deux techniques essentielles permettant de dégager la structure sémantique d'une phrase à partir de la structuration syntaxique.

La théorie des dépendances conceptuelles est introduite par un chercheur en Intelligence Artificielle, Schank, dans les années 70. Cette théorie repose sur l'idée que le sens de toute expression peut être construit à l'aide d'un ensemble relativement restreint de foncteurs, c'est-à-dire des symboles fonctionnels intervenant comme arguments des prédicats : reprenons la phrase *Le duc Alexandre prend la défense de son cousin.* de l'exemple [4] : elle est représentée ainsi <prendre (duc Alexandre, défense, cousin)>. Cette description repose sur le rôle que joue le prédicat et n'exprime pas clairement celui des foncteurs dans la construction du sens. La solution consiste à décrire une combinatoire fonctionnelle détaillée en se reposant non sur les prédicats mais sur les événements. « Le sens est construit par l'insertion de la description conceptuelle de chaque mot dans le schéma conceptuel en cours de construction pour la phrase » (Yvon 2007 : 16). La représentation des événements est au centre des dépendances conceptuelles. En effet, chaque événement est décrit sous la forme d'une structure comportant une action, un acteur, un objet et une direction, l'action étant ici une primitive sémantique différente du verbe et par conséquent indépendante de toute référence linguistique. Prenons un exemple :

[7] (a) Le duc Alexandre prend la défense de son cousin.
(b) Le duc Alexandre est chargé de la défense de son cousin.
(c) Le duc Alexandre décrit la défense de son cousin.

Nous remarquons que ces trois énoncés, bien qu'ils avaient la même structure syntaxique (même forme active, même forme pronominale, ...), n'ont pas les mêmes représentations conceptuelles: dans (a), c'est le duc qui défend volontairement son[?142] cousin, dans (b), le duc est chargé de défendre son cousin (il joue le rôle d'un avocat), et enfin dans (c) le duc n'est pas participant dans la défense de son cousin, il joue le rôle d'un rapporteur. C'est la représentation conceptuelle des trois verbes, stockée dans le lexique, qui permet ces interprétations. En dépendance conceptuelle, la représentation est un graphe

[141] Cf. chapitre 2 de la première partie.
[142] Nous ne pouvons, dans un contexte isolé comme le présente l'exemple [6], qu'admettre que *son* se réfère au duc.

dont les sommets sont des entités qui dérivent de quatre catégories conceptuelles de base :

- Les générateurs d'images (GIs) correspondent « en gros » aux noms, mais cette dénomination évoque leur fonction psychologique plutôt que leur rôle syntaxique.
- Les modificateurs d'images (MIs) correspondent « en gros » aux adjectifs, leur rôle est de modifier l'image d'une entité.
- Les actes sont les actions conceptuelles fondamentales. Leurs équivalents linguistiques les plus proches sont les verbes.
- Les modificateurs d'actions (MAs) modifient le concept correspondant à une action, "en gros" comme des adverbes.

Le nombre de primitives sémantiques devait rester faible selon Schank (1975). Il en a utilisé couramment onze :

- **ATRANS** : transfert d'une propriété abstraite (ex : *donner*).
- **TRANS** : modification du lieu d'un objet (ex : *aller*).
- **MTR ANS** : transfert d'une information (ex : *dire*).
- **MBUILD** : création de nouvelles informations (ex : *décider*).
- **PROPEL** : application d'une force sur un objet (ex : *pousser*).
- **ATTEND** : perception par un sens (ex : *écouter*).
- **SPEAK** : production de sons (ex : *parler*).
- **GRASP** : préhension d'un objet (ex : *prendre, saisir*).
- **MOVE** : mouvement d'une partie du corps (ex : *frapper*).
- **INGEST** : ingestion d'un objet (ex : *manger*).
- **EXPEL** : expulsion de quelque chose du corps (ex : *pleurer*).

La phrase *Le duc Alexander prend la défense de son cousin* sera représentée par exemple à l'aide du graphe de la figure 24. Les flèches indiquent une dépendance, la double flèche un lien privilégié entre l'acteur et l'action, **O** l'objet et **D** sa direction :

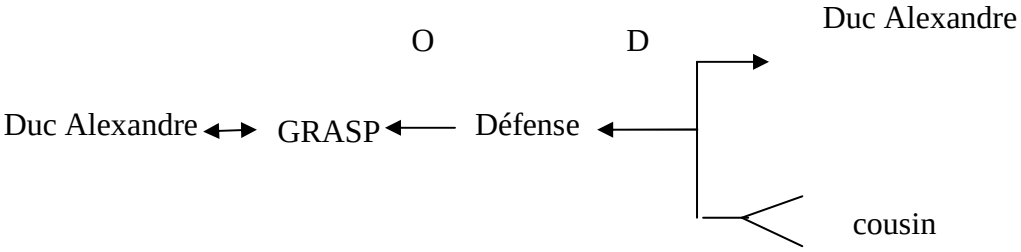


Figure 24 : Graphe de dépendance conceptuelle de la phrase *Le duc Alexander prend la défense de son cousin*

En plus des événements, les dépendances conceptuelles traitent aussi des états et de leur changement. Les figures 25a et 25b décrivent respectivement les phrases *Le duc est au palais* et *Le duc est mort.*

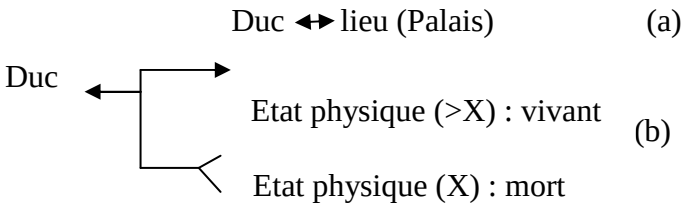


Figure 25 : Graphe de dépendance conceptuelle décrivant des états et des changements d'états

Les dépendances conceptuelles conduisent à une analyse de la phrase très voisine de celle employée par la grammaire de cas. A la différence de la grammaire de cas, les phrases sont transformées par l'analyseur en graphe de dépendances conceptuelles sans passer par une structure syntaxique. C'est également grâce à ces représentations conceptuelles que l'on peut résoudre correctement l'anaphore pronominale *il* dans l'exemple :

> [8] Le duc reçoit le cardinal Valori, de retour de Rome : ce dernier l'informe que le pape Paul III est irrité des désordres auxquels se livre **Lorenzo**, **que** le peuple surnomme. Sire Maurice, un chancelier abonde en ce sens **Lorenzaccio**. Le duc Alexandre prend la défense de **son cousin**. C'est à ce moment qu'apparaît **Lorenzo**. **Il** se moque du chancelier, qui **le** provoque en duel. Ce duel amuse Alexandre. **Lorenzo** s'évanouit à la vue de l'épée. (Résumé Lorenzaccio)

Le lecteur, qui lit le texte pour la première fois, ne peut pas savoir si *Lorenzo*, *Lorenzaccio* et *cousin du duc* sont les référents d'un même personnage. La résolution du pronom *il* est difficile car, dans ce contexte typique, c'est soit *le duc Alexandre* soit *Lorenzo* qui se moque du chancelier. C'est grâce au contexte d'énonciation et la lecture du

résumé que l'on peut ainsi reconstituer le sens de certains énoncés sous la forme d'une expression conceptuelle :

> [9] Le duc est jeune et mène une vie de débauché. Il règne sur la ville par la terreur, ne tenant compte ni du peuple, ni des autres grandes familles de Florence. On le déteste, mais pas autant que son cousin, son âme damnée : Lorenzo de Médicis, méchamment surnommé Lorenzaccio. (Résumé Lorenzaccio)

Toutefois, les glissements de sens, omniprésents dans l'usage linguistique, restent très difficiles à analyser automatiquement avec les techniques actuelles. Nous n'allons pas utiliser le traitement sémantique lors de la création de notre logiciel, bien que ce niveau nous paraisse plus puissant que le traitement morphosyntaxique[143].

2.4. Le niveau pragmatique

L'analyse sémantique d'un groupe de mots ou d'une phrase ne conduit à représenter que la partie de la signification des mots. Par conséquent, elle ne donne pas une signification complète des phrases ou des énoncés telle qu'un humain l'appréhende lors d'un processus de compréhension. D'où la nécessité d'un niveau pragmatique qui vise donc à trouver la signification réelle des phrases en fonction du contexte.

Cette étape travaille ensuite au niveau du texte dans son ensemble. Les informations syntaxiques extraites à l'étape précédente sont interprétées dans le cadre de règles liées à une connaissance du monde, ou tout au moins, du domaine traité. Le résultat sera une représentation des connaissances qui va servir par la suite à l'élaboration des raisonnements sur ce qui est connu ou à l'extraction d'informations selon les objectifs. Voici un exemple issu de notre corpus :

> [10] Il est plus fort que nous, Antigone, il est le roi. (Résumé Antigone)

Cette phrase sera interprétée comme un conseil si l'on sait qu'Antigone n'a pas respecté l'ordre du roi Créon. Cette interprétation n'est pas de nature sémantique. À partir de la compréhension du sens de l'intervention d'Antigone, Ismène, la sœur d'Antigone, réalise une inférence logique en utilisant une connaissance contextuelle, le conflit entre Antigone et le roi Créon. Il résulte que cette opération n'est pas une construction conceptuelle, c'est une opération logique. Elle appartient donc à la pragmatique. Nous partageons le point de vue d'Yvon (2007 : 20) :

---

[143] Ce niveau nécessite un système qui peut identifier les rôles sémantiques des groupes nominaux et le rôle sémantique de l'anaphore.

> Le niveau pragmatique est aussi invoqué à un niveau plus élevé, celui de l'organisation de larges tronçons de textes ou de discours. Il s'agit alors de repérer les relations rhétoriques et structurelles entre les passages. Les techniques correspondant à ce niveau de traitement sont encore très mal maîtrisées. Le niveau pragmatique, même si les techniques qui lui correspondent ne sont pas encore stabilisées, apparaît moins difficile à aborder que le niveau sémantique.

Il résulte des observations et des analyses proposées ci-dessus qu'une analyse linguistique est un mécanisme plus ou moins selon suivant les besoins, basée généralement sur un processus à quatre étapes nécessitant des connaissances préalables du domaine étudié, ces connaissances étant d'ordre lexical, syntaxique, sémantique et pragmatique. Nous avons exposé les différents niveaux pris en compte dans les systèmes de TALN ainsi que l'architecture générale du fonctionnement de ces derniers. Dans ce qui suit nous présenterons un traitement automatique à vocation formelle, s'intéressant au niveau de la morphologie et de la syntaxe, et en pensant à un niveau cognitif[144].

Pour étiqueter notre corpus, il a fallu adapter l'analyseur Fips (Werli 1997). Nous avons divisé RESUMAN en deux sous-corpus, eux aussi « petits » : le premier pour élaborer le logiciel, le deuxième pour le tester et le valider. De ce fait, et comme « la langue est constituée en grande partie de *pré-fabriqués* dont on peut faire l'analyse en interrogeant les corpus en s'appuyant sur des méthodes statistiques », nous avons crée un algorithme qui s'appuie sur le calcul de saillance (Landragin 2011) comme facteur principal de résolution des anaphores pronominales dans notre corpus. En prenant en compte différents facteurs syntaxiques et cognitifs, cet algorithme fait recourt à un modèle permettant d'évaluer d'une manière efficiente la saillance d'un antécédent potentiel. Ces facteurs comportent chacun un indice différent en fonction de leur utilité dans la résolution. Notre interrogation est la suivante : notre méthode statistique est-elle performante ?

[144] Nous reviendrons sur le traitement cognitif dans le premier chapitre de la troisième partie.

Troisième partie
Résolution automatique de l'anaphore pronominale dans RESUMAN$_O$

Afin de résoudre le problème des ambigüités que créent les anaphores en langue naturelle, plusieurs chercheurs du domaine du traitement automatique des langues (TAL) ont mis en place diverses approches. Celles-ci consistent à trouver la référence d'un groupe nominal qui peut se présenter sous la forme d'un nom avec un déterminant, un pronom personnel, un pronom démonstratif ou réfléchi et qui doit être interprété par rapport à un élément apparaissant avant lui dans le discours. Nous pouvons décomposer la résolution du problème de l'anaphore en deux étapes. Tout d'abord, la résolution de la coréférence[145] où deux groupes nominaux pleins faisant référence à un même élément dans le discours sont mis en relation. Ensuite, la résolution d'anaphores pronominales où il y a recherche de la référence d'un pronom (élément qui doit obligatoirement être interprété selon son contexte). Plusieurs formes linguistiques telles que les noms, les groupes nominaux, les verbes, ou des phrases (anaphore abstraite) peuvent être reprises par les pronoms anaphoriques.

Nous pouvons regrouper les algorithmes traitant le problème de l'anaphore en deux familles dans le domaine du traitement automatique des langues. Dans une première famille, nous pouvons regrouper les algorithmes qui, au niveau de leur fonctionnement, se basent sur les connaissances linguistiques. Ces connaissances utilisées peuvent être incomplètes ou erronées, ce qui engendre un défaut de fiabilité de ces algorithmes. Dans une deuxième famille, on regroupe les algorithmes qui ont montré leur limite en cas d'anaphores complexes, comme les algorithmes de Lappin/Leass, ceux-ci se basant sur des méthodes d'apprentissage automatique et sur des indices de surface. L'amélioration des performances de ce deuxième type d'algorithmes pour des textes avec de forts liens d'anaphores a fait l'objet de plusieurs travaux scientifiques.

Nous allons, dans cette partie, présenter une nouvelle version d'algorithme permettant d'améliorer les performances d'une résolution automatique en cas de textes à fortes liens d'anaphores. Celle-ci se base sur des connaissances linguistiques avec une approche statistique. $\text{RESUMAN}_{\text{O}}$ doit attribuer à chaque pronom un antécédent. Néanmoins, la résolution automatique de l'anaphore peut s'avérer difficile si l'antécédent se trouve très loin du pronom dans le texte ou s'il y a plusieurs candidats possibles ou si une grande quantité d'information est insérée entre les deux termes. Nous montrerons, dans

[145] Nous ne traitons pas ce genre de résolution dans notre travail. Nous nous intéressons exclusivement, comme énoncé dès le début, à la résolution de l'anaphore pronominale.

ce qui suit, qu'à l'aide du calcul de la saillance, nous parviendrons à avoir un bon résultat de résolution.

Chapitre 1 : Intelligence artificielle, linguistique et cognition

Les chapitres précédents ont permis de mettre en évidence l'état de recherche de la problématique du traitement automatique de l'anaphore ainsi que les difficultés d'interprétation de l'antécédent. Nous partons ici de l'hypothèse que les outils linguistiques permettent la résolution de certains de ces problèmes. C'est pourquoi nous proposons dans ce chapitre de montrer le cheminement scientifique qui a permis de développer des outils automatiques de résolution de l'anaphore. Tout d'abord, nous mettrons en lumière l'état actuel des connaissances dans ce domaine après un bref historique retraçant l'évolution des différentes approches du traitement automatique des langues dans la perspective «cognitiviste» de l'intelligence artificielle. Nous tâcherons d'énumérer les divers types de connaissances disponibles en informatique linguistique pouvant être mises en œuvre pour une compréhension automatique correcte des textes, voire une performante résolution de l'anaphore ; nous présenterons aussi les limites de l'intelligence artificielle actuelle et examinerons les raisons de ces limites. Dans un deuxième temps, nous mettrons en lumière l'importance de la résolution de l'anaphore pour la communication homme-machine, pour le développement de l'intelligence artificielle et pour la compréhension automatique des textes.

1. Evolution de l'intelligence artificielle

1.1. Facteurs d'apparition de l'IA

Le souci d'intégrer au traitement automatique des langues des connaissances générales, telles que le contexte socio-culturel, les différentes situations de communication ou encore la pratique des interactions, s'est développé très progressivement dans ce domaine de recherche. En effet, ces aspects ont pu tout d'abord sembler indépendants des connaissances linguistiques nécessaires au traitement de l'information (règles de grammaire ou autres). Nous mettons ici en exergue l'évolution de l'intelligence artificielle vers la nécessité d'une intégration progressive des connaissances (générales et linguistiques) dans le traitement automatique des langues et donc plus particulièrement des anaphores.

Poibeau (2014) présente un aperçu historique du TAL. Les premières recherches datent des années 50 et ne prennent pas en compte la complexité de la langue. A cette

époque, comme il le précise, la volonté des chercheurs américains pour déchiffrer les documents russes a été un moteur dans la mise au point de traducteurs automatiques, l'informatique étant à ses débuts essentiellement à usage militaire et industriel. Bar-Hillel a organisé la première conférence sur ce sujet en 1952 au MIT[146] et en 1954 les premières phrases russes ont été traduites automatiquement en suivant le modèle des automates de Turing[147]. Ces évènements ont été un déclencheur majeur dans le lancement de la recherche dans ce domaine. Ainsi, de nombreux projets furent lancés sans pour autant reposer sur une réelle évaluation, il faut le reconnaître, de l'ampleur de la tâche :

> En janvier 1954 eut lieu à New York la première démonstration sur ordinateur, une machine IBM 701, qui déclencha une accélération des recherches. Il s'agissait de la traduction de russe en anglais de phrases utilisant un vocabulaire de 250 mots et six règles de syntaxe mises au point par la *Georgetown University*. Bien que très limitée, cette démonstration fut montée en épingle par la presse et fit grande impression sur le public et certains scientifiques. (Léon 2002 : 85)

Ces principaux travaux, assez simplistes, ont ainsi consisté en la mise au point de dictionnaires électroniques. En effet, la traduction ne se faisait que mot par mot avec une possible réorganisation de l'ordre des mots. Concrètement, ces traitements, essentiellement restreints aux mots, ne prenaient pas en compte d'autres dimensions linguistiques.

Rapidement, des limites portant sur la question fondamentale de la représentation des connaissances et de leur utilisation ont été constatées. En effet, Bar-Hillel (1964) démontra qu'une bonne traduction de phrases nécessitait à la fois des connaissances sur la situation décrite (connaissances contextuelles) et des connaissances portant sur le monde en général (connaissances encyclopédiques). A cette époque, les aspects cognitifs commençaient à apparaître, mais étaient impossibles à prendre en considération. Ces résultats non satisfaisants[148] associés au coût très élevé de la traduction automatique ont, semble-t-il, amené le gouvernement américain à suspendre les financements publics.

Malgré cet échec, ces projets ont néanmoins contribué à la naissance d'idées fondamentales dans ce domaine. Ainsi, l'idée de l'intelligence artificielle serait née en 1956 à l'école d'été de Dartmouth lorsque « M. Minsky (psychologue), A. Newell (psychologue), H. A. Simon (psychologue, économiste), J. McCarthy (informaticien), C. Shannon (mathématicien) » (Rialle 1996) des personnalités marquantes de l'époque, ont

[146]Massachusetts Institute of Technology : Institut de recherche américain spécialisé dans l'IA.
[147]Il s'agit d'un modèle abstrait de machine introduit en 1936 par le chercheur anglais Allan Turing. Nous renvoyons pour plus d'informations au sujet de ce modèle à l'article en ligne (http://zanotti.univ-tln.fr/turing/) consulté le 15 Mai 2016.
[148]Révélés dans le rapport ALPAC (*Automatic Language Processing Advisory Council*) en 1966.

postulé qu'une machine pourrait simuler les aspects de l'intelligence humaine si ceux-ci sont décrits de façon précise, et créer ainsi des programmes informatiques que l'on pourrait appeler intelligents. Cette focalisation sur les processus de pensée humains permit le développement du cognitivisme qui influença considérablement la psychologie, la linguistique, l'informatique et la philosophie.

Le premier modèle à intégrer la dimension discursive de la langue est celui de Shannon (1940) et Weaver (1955). Ce premier modèle est basé essentiellement sur des caractéristiques de fluidité des échanges langagiers et sur une conception du langage ne prenant pas en compte la complexité des relations entre langue et contexte. En effet, selon ce modèle, lors de l'énonciation du message, l'énonciateur produit une expression codant le sens du message en appliquant des règles d'encodage. L'auditeur, de son côté, combine tous les éléments identifiés en utilisant un processus de décodage afin de reconstruire le sens du message (les sons produits, les structures syntaxiques utilisées, les relations sémantiques correspondantes). La compréhension du message ne dépend en aucune façon de caractéristiques contextuelles. Les premiers modèles de traitement automatique des langues développés par l'intelligence artificielle sont fondés sur cette conception de la communication : les phrases de la langue sont supposées correspondre à des faits réels, la compréhension est vue comme un ensemble de transformations successives d'un langage de représentation dans un autre[149].

A partir de ces constats, nous émettons, dans notre recherche, l'hypothèse suivante : la mise en place d'un système formel de résolution de l'anaphore inclut le fait que chaque cas anaphorique se prête à une formule de ce système. Chaque contexte, dans lequel figure l'anaphore, peut être soumis, aussi, à cette formule et, enfin, une étude statistique sur ces cas donne lieu à des raisonnements sur la référence réelle.

Dans les modèles qui ont suivi, le processus cognitif que l'on a cherché à modéliser est complètement modifié par le programmeur qui sert d'intermédiaire lors de la construction des représentations. Certes, ces systèmes d'intelligence artificielle sont généralement dépourvus de moyens de perception et d'action sur le monde réel. Néanmoins, des formes de compensations possibles de cette absence de perception et d'action apparaissent, permettant l'analyse et le traitement des inférences fondées sur des systèmes symboliques et ce qu'ils peuvent entraîner d'interprétations humaines implicites

[149] En effet, les contraintes temporelles et la rapidité de la compréhension impliquent que l'architecture sous-jacente aux processus de compréhension soit séquentielle, car ces processus sont indéniablement amorcés dès le début de la production

et externes. Nous pouvons alors nous demander dans quelle mesure ces derniers processus ne sont pas une phase incontournable pour toute signification.

Dans les évolutions récentes du traitement automatique des langues, trois caractéristiques se distinguent désormais : une recherche de la consistance et performance des analyses et une orientation vers les linguistiques de corpus liée à une volonté aigue d'évaluation :

- une recherche de la consistance et performance des analyses : en effet, quelles que soient les situations non prévues (fautes d'orthographe, de grammaire, omissions ou mots non connus du système), les systèmes actuels cherchent le plus possible à trouver une interprétation ; c'est ce qu'on appelle une analyse consistante, c'est-à-dire la capacité pour un système à proposer une réponse. Lorsqu'une analyse complète n'est pas possible, une stratégie d'analyse partielle est alors mise en place. Un analyseur consistant et performant se doit de présenter une analyse syntaxique pour chaque phrase du texte et pour pouvoir analyser des phrases qui ne sont pas intégrées par la grammaire du système. Dans cette stratégie qui permet de limiter l'explosion combinatoire fréquente en analyse syntaxique, l'analyseur traite autant que possible chaque partie de phrase en entrée et renvoie une analyse construite à partir de ces éléments partiels. En restant dans le domaine symbolique, divers analyseurs ont été proposés en ce sens ; on peut citer Wehrli (1997), Hull *et al.* (1997), Gaudinat *et al.* (1998), Vergne (2003) pour le français. Néanmoins, pour ce qui est du traitement de l'anaphore, aucun système n'a encore été proposé[150]. Grâce à l'énorme quantité de textes électroniques maintenant disponibles, une autre approche s'est développée, permettant de renforcer l'analyse consistante. Elle permet d'appliquer sérieusement des méthodes statistiques (et d'apprentissage) au traitement automatique des langues. En effet, l'absence d'une grammaire complète couvrant toutes les formes possibles d'une langue est un obstacle essentiel à la réalisation d'un analyseur robuste. Afin d'obtenir des indications sur le contexte d'apparition de mots, l'analyse statistique consiste à supprimer les mots grammaticaux et à évaluer les fréquences d'apparition des mots ou des groupes de mots qui indiquent les structures associatives privilégiées. Avec des taux actuels de reconnaissances de 95 % à 98 %, ces techniques statistiques sont d'une grande aide dans le domaine des étiqueteurs automatiques, (Manning/Schütze 1999). Afin de sélectionner, éventuellement, l'analyse la plus probable, les analyseurs statistiques présents donnent lieu, généralement, à toutes les analyses possibles pour chaque phrase

[150] Nous reviendrons sur l'état des systèmes de résolution de l'anaphore dans le chapitre suivant.

d'un texte et utilisent ensuite des fréquences d'occurrence des séquences de mots. Par exemple, à partir d'un corpus de textes syntaxiquement annotés, Charniak (1997), décrit un analyseur statistique qui conduit à une grammaire particulière.

- le développement d'une linguistique de corpus : les corpus exhaustifs de textes sur support électronique annotés avec des informations linguistiques, sont un matériau très utile pour les recherches linguistiques ; ils le sont également pour les applications en traitement automatique des langues. De nos jours, ces corpus de plus en plus nombreux, sont disponibles pour différentes langues. Pour l'anglais par exemple, on observe le développement d'une linguistique de corpus et une amélioration considérable des programmes de traitement automatique. En effet, de nombreux travaux sur ce type de corpus existent depuis une dizaine d'années. Les premiers corpus annotés commencent également à être disponibles en français. En effet, leur utilité est reconnue tant en linguistique (Blanche-Benveniste 1996) qu'en traitement automatique (Habert *et al.* 1997).

Des techniques d'évaluation très pointilleuses ont vu le jour grâce à ces corpus. Cela se manifeste par exemple pour l'analyse syntaxique. En effet, l'application des analyseurs syntaxiques, sur le plan de l'écrit, était plus difficile que les systèmes d'étiquetage morpho-syntaxiques (*cf.* campagne GRACE, Chibout *et al.* 2000). On commence toutefois aujourd'hui à considérer des protocoles pour l'évaluation des systèmes de compréhension (Blache *et al.* 1997). Il faut noter que ces mesures n'ont rien d'absolu et comportent une certaine part d'arbitraire, car elles restent principalement basées sur des comparaisons avec des étiquetages réalisés par des experts.

L'évolution du traitement automatique des langues, que nous venons de retracer brièvement a des conséquences sur le plan linguistique comme sur le plan informatique. C'est ce que nous proposons à présent d'étudier

1.2. Evolution des aspects linguistiques et informatiques du TAL

La théorie des grammaires formelles et transformationnelles constitue un point d'ancrage important dans la recherche en linguistique. Néanmoins, en visant essentiellement une formalisation de la compétence linguistique des sujets parlants, elle aboutit à des traitements purement syntaxiques et n'est pas adaptable à l'intelligence artificielle. Cette contrainte a caractérisé le début de cette théorie et on cite, Petrick (1973 : 30), auteur de cette époque initiale :

> Cette théorie, qui se veut une formalisation de la compétence linguistique d'un individu, débouche sur des traitements **purement syntaxiques**, mais reste

> difficilement applicable à l'époque en intelligence artificielle : dans un cas extrême, une phrase de 17 mots a, par exemple, donné lieu à 572 interprétations syntaxiques.

Cette théorie a permis le développement de grammaires comme celle de Harris ou encore Salkoff, essentiellement axées sur un traitement syntaxique. La grammaire en chaîne[151] de Harris était à la base des premiers analyseurs automatiques de l'anglais. Elle confère une souplesse importante pour les réalisations pratiques, notamment pour exprimer l'ordre relatif des divers constituants de la phrase. Mais elle s'identifie par certains aspects à une grammaire formelle non contextuelle. Salkoff (1973), quant à lui, s'intéresse particulièrement à l'analyse des textes scientifiques et développe une grammaire en chaîne du français. Sans utiliser de structures profondes, son objectif principal est de décrire des phénomènes de surface avec un appareil théorique minimal. Son intérêt pour l'intelligence artificielle ne peut être vérifié que si l'on cherche à réaliser une analyse purement syntaxique, avant tout autre traitement ; autrement, elle est très peu utilisée dans des mécanismes de simulation de la compréhension automatique des textes.

Se sont alors développées à partir des années 70 des grammaires syntagmatiques capables de traiter des aspects plus sémantiques. Ainsi, le logicien Montague (1970), parallèlement au développement de la théorie de Chomsky (1971), développa un modèle pour traiter principalement les aspects sémantiques, considérant le langage comme un processus logique. Montague ne considère la notion de vérité que par rapport à une interprétation dans le cadre d'une logique intensionnelle (niveau de la langue) et introduit ainsi dans le schéma qu'il propose une caractéristique significative, qui le différencie de l'approche de Chomsky. En effet, à chaque mot de la langue, il attache deux informations : sa catégorie syntaxique et sa catégorie sémantique logique (sa définition intensionnelle, qui représente son « sens »). Par ailleurs, Montague vise à établir une correspondance absolue entre syntaxe et sémantique (à chaque règle syntaxique correspond une règle sémantique) et porte ainsi son intérêt sur une grammaire catégorielle. En parallèle de la structure syntaxique, les règles sémantiques visent particulièrement à construire une structure logique dont la fonction est de représenter le sens de la phrase. L'interdépendance de ces deux types de règles est un point essentiel de la théorie, l'une n'est autorisée que si l'autre est applicable. Cela fait apparaître l'importance, pour cette théorie, du principe de

[151]Une grammaire en chaîne est « un exemple de programme d'analyse de toutes les phrases d'une langue donnée ; elle fournit une analyse exprimée dans une métalangue abstraite par rapport à la langue objet. » (Salkoff, 1971)

compositionnalité[152] qui gouverne beaucoup de mécanismes représentés par des symboles. Les représentations des sens des groupes de mots et des phrases sont obtenues par des combinaisons des représentations des mots, combinaisons guidées par les contraintes syntaxiques.

On retrouve dans les grammaires de cas l'écho du choix de la sémantique. En effet, dans ce type de grammaire, l'essentiel se trouve dans l'identification du type de relation qu'il peut y avoir entre le verbe, considéré comme le centre de la phrase, et ses divers compléments ; on y considère ainsi plus que la structure elle-même. En mettant l'accent sur la structure profonde des phrases, les grammaires de cas ne peuvent laisser de côté leur sens. De plus, les grammaires de cas analysent les dépendances conceptuelles entre les phrases d'un même texte, ce qui les amène à prendre en compte leur sens général. Au vu de leurs avantages, ces grammaires ont eu un impact majeur sur les travaux d'intelligence artificielle. Sager et *al.*, supposent que :

> Si on applique à un corpus de textes d'un secteur scientifique des méthodes de linguistique descriptive similaires à celles utilisées pour le développement d'une grammaire d'une langue dans son ensemble, on obtient des motifs précis de cooccurrences de mots à partir desquels on peut définir des sous-classes de mots et des séquences de ces sous-classes qui sont caractéristiques [...]. Ces catégories lexicales et formules syntaxiques de la grammaire du sous-langage sont étroitement corrélées aux classes d'objets du monde et aux relations qui sont propres à ce domaine. (Sager *et al.* (1987), cité par Bourigault/Fabre (2000 : 133).

En effet, tout d'abord, elles donnent un modèle de la structure de base d'une phrase où la sémantique est déterminante et où elle reste facile à utiliser. L'interprétation des énoncés est ainsi affinée par des catégories sémantiques. Ensuite, tout en prenant en compte les contraintes syntaxiques, elles s'apparentent à un processus d'analyse purement sémantique qui tient considérablement compte des restrictions de sélection dérivant des verbes. De ce fait, l'analyse casuelle reste largement reprise dans de nombreux systèmes de traitement automatique du langage même si, seule, elle ne résout pas tous les problèmes de compréhension. Il existe néanmoins des différences de conception entre des chercheurs comme Fillmore (1968) et Schank (1975), représentants tous deux de la grammaire de cas. Le premier pense qu'une structure émanant de cette notion de cas est relative à chaque verbe du lexique tandis que le second ne définit de structure de cas que pour les onze actions primitives composant la base de sa dépendance conceptuelle. Ces actions

[152] La signification d'une expression « doit être fonction de celle de ses composantes et des relations syntaxiques entre ces dernières, et de rien d'autre » (Vallée 2003 : 359).

primitives sont en effet des catégories sémantiques de base. L'avantage, d'après Schank, est que, même dans le cas où la syntaxe d'une phrase n'est pas correcte, c'est-à-dire même lorsqu'elle présente un écart par rapport à la norme, ce type de théorie pourrait aboutir à une application automatique pour la compréhension et la traduction des textes. Néanmoins, ce traitement automatique des textes a jusqu'à maintenant rencontré des échecs partiels et il reste encore difficile d'appréhender un texte dans sa totalité.

D'autres grammaires prenant en compte le contexte se développent à peu près à la même époque. Ce sont des grammaires (Halliday 1973) qui ne considèrent pas le langage comme un système à part mais intègrent au traitement le contexte d'utilisation : elles se sont particulièrement intéressé à l'organisation fonctionnelle du langage et aux liens qui relient la forme d'un texte au contexte. Il s'agit donc de grammaires systémiques dans la mesure où elles intègrent les données contextuelles dans un système global. Ces grammaires confèrent aux phrases un ensemble de traits qui peuvent ensuite être utilisés dans d'autres configurations contextuelles. Ainsi, elles s'ancrent dans les grammaires descriptives et ne sont pas des grammaires génératives. Le passage de traits d'une procédure à une autre est l'aspect le plus caractéristique de ces grammaires pour l'intelligence artificielle. Tout en gardant une certaine modularité, le procédé en question intègre les aspects contextuels et prend des décisions qui tiennent compte de plusieurs processus interactifs ; il se réalise ainsi aisément.

Le fait que la spécification des règles syntaxiques soit beaucoup plus aisée et plus économique que dans les grammaires formelles procure aux grammaires systémiques un atout majeur dans la mesure où il rend beaucoup plus fiable l'interprétation. De plus, pour le traitement d'un même phénomène, le niveau d'abstraction de ces grammaires donne la possibilité de comparer efficacement les mécanismes divers auxquels recourent des langues différentes. Ces grammaires sont très intéressantes pour l'intelligence artificielle sur les plans théorique et pratique. Sur un plan théorique, elles permettent une approche de la sémantique et même de la pragmatique, notamment pour tout ce qui est en lien avec la gestion des dialogues homme-machine, qui rend indispensable l'intelligence artificielle. Sur le plan pratique, l'application des techniques de gestion d'arbres est très facile dans ces grammaires puisque ces derniers sont particulièrement bien maîtrisées en informatique.

Dans la description des expressions linguistiques et de façon similaire aux grammaires systémiques, les grammaires fonctionnelles, représentées en particulier par Kay, Bresman ou Kaplan, accordent un rôle mineur aux notions catégorielles des

grammaires formelles au profit des aspects fonctionnels et relationnels. De même, la langue n'est pas ici perçue comme la description statique d'un ensemble de phrases mais comme un outil d'interaction sociale. L'utilisation qui est faite d'une langue est privilégiée par rapport à la compétence théorique : ainsi, plus que la seule expression de pensées, sa fonction principale est la communication. Les connaissances lexicales, les connaissances sur les structures et les règles de grammaire sont alors considérées de façon uniforme par les grammaires fonctionnelles des expressions de contraintes. En se basant dans une certaine mesure sur l'idée d'intégration de descriptions partielles et dans sa tentative de développer un formalisme unique qui prendrait en considération tous ces aspects, Kay aboutit à la notion de description fonctionnelle puis, à la notion de « grammaire d'unification ». Bresnan/Kaplan (1981), quant à eux, développent les grammaires lexicales fonctionnelles en s'appuyant sur les mêmes idées que Kay :

> Bresnan et Kaplan (1981) ont suivi, à partir des mêmes idées, une autre direction et ont développé les *grammaires **lexicales** fonctionnelles* qui utilisent la notion d'équations simultanées permettant d'interpréter sémantiquement une structure construite par une grammaire non contextuelle. (Sabah 1996 : 21).

Ces grammaires recourent à la notion d'équations simultanées qui permet d'expliciter une structure construite par une grammaire non contextuelle. Cette théorie cherche surtout à formaliser la connaissance syntaxique nécessaire à la définition des relations entre les aspects sémantiques prédicatifs importants dans le sens d'une phrase et les choix des mots et des structures des phrases qui pourraient exprimer ces relations.

Les théories fonctionnelles intègrent une « lexicalisation » de la grammaire. Abeillé (1993) a procédé à une synthèse des grammaires modernes lexicalisées. On peut citer notamment le modèle de grammaires HPSG[153] où les structures à caractère complexe ont un rôle omniprésent, dans le lexique, les règles et les représentations construites sont essentiellement remplacées par des possibilités de combinaisons d'arbres élémentaires issus du lexique. Estival (1994) décrit l'ouvrage d'Abeillé comme une vulgarisation de l'informatique pour les linguistes. Il le présente de la manière suivante :

> Dans cet ouvrage, A. A. veut présenter au public français les quatre théories syntaxiques qui dominent depuis quelques années le domaine de la linguistique formelle, en particulier dans ses applications en linguistique informatique, à savoir la grammaire lexicale fonctionnelle (LFG), la grammaire syntagmatique généralisée (GPSG), la grammaire syntagmatique généralisée guidée par les têtes (HPSG) et la grammaire d'arbres adjoints (TAG). L'ouvrage

[153]HPSG *: Head-driven Phrase structure Grammar*, parfois traduit « Grammaire syntagmatique guidée par les têtes ». Pour une présentation en français du modèle, *cf.* Bonami et Godard (2001 : 166-174).

> s'adresse à des linguistes, sans présupposer qu'ils aient une formation mathématique ou informatique. (Estival 1994 : 42)

Grâce à l'introduction de notions de sémantique formelle, les grammaires lexicales fonctionnelles représentent une des théories linguistiques les plus développées. Elles présentent de nombreuses applications en intelligence artificielle dans les années 1980 et 1990. Elles diffèrent néanmoins des grammaires transformationnelles sur les points suivants :

- Aspects psychologiques. La théorie de Kaplan/ Bresnan (1981) accepte des aspects beaucoup plus divers que ceux des linguistes qui ont suivi Chomsky. En effet, quand ces derniers recherchent la définition d'une grammaire universelle propre à la faculté du langage chez l'homme (précisant dans quelle mesure les hommes, et eux seuls, sont capables d'apprendre une langue), la théorie de Kaplan et Bresnan explicite la manière dont les facultés langagières interagissent avec d'autres processus mentaux lors des processus de compréhension et de production du langage ;

- Structures et fonctions : Contrairement aux grammaires fonctionnelles où la fonction grammaticale prime sur la structure, dans une grammaire transformationnelle, c'est la notion de structure qui est primordiale et les rôles grammaticaux des divers constituants en sont déduits.

- Rôle du lexique : Tout comme les grammaires systémiques, les grammaires fonctionnelles visent à expliciter les articulations entre les niveaux lexicaux, syntaxiques et sémantiques. Elles donnent généralement naissance à des modèles plus riches et plus souples que les grammaires génératives d'un point de vue informatique.

Fondamentalement, dans ces grammaires, chaque forme de mot peut accepter plusieurs entrées distinctes si celles-ci jouent des rôles différents. Dans les grammaires fonctionnelles le passif est abordé par l'intermédiaire du lexique (une entrée spéciale correspondra, pour chaque verbe, à l'utilisation passive du participe passé) alors qu'il est traité par la grammaire dans les grammaires transformationnelles.

Gross (1996) va dans le même sens d'une critique des grammaires transformationnelles en mettant en évidence leur incapacité à rendre compte de tous les phénomènes relevés dans les langues. En effet, selon lui, aucune de ces théories n'est construite à partir du recensement exhaustif des faits à expliquer. Ainsi en n'introduisant dans le langage que les abstractions nécessaires à l'explication des observations et en s'appuyant sur un modèle aussi limité que possible, il accorde une importance primordiale aux faits attestés. Il aboutit ainsi essentiellement à une constitution de lexiques-grammaires

explicitant les possibilités de combinaisons des mots entre eux (l'hypothèse de base de Gross étant qu'aucune convergence ni règle générale n'est possible, il faut relever tous les exemples possibles...). Ainsi, dans une optique différente, Gross tente également d'intégrer les phénomènes syntaxiques dans les caractéristiques lexicales des mots considérés :

> C'est d'ailleurs à partir de ces contributions à un domaine appliqué de la linguistique que Gross va ouvrir à partir des années 80 de nouvelles brèches en linguistique théorique, en dévoilant des domaines auxquels la tradition n'avait guère accordé d'importance et qui pourtant s'avéreront d'un intérêt fondamental, notamment les expressions figées, les verbes supports, le traitement des incises dans le discours ou encore, l'incroyable polysémie du langage naturel. (Lamiroy 2003 : 145)

En résumé, puisque les sciences cognitives visent à rendre indissociables la forme et le sens, c'est-à-dire à rendre explicite la façon dont la forme reflète les notions sémantiques, il est indéniable que ces travaux présentent un intérêt considérable pour cette discipline (même si la sémantique n'est pas toujours au centre de leurs préoccupations). Il est donc intéressant de mentionner toutes ces grammaires intégrant la dimension sémantique qui, plus qu'un outil linguistique, sont dans une certaine mesure une façon pragmatique de mettre en place des programmes efficaces. En se basant essentiellement sur le genre et la sphère textuelle (textes littéraires, textes scientifiques, etc), les concepteurs élaborent des classes sémantiques qui représentent les parties non terminales de la grammaire. Ainsi, sans se préoccuper davantage de la syntaxe, qui reste implicite et correspond simplement à l'ordre dans lequel peuvent apparaître les diverses catégories sémantiques, le traitement d'une phrase correspond à une vérification des correspondances sémantiques avec la grammaire. Même si le produit final ne se prête pas facilement à d'autres domaines et que le manque de nuances rend la compréhension assez limitée, ces grammaires ont l'avantage, dans des domaines restreints, d'être très souples et faciles à mettre en œuvre. En effet, elles peuvent traiter des énoncés non normés.

L'évolution des théories imprégnant le traitement automatique des langues s'est accompagnée de tentatives d'application dans le traitement du sens en informatique. C'est ce que nous allons à présent aborder, à travers les mots-clefs, même si les systèmes proposés sont peu satisfaisants. En se fondant sur l'utilisation de mots clefs, les disciples de Minsky (1974) au MIT développent divers systèmes traitant des textes en anglais et relancent les recherches sur la compréhension automatique du langage. Nous pouvons noter ces programmes même si un grand nombre d'entre eux ne sont valables que dans des domaines très limités et n'autorisent qu'une syntaxe très pauvre afin de tenter de

contourner les difficultés (surtout les domaines scientifiques comme la médecine, les mathématiques, etc.) car aucun problème complexe relatif au langage n'est réellement abordé. En effet, ces systèmes rudimentaires supposaient qu'il suffisait de quelques mots et règles syntaxiques pour exécuter certaines taches utilisant le langage (surtout lorsque les questions sont posées en anglais).

Pour une application donnée et afin de mettre en place un programme qui exécutera les actions appropriées à partir des mots clefs identifiés (par exemple, construire une requête pour une base de données), il faudra d'abord lister les mots clefs de ce domaine et, ensuite, écrire un analyseur qui pourra les filtrer en considérant leurs variantes morphologiques. Parmi ces programmes, les plus connus sont : BASEBALL (Green *et al.* 1961) - réponses à des questions avec une base de données -, STUDENT (Bobrow 1968) - résolvant des exercices d'algèbre élémentaire - et ELIZA (Weizenbaum 1966) - modèle des systèmes « écholaliques ». Néanmoins, ces programmes restent limités aux mots et ne traitent que de considérations syntaxiques très élémentaires. Ils illustrent bien les limites de la technique des mots clefs puisque des complications apparaissent si on veut établir la liste complète des mots clefs pertinents d'un domaine plus étendu. Il apparait donc nécessaire de recourir à des méthodes plus élaborées quand par exemple un mot clef a plusieurs significations ou que l'absence d'un mot clef est significative.

2. Intelligence artificielle et communication homme-machine

Tout d'abord, les traitements purement formels ne tiennent pas compte du caractère fluide des langues, qui est à prendre en considération. Ainsi, afin de garantir la fluidité de la communication, il serait plus judicieux de faire face à ce problème plutôt que de le considérer comme un problème à résoudre par une limitation adéquate des domaines d'application (ce qui est la position de nombre d'approches actuelles). En fait, il est pour ainsi dire pratiquement impossible, sauf à quelques exceptions près, de définir des sous-langages limitatifs qui conservent cette flexibilité. Cela signifie en quelque sorte qu'il faut permettre l'usage de tous les phénomènes de la langue (des anaphores aux métaphores et métonymies, en passant par les ellipses, déictiques...).

Même s'il semble souhaitable de mettre en place une sémantique objective et universelle dans laquelle on souhaiterait développer des méthodes qui soient facilement généralisables et qui considèreraient les connaissances comme des axiomes, cette sémantique n'est pas très utile. Pour qu'une intelligence (artificielle ou non) soit acceptée comme telle par l'être humain, il est impératif qu'elle lui paraisse analogue à la sienne. En

effet, il est impératif que la notion de sens que véhicule une entité avec laquelle nous communiquons soit vectrice de sens pour l'humain.

Le système exploité doit présenter des connaissances relatives à la cognition humaine (il est certain que le système doit avoir une bonne représentation de son interlocuteur et de son fonctionnement pour dialoguer efficacement) afin de permettre une facilité et une accessibilité des interprétations mises en place par la machine : en d'autres termes, leur réponse aux attentes des utilisateurs. On utilisera ce que l'on connait du fonctionnement cognitif humain face au langage analogiquement au système lui-même (même si ce n'est pas une obligation, c'est souvent une source d'inspiration extrêmement utile) grâce à l'accessibilité des connaissances. Il n'est pas toujours aisé de décrire une langue d'une façon algorithmique.

C'est un mécanisme prédictif techniquement très différent des analyses classiques, réalisé par des processus totalement automatiques (c'est-à-dire non contrôlés, ni réflexifs). En effet, le contexte doit diriger le système vers une interprétation résultante, souvent unique car l'énoncé en cours de traitement peut se prêter à plusieurs interprétations construites en parallèle. L'état du contexte s'apparente alors à un ensemble d'hypothèses qui permet l'élargissement du nombre des interprétations les plus cohérentes.

Afin de garantir la construction de modèles du langage plus abstraits, il serait préférable de fonder le processus d'interprétation sur des principes généraux comme : « l'attachement minimal » (ne pas retenir des nœuds de l'arbre syntaxique potentiellement inutiles) et de la « clôture différée » (relier les nouveaux éléments au syntagme en cours de traitement). Par exemple, en analyse syntaxique, les interprétations les plus retenues généralement sont celles qui se conforment à ces principes. Malgré le fait qu'isoler les éléments de jugement pertinents (si l'étude statistique de corpus permet de révéler les règles générales, elle ne donne néanmoins pas le moyen de traiter les cas particuliers) est une tâche assez délicate, il faudrait mettre en place un moyen d'identifier leur exception. Cela expliquerait ainsi le fait que ces régularités ne peuvent pas être utilisées comme des règles formelles d'analyse. En revanche, il est plus facile de les expliquer comme un effet découlant de l'organisation concurrentielle des processus interprétatifs : les interprétations qui vérifient l'attachement minimal et la clôture différée sont généralement les plus simples à réaliser et donc les premières à être perçues.

Même si cela dépend nécessairement d'une perception spontanée du sens (cette distinction permet de différencier les « vraies » ambiguïtés relevées par la communication,

qu'une planification dynamique devrait résoudre, et les ambiguïtés artificielles, qui restent imperceptibles sans une étude linguistique approfondie), la pensée rationnelle joue un rôle très important dans la compréhension. Ce deuxième aspect autocontrôlé et planifié permet, en particulier, le traitement de tous les imprévus et aboutit à un apprentissage de nouvelles connaissances et de nouveaux processus. Une véritable compréhension doit préciser le rôle de l'apprentissage dans l'appropriation de la langue et engage une comparaison continue des énoncés reçus et des connaissances antérieures.

Les linguistiques cognitives sont un outil très efficace pour l'intelligence artificielle et constituent ainsi une approche ciblant son intégration (Sabah 2004). Dans ce contexte, on peut poser la question de la structure prédéfinie, et donc de l'architecture qui permettrait à ces divers niveaux de connaissances de collaborer de façon cohérente. Nous en venons ainsi à présenter certains éléments de réflexion sur l'aménagement de l'architecture générale de RESUMAN, l'outil créé pour résoudre les anaphores pronominales ambiguës de notre corpus. Le caractère divers et complexe des connaissances indispensables à un système de compréhension automatique des langues a été mis en relief par de nombreuses études menées sur le langage. Les questions essentielles sont les suivantes : comment toutes les sources de connaissances collaborent-elles ? Quelles relations entretiennent-elles les unes avec les autres ? Quelles architectures informatiques permettent de les mettre en œuvre de la façon la plus efficiente possible ?

Même si un ordre précis des opérations à effectuer ne peut être valable dans tous les cas, rappelons que les tous premiers programmes de traitement automatique des langues recoururent à des architectures en série qui imposaient des communications fixes et limitées entre les modules. Ainsi, même si la mise en pratique de RESUMAN reste difficile, au stade où nous en sommes, il apparaît donc qu'une certaine intégration des différents types de connaissance (permettant l'utilisation simultanée, dans un seul module, de toutes les connaissances est préconisée. Néanmoins, les restructurations s'avèrent relativement difficiles. En effet, d'une part, il est nécessaire d'expliciter, dans les règles mêmes de traitement, le mode d'interaction des diverses connaissances et d'autre part, il n'existe encore aucune théorie linguistique qui intègre réellement toutes les connaissances essentielles à la compréhension.

L'interactivité de différentes sources de connaissances plus ou moins indépendantes permet une expression plus claire des connaissances : les connaissances de même nature seraient réunies dans des modules qui coopèrent ; l'utilisation des connaissances ne dépend

pas des connaissances elles-mêmes, elle est gérée indépendamment par un contrôleur. Cette interactivité s'avère donc pratiquement nécessaire. Nous pourrions alors aboutir à une structure de mémoire élaborée permettant de relier des processus automatiques et des processus réflexifs, en d'autres termes, à des architectures modernes de systèmes multi-agents qui doivent être augmentés par la prise en considération de ces deux processus.

Les approches présentées précédemment semblent se rencontrer dans le fait qu'elles se situent toutes dans le cadre d'une intelligence formelle sans rapport direct avec les perceptions du monde dans lequel elle évolue. Il devient nécessaire d'utiliser un large éventail de connaissances, de mécanismes de planification et de prise de décision, et une mémoire permettant un stockage et une recherche très efficiente (ce qui représente d'ailleurs l'essentiel des travaux d'intelligence artificielle et de traitement automatique des langues, qui, soulignons-le, n'ont pas réellement essayé de trouver d'autre solution). En effet, l'étude des interactions avec les autres phénomènes est mise au second plan et le langage et les phénomènes de compréhension sont tellement complexes qu'ils sont étudiés de façon isolée. Les caractéristiques suivantes de nos connaissances expliquent que les raisonnements et les mécanismes de planifications, basés essentiellement sur le raisonnement formel, sont confrontés à un certain nombre de problèmes :

- leur caractère incomplet : il est impossible de savoir tout ce qui est pertinent dans une tâche réelle ;

- le manque de précision : il est difficile de connaître avec une précision exacte toutes les variables pertinentes ;

- la variabilité : ce qui est valable à un moment donné ne l'est pas forcément au moment de l'action, il n'est pas sûr d'avoir une représentation exacte du monde car il est pratiquement impossible d'énumérer tout ce qui n'est pas modifié par une action (le fameux *frame problem*) ;

- le temps d'accès : même si on pouvait tout savoir, en pratique, cela entraînerait un fonctionnement impossible de la mémoire (stockage, recherche et calcul).

D'autre part, la mémoire humaine, influencée par les choses importantes pour la vie, est prospective et réflexive et n'est pas seulement associative. Grâce à elle nous organisons le monde en oubliant ce qui est inutile : le monde, avec tout ce qui le compose d'aléas ne manque pas de nous les révéler selon les besoins (loin d'être une organisation a priori, ce sont plutôt des moyens d'accéder directement à l'information requise : comment une

description peut-elle nous permettre de naviguer dans la mémoire et de retrouver très rapidement les éléments congruents ?).

Deux précisions sont, néanmoins, obligatoires : tout d'abord, les machines ne peuvent pas bénéficier des sources d'information les plus riches à cause de la désincarnation. Ensuite, le mécanisme de la catégorisation est capital pour tout ce qui est en rapport avec le langage : il apparaît essentiel d'établir des relations entre objets pour structurer le monde. On met en place des classes d'objets identiques, on les nomme et l'on construit à nouveau des classes linguistiques au niveau méta pour ces nouveaux éléments.

Par ailleurs, lors de la résolution d'un problème de façon collective, une communication élaborée (même dans le seul cadre de l'intelligence artificielle distribuée) est indispensable car des idées nouvelles naissent autant de l'interaction que d'un individu seul[154]. Ainsi, le langage, l'apprentissage et la communication jouent un rôle déterminant dans l'exploitation des connaissances déjà accumulées par des siècles d'expériences précédentes. Avantageuse en tous points, cette communication nous permet d'acquérir un savoir par l'expérience, mais aussi par la communication, par la lecture. Il est donc important de manifester une disposition à apprendre différemment selon les circonstances. Une véritable intelligence artificielle doit donc être capable d'évaluer et de modifier ses propres programmes.

En résumé, l'ensemble des sciences cognitives sont pertinentes pour rendre compte de l'activité de traitement du sens : attribuer un sens aux expressions linguistiques n'est qu'une activité dans la foulée de tant d'autres. Cela interagit avec tous les aspects de la cognition et de la vie sociale comme les activités de perception, de raisonnement, de mémorisation. Les approches symboliques s'inspirent directement de l'hypothèse de représentations mentales. Ceci nous conduit, en dépit de leurs grandes différences avec les représentations des humains, à une analogie très étroite entre les représentations déjà acquises et les représentations de l'intelligence artificielle. Celles-ci présentent de nombreuses similitudes à un certain niveau de description. En impliquant un niveau d'analyse totalement séparé du niveau neurobiologique comme du niveau sociologique et culturel, cette hypothèse est concluante.

Pour ce qui est de la compréhension du langage, il est judicieux de souligner l'importance de la notion de réflexivité (la capacité pour un système de raisonner sur son propre comportement), tout en gardant ce cadre purement symbolique. On ne peut pas nier

[154] Dans notre cas, nous travaillerons avec une informaticienne pour la création de RESUMAN.

les similarités avec des idées qui apparaissent dans le domaine de la métacognition, même s'il est vrai que ces programmes sont loin d'être un modèle du fonctionnement de la conscience. Lorsqu'il est question de l'aspect fonctionnel du contrôle, on peut relever une certaine ressemblance entre le modèle réparti et réflexif et une certaine conception de la conscience, et ceci malgré les nombreuses différences dues à l'organisation et aux distinctions entre les composants des machines humaine et informatique. Une qualité déterminante de l'intelligence qui doit être impérativement prise en compte par les programmes d'intelligence artificielle est sa caractéristique d'autoreprésentation et d'autoréférence.

En résumé, le point de vue purement symbolique est possiblement remis en cause. En effet, grâce à des techniques efficaces pour le traitement des informations incertaines ou ambigües, les recherches travaillent actuellement à pourvoir quelques éléments convergents vers cet objectif. Les collaborations entre les linguistes et les informaticiens sont assez prometteuses. A plus forte raison, comme l'être humain pour qui le langage est l'outil essentiel qui lui permet d'exprimer ses compétences cognitives, on peut dire que le traitement automatique des langues et la communication homme-machine doivent travailler à l'essor de bases nécessaires pour tous les autres processus de raisonnement.

Chapitre 2 : Systèmes de résolution automatique d'anaphores

Dans les années 90, il y a eu, suite aux développements de la linguistique informatique, le développement de systèmes de résolution automatique des pronoms, effectivement implantés en machine et capables de fonctionner sur des textes quelconques. Plusieurs travaux scientifiques portaient sur les algorithmes de résolution de l'anaphore (dit aussi calcul de la référence ou calcul référentiel) afin d'améliorer les performances des algorithmes proposés. Nous présentons ici quelques-uns de ces systèmes puisque l'objectif du présent travail est d'implanter un tel système pour le français avec comme support textuel $RESUMAN_c$. Nous nous intéresserons en premier lieu aux systèmes qui ont été effectivement évalués, puisque l'évaluation est le seul moyen de juger de la pertinence des informations utilisées[155].

Nous allons présenter, dans ce premier chapitre, les principaux algorithmes de résolution de l'anaphore trouvés dans la littérature, les mieux documentés et les plus complètement évalués. Nous évoquerons ces systèmes en détaillant leur fonctionnement, performances et en déterminant leurs points faibles. Nous les répartirons en deux types : les systèmes qui utilisent principalement des connaissances linguistiques pour identifier l'antécédent d'une anaphore donnée comme celui de Hobbs[156] (1976), Lappin/Leass (1994), Grosz *et al.,* (1995), Kennedy/Boguraev (1996), Baldwin (1997), Mitkov (1998) et quelques systèmes français[157].

Nous présenterons ensuite les systèmes d'apprentissage statistique comme celui de Conolly *et al.*, (1994), Ge *et al.*, (1998) et BART (2010). Enfin, nous discuterons les limites de ces systèmes à travers lesquelles nous nous positionnerons pour donner notre algorithme général en fin de ce chapitre.

---

[155] Ces considérations nous ont conduit à mettre de côté bon nombre de travaux, en particulier :

- Les premiers systèmes effectivement implantés, qui s'appliquent à des domaines et langages très restreints, par exemple le système SHRDLU de Winograd (1972) restreint à un univers de blocs, le système de Günthner et Lehmann (1983), restreint à l'interrogation d'une base de données ;
- Les propositions qui se concentrent plus sur une architecture que sur la définition effective de règles de résolution, par exemple les propositions de Carbonnel et Brown (1998) ou encore Byron et Tetreault (1999) ;
- Bon nombre de systèmes que nous jugeons insuffisamment documentés, en l'occurrence, les systèmes utilisés par les participants aux conferences MUC *(Message Understanding Conferences)* (1995).

[156] Il est le seul qui n'a pas donné lieu à une implémentation par son auteur, nous l'évoquerons tout de même car il constitue classiquement une référence à laquelle on compare des approches plus récentes.

[157] A notre connaissance, il n'y en a pas beaucoup.

1. Systèmes à base des connaissances linguistiques[158]

Les auteurs des premiers systèmes de résolution de l'anaphore pronominale essayaient d'exploiter des connaissances linguistiques complexes comme les contraintes syntaxiques et sémantiques régissant l'anaphore. Afin de résoudre une anaphore donnée, ils ont eu recours à des règles morphologiques, syntaxiques et/ou sémantiques développées à partir d'un corpus bien précis. Nous discuterons dans cette section les principaux systèmes à base de connaissances linguistiques.

1.1. Systèmes anglais

1.1.1. Systèmes exploitant des connaissances linguistiques profondes

1.1.1.1. Algorithme de Hobbs (1976)

Le premier algorithme de Hobbs, nommé aussi approche naïve de Hobbs (*Hobbs's naive approach*), est l'un des premiers travaux sur la résolution des anaphores. Il est qualifié de naïf car il utilise une approche purement syntaxique. Il prend en compte les contraintes imposées par la syntaxe de l'anglais. En effet, il résout les relations anaphoriques pour les pronoms *he, she, it* et *they* ainsi que pour les pronoms possessifs. L'auteur ne soulève pas les figures d'emploi du pronom *it* qui peut être impersonnel et les relations anaphoriques *clausales* pour ce même pronom. Tout en tenant compte de l'accord grammatical entre pronom et antécédents, au niveau de cet algorithme, c'est grâce au parcours de l'arbre syntaxique que se fait la recherche de l'antécédent d'un pronom. La recherche commence par un parcours en largeur qui traduit la préférence des constituants immédiats sur les constituants emboîtés et elle est effectuée de gauche à droite ce qui traduit la préférence du sujet sur l'objet. Il existe une préférence pour des antécédents proches plutôt que pour ceux qui sont éloignés. En effet, l'algorithme parcourt les arbres syntaxiques des phrases précédentes, en commençant par celle qui précède immédiatement, quand aucun antécédent n'est trouvé dans la phrase courante.

Cet algorithme suppose une connaissance parfaite de la structure syntaxique résultant de l'application d'une grammaire contextuelle. La structure de l'arbre doit correspondre à la structure grammaticale réelle de la phrase. Tous les syntagmes doivent apparaître, comme les anaphores zéro qui doivent figurer dans la structure, et leurs liens de dépendance doivent être corrects. Un syntagme nominal concordant avec le genre et le nombre du pronom est recherché par l'algorithme qui explore toute la structure. Cet

[158] Ou *Knowledge-based approaches*

algorithme traite, selon une préférence dérivée de la linéarité du discours, les anaphores interphrastiques et intègre des contraintes syntaxiques sur la coréférence pronominale intraphrastique. En privilégiant l'antécédent le plus proche de l'anaphore, il parcourt en largeur de gauche à droite, pour ce qui est du niveau intraphrastique. En privilégiant les sujets comme antécédents, il parcourt également en largeur au niveau interphrastique. L'algorithme répertorie, au niveau de ce parcours, les antécédents possibles qu'il vérifie ensuite en appliquant des contraintes d'accord morphologique (traits de genre et nombre). Par ailleurs, il soumet des contraintes syntaxiques basées sur la condition B de la théorie du liage : ne peuvent apparaître dans la même phrase simple un pronom non réfléchi et son antécédent et l'antécédent doit figurer avant le pronom ou le commander.

Un travail d'évaluation mené manuellement sur 300 textes constitués de récits historiques, de romans littéraires et d'articles journalistiques montre que l'algorithme de Hobbs a eu un taux de succès[159] de 88,3%. Ses performances frôlent les 92% de succès sur la totalité des anaphores et 82% en cas de simples anaphores « non-triviales ». Ainsi cet algorithme est couramment utilisé par la communauté, celui-ci ayant servi et sert encore de référence dans l'évaluation des systèmes de résolution. Cependant, cet algorithme échoue sur certains cas de reprise d'éléments phrastiques tels que :

[1] Pierre avait des ennuis$_i$ et il le$_i$ savait.

Par ailleurs, l'algorithme recherche également des antécédents de pronoms qui apparaissent après le pronom. Il ne cherche pas toutefois en dessous des niveaux des groupes nominaux ou des propositions. Ainsi, en dépit de ses résultats satisfaisants, l'approche de Hobbs ne peut pas résoudre les pronoms dans certaines constructions qui contiennent une anaphore événementielle comme dans [1]. Des approches plus récentes ont pu parfaire la couverture du système de résolution, citons par exemple le système RAP (*Resolution of Anaphora Procedure*[160]), détaillé dans la sous-section suivante, basé sur la syntaxe (Lappin/Leass, 1994) et consacré à la résolution du pronom anaphorique de la 3ème personne.

[159]Le taux de succès est donné par le rapport entre le nombre de bonnes réponses et le nombre total de réponses retournées.
[160] Procédure de Résolution d'Anaphores.

1.1.1.2. Algorithme de Lappin/Leass (1994)

L'algorithme de Lappin/Leass (1994), ou RAP, identifie les antécédents des pronoms de la 3ème personne (*he, she, they, it*) et d'anaphores réflexives et réciproques (*himself, herself, themselves, itself*) en anglais. Il dépasse la performance de l'algorithme naïf de Hobbs de 4%. Pour cela, l'algorithme exploite des informations de nature syntaxique et morphologique et il se base sur une mesure de saillance[161] à partir de la structure syntaxique et d'un modèle attentionnel. Afin de restreindre les candidats potentiels, l'algorithme les soumet à trois filtres. Tout d'abord, il les soumet à un filtre relatif à la cohérence morphologique entre l'anaphore et le candidat. Ensuite, afin de dégager les dépendances syntaxiques entre le pronom et certains candidats, il les soumet à un filtre relatif à la structure syntaxique de la phrase contenant le pronom. Enfin, un filtre relatif aux pronoms réflexifs (*himself, myself...*) et réciproques (*each other, one other...*) est appliqué. Les candidats restants, après exclusion des pronoms non-anaphoriques[162], seront évalués selon un poids de saillance calculé selon les critères suivants :

- la distance du candidat (par rapport au pronom),
- son rôle syntaxique,
- l'ordre de préférence Sujet>COD>COI,
- son appartenance dans les syntagmes nominaux et adverbiaux.

En effet, en prenant en compte ces différents facteurs, il fait recourt à un modèle permettant d'évaluer d'une manière efficiente la saillance d'un antécédent potentiel. Ces facteurs comportent chacun un indice différent en fonction de leur utilité dans la résolution. Des classes d'équivalence regroupent tous les référents qui constituent une chaîne anaphorique. On attribue à ces classes un poids en rapport avec la somme de tous les facteurs de saillance relatifs à au moins un membre de la classe d'équivalence.

Les valeurs de saillance des antécédents potentiels sont réduites conformément à certaines règles au fur et à mesure de la résolution. C'est l'antécédent le plus saillant qui est choisi comme le référent d'un pronom dans une classe d'équivalence et c'est l'antécédent le plus proche du pronom qui est retenu dans les situations où deux ou plusieurs d'entre eux ont la même mesure de saillance. Comparativement aux antécédents interphrastiques, les

[161] Cet algorithme mérite qu'on s'y arrête dans la mesure où notre propre système fait usage d'une procédure et/ou d'une information similaire.

[162] Les pronoms pléonastiques sont aussi mis à l'écart, comme *it* dans des constructions avec un adjectif modal: *it is necessary/essential/sufficient to/that.* Les cas où *it* est avec un verbe cognitif ou d'attitude propositionnelle sont aussi écartés : *it is recommended/ believed/ known/ expected that.*

antécédents intraphrastiques sont privilégiés. Les propriétés structurales ou syntaxiques représentent les facteurs de saillance principalement utilisés dans l'algorithme. Selon sa pertinence, chaque facteur permet d'augmenter le score des antécédents potentiels comme décrit ci-dessous (*cf.* Figure 26) :

- **Récence de la phrase** (*Sentence recency*) (100) **:** la saillance de l'antécédent est grande quand la proximité avec l'anaphore est grande. Pour chaque nouvelle phrase, la valeur de ce facteur est diminuée de moitié.

- **Emphase sur le sujet** (*Subject emphasis*) (80) **:** la saillance dépend du rôle grammatical : un antécédent en position sujet est le plus saillant.

- **Emphase existentielle** (*Existential emphasis*) (70) **:** un élément nominal dans une construction existentielle (« *There is...* ») est saillant.

- **Emphase accusative** (*Accusative emphasis*) (50) **:** un élément objet direct est saillant, mais moins qu'un sujet.

- **Objet indirect** (*Indirect object, oblique complement emphasis*) (40) **:** un élément objet direct est plus saillant qu'un objet indirect[163].

- **Tête d'un SN** (*Head noun emphasis*) (80) : dans un groupe nominal complexe, un nom en position de tête nominale est plus saillant. Un nom non tête est pénalisé.

- **Emphase non adverbiale** (*Non adverbial emphasis*) (50) : dans des constructions adverbiales, il pénalise les SN.

Condition	V
SN_i est dans la phrase courante	$+100$
SN_i est sujet	$+80$
SN_i apparaît dans une construction existentielle (ex. *There are* SN_i)	$+70$
SN_i est complément d'objet direct	$+50$
SN_i est complément d'objet indirect ou oblique	$+40$
SN_i n'est pas enchâssé dans un autre syntagme nominal	$+80$
SN_i n'est pas enchâssé dans un syntagme prépositionnel adverbial	$+50$
SN_i suit P_i (cataphore)	-175
SN_i et P_i occupent la même fonction (parallélisme des fonctions)	$+35$

Figure 26 : Scores de saillance en fonctions des critères syntaxiques selon Lappin/Leass (1994)

[163] *Cf.* deuxième chapitre de la deuxième partie.

Pour résumer, cet algorithme a pour caractéristiques principales :

- Comparativement à l'algorithme de Hobbs, il possède des filtres syntaxiques et morphologiques plus fins,
- Des pronoms pléonastiques y sont pris en compte,
- Un calcul de la saillance qui n'est pas seulement basé sur la linéarité mais aussi sur d'autres critères syntaxiques,
- En cas d'ambiguïté sur la saillance, le choix sur des critères de proximité,
- La préférence des anaphores intra- phrastiques sur les anaphores inter-phrastiques
- La préférence des anaphores sur les cataphores,
- Il n'exige pas des connaissances sémantiques ou pragmatiques, comme l'algorithme naïf.

Afin d'évaluer les performances des deux approches, les auteurs ont adapté l'algorithme de l'approche naïve de Hobbs dans leur analyseur syntaxique. Le système a été appliqué sur un corpus composé de 355 phrases extraites aléatoirement de 47 manuels d'informatique. Les résultats démontrent une meilleure performance du système RAP avec 86% des anaphores résolues contre 82% pour leur adaptation à l'algorithme de Hobbs. Les auteurs l'expliquent par un meilleur traitement des anaphores intra-phrastiques qui constituent plus de 80% des pronoms du corpus (*Cf.* Tableau 1). Le corpus utilisé pour les tests compte peu d'anaphores inter-phrastiques (20%), ainsi, ces résultats restent à reconsidérer même s'ils sont satisfaisants. Par ailleurs, le taux de performance de ce système sur ce type d'anaphore est inférieur à 74% contre 87% pour l'algorithme de Hobbs.

	Total	**Interphrastique**	**Intraphrastique**
Taux de succès pour l'algorithme RAP	86%(310)	74% (52)	89%(258)
Taux de succès pour l'algorithme naïf (Hobbs)	82%(295)	87%(61)	81%(234)
Nombre d'occurrences de pronoms	360	70	290

Tableau 12 : Comparaison entre les algorithmes Hobbs et RAP

1. créer une liste pour tous les GN de la phase courante
2. pour chaque GN :

Classification selon le type {indéfini, défini, pronom pléonastique, pronom personnel, pronom réflexif. etc.}

si GN= {indéfini ou défini} alors

calcul de la saillance

sinon si GN= pronom réflexive alors

calculer une liste de paires {GN; pronom} pour lesquelles il peut y avoir coréférence : si plusieurs possibilités, l'antécédent est choisi sur le critère de saillance des GN

sinon

si GN= pronom personnel de 3ème personne alors

- calculer une liste de paires {GN: pronom} pour lesquelles il ne peut pas y avoir coréférence
- créer la liste des antécédents possibles : celle-ci contient le référent discursif le plus récent pour chaque classe d'équivalence avec son facteur de saillance
- modifications locales de la saillance : si le GN antécédent se trouve après le pronom, la saillance décroît (pénalisation des cataphores). Si le GN antécédent a le même rôle grammatical que le pronom, sa saillance augmente
- définir un seuil de saillance et filtrage
- appliquer le filtre morphologique si plusieurs candidats alors

choix sur la saillance du GN antécédent si égalité de saillance du GN antécédent choix selon proximité

fsi

fsi

fsi

fsi

fin pour chaque GN

Figure 27: Algorithme RAP de Lappin et Leass (1994)

L'approche à base de connaissances linguistiques profondes s'avère assez coûteuse en temps et en code. En effet, l'algorithme RAP et celui de Hobbs, nécessitent des niveaux d'analyse variés, notamment syntaxiques. De plus, il faut leur fournir, comme entrée, un corpus parfaitement analysé syntaxiquement[164] et dont l'influence de chaque élément contextuel sur le poids de la saillance a été traité au préalable. Pour parer à cette difficulté, certains auteurs ont proposé des approches qui ne se basaient pas sur une analyse linguistique fine mais seulement sur des indices de surfaces. Nous les examinerons dans la sous-section suivante.

1.1.2. Systèmes à base d'indice de surface[165]

1.1.2.1. Approche de Grosz et al. (1995)

Dans ce contexte, on peut poser la question de la structure prédéfinie, et donc de l'architecture qui permettrait à ces divers niveaux de connaissances de collaborer de façon cohérente. C'est ce que nous éclaircirons à travers la théorie du centrage d'attention (Grosz *et al*. 1995 et Walker *et al.* 1998), qui est une théorie cognitive « qui relie le focus d'attention, le choix d'une expression référentielle et la cohérence des énoncés à l'intérieur d'un segment de discours » (Grosz *et al*. 1995 : 204, reprise dans Walker 2000 : 31-32). Cette théorie a permis de clarifier un ensemble de contraintes que l'énonciateur est obligé de prendre en considération lorsqu'il emploie une expression référentielle. En effet, selon cette théorie, le sens de tout énoncé repose sur un élément central, *le centre préféré,* « défini comme étant le référent de discours psychologiquement le plus saillant pour le locuteur et l'interlocuteur au moment où l'expression qui le représente se trouve énoncée » (Kleiber 2002 : 8).

Les auteurs définissent le *Discourse Purpose* comme les informations communiquées par chaque discours. Ensuite, le discours, est découpé en segments *Discourse Segment Purpose* (*DSP*). Le partage des informations est réalisé grâce à la contribution de chaque DSP. La cohérence globale du discours dépend de la cohérence de chaque DSP, tributaire des liens entretenus entre les propositions d'un même segment, et celles de leurs relations. Cette cohérence est spécifiée dans la théorie du centrage. La cohérence est affaiblie quand, pour comprendre l'enchaînement de deux énoncés, un interlocuteur doit recourir à des

[164] Ils nécessitent une analyse des rôles sémantiques des groupes nominaux dans le texte et impliquent d'office une analyse syntaxique et morphologique.
[165] Ou *knowledge-poor approaches* (Mitkov, 1998).

inférences. Les énoncés sont incohérents si une inférence ne peut pas aboutir. Les auteurs donnent en exemple deux contextes :

[2] 1A. John went to his favorite music store to buy a piano.
1B. He had frequented the store for many years.
1C. He was excited that he could finally buy a piano.
2A. John went to his favorite music store to buy a piano.
2B. It was a store John had frequented for many years.
2C. He was excited that he could finally buy a piano.

Deux centres sont possibles, *John* ou *le magasin*, en revenant à la première phrase (1A ou 2A). Nous pouvons remarquer que, tout le long du segment, le premier contexte met l'accent sur le centre préféré *John*, alors que le second oriente le lecteur sur *le magasin* puis change brutalement de centre pour *John*. En effet, il contraint le lecteur à une inférence supplémentaire indispensable à la résolution du pronom anaphorique *he* et à la compréhension du sens du segment. Ainsi, le premier contexte semblerait plus cohérent que le second.

Grosz *et al*. (1995), à l'appui de la théorie du centrage, ont établi un système de marqueurs référentiels composé de règles garantissant la cohérence d'un segment de discours en résolvant ses anaphores. Cette théorie est jugée « mal centrée[166] » selon Kleiber (2002 : 19) qui affirme que :

> pour ce qui est de l'explication de l'emploi des expressions référentielles, la chose est beaucoup moins sûre, même si, nous le reconnaissons bien volontiers, la théorie du centrage arrive à régler certains délicats problèmes de distribution.

1.1.2.2. Algorithme de Kennedy/Boguraev (1996)

Dans une version révisée et plus importante de par sa teneur que celle qui a été développée par Lappin/Leass (1994), Kennedy/Boguraev (1996) avancent une nouvelle approche de résolution des anaphores pronominales. Leur système utilise la sortie d'un tagger, celui de Voutilainen *et al.* (1992) comme texte d'entrée. Ces textes sont enrichis seulement avec des annotations de fonction grammaticale des éléments lexicaux et un étiquetage morpho-syntaxique, donc ne nécessitent pas d'analyse syntaxique exhaustive. Même si la logique de base de leur algorithme est comparable à celle de l'algorithme de Lappin/Leass, nous pouvons toutefois noter des différences majeures. Celui de

[166] Pour plus d'informations, *cf.* Kleiber (2002).

Lappin/Leass s'appuie essentiellement sur l'information basée sur la configuration syntaxique, alors que, à défaut d'une telle information, celui de Kennedy et de Boguraev se réfère plutôt à des inférences de fonction et de priorité grammaticales. L'ensemble des référents de discours restant formeront l'ensemble d'antécédents candidats pour le pronom une fois les filtres morphologiques et syntaxiques appliqués. Un procédé final évalue alors le positionnement d'un candidat dont la saillance est supérieure et celui-ci sera retenu comme bon antécédent. Ce procédé augmente la saillance des candidats qui répondent à une localité ou un état de parallélisme s'appliquant aux candidats intraphrastiques et réduit la saillance des candidats que le pronom précède : deux facteurs[167] de saillance, ajoutés par les auteurs, ajoutent, à ceux définis dans Lappin/Leass (1994), une valeur de 65 à un syntagme nominal possessif (codé *POSS-S*) et une valeur de 50 à un syntagme figurant dans le contexte courant (*CNTX-S*) (*Cf.* Tableau 12)

SENT-S:	100	iff[16] in the current sentence
CNTX-S:	50	iff in the current context
SUBJ-S:	80	iff GFUN = *subject*
EXST-S:	70	iff in an existential construction
POSS-S:	65	iff GFUN = *possessive*
ACC-S:	50	iff GFUN = *direct object*
DAT-S:	40	iff GFUN = *indirect object*
OBLQ-S:	30	iff the complement of a preposition
HEAD-S:	80	iff EMBED = NIL
ARG-S:	50	iff ADJUNCT = NIL

Tableau 13 : Saillances et poids associés[168] selon Kennedy/Boguraev (1996)

L'algorithme, évalué sur 27 textes de divers genres (page web, publicité, reportage), a eu un taux de succès de 75 %. Ce taux est sensiblement plus bas que celui obtenu par l'algorithme de Lappin/Leass (10% de moins). Le fait que l'information en entrée du système est obtenue de manière entièrement automatique et est moins riche pourrait expliquer cette différence. De plus, selon les auteurs, leur corpus d'évaluation est plus complexe à analyser que celui composé de manuels informatiques[169] utilisé dans Lappin/Leass (1994). Malgré cette légère défaillance, Mitkov (1999 : 16) affirme que :

[167] Peu de détails sont donnés pour ces deux facteurs.
[168] Iff équivaut à Si et seulement si.
[169] Nous rappelons que nous avons pensé à ce genre textuel comme corpus pour notre travail actuel, et nous avons y renoncé à cause de l'inexistence des cas relevant une ambiguïté anaphorique.

> Les rapports d'évaluation ont une précision de 75%, mais il faut leur donner un «bonus» pour que ces résultats couvrent une couverture très large : l'évaluation a été basée sur une sélection aléatoire de genres[170].

1.1.2.3. Système CogNIAC de Baldwin (1997)

Le système CogNIAC, conçu par (Baldwin 1997)[171], est un programme de résolution de l'anaphore dans lequel ont été implantées uniquement des règles décrivant l'interprétation des pronoms. Baldwin (1997) a choisi de ne résoudre que les anaphores non ambiguës, en garantissant une haute précision[172] par rapport aux autres systèmes de résolution d'anaphore pronominale connus : il vise à fournir un antécédent unique pour chaque pronom et se distingue ainsi des autres systèmes présentés ici.

Ce système requiert une analyse préliminaire du texte : étiquetage morphosyntaxique, puis détection des propositions et reconnaissance des syntagmes nominaux. Pour chaque pronom trouvé, un ensemble d'antécédents possibles est accordé. Ces antécédents candidats sont compatibles avec le genre, le nombre et les restrictions de coréférence associées au pronom et associées au texte qui précède. Pour la résolution d'une anaphore pronominale, les règles utilisées sont les suivantes :

1. *Unique in Discourse* : Si la liste des antécédents potentiels n'en comprend qu'un seul, celui-ci est attribué au pronom et la résolution est considérée comme réussie.
2. *Reflexive* : Si un pronom est réfléchi, l'antécédent intraphrastique le plus proche sera sélectionné ;
3. *Unique in Current + Prior* : S'il existe un seul antécédent possible dans la phrase précédente et dans la phrase courante, sélectionner cet antécédent ;
4. *Possessive Pro* : Si un pronom est possessif et s'il existe dans la phrase précédente un SN déterminé par un possessif, ce SN sera sélectionné comme antécédent ***;***
5. *Unique Current sentence* : S'il y a un seul antécédent possible dans la phrase courante, sélectionner cet antécédent ;
6. *Unique Subject/Subject Pronoun* : Si un pronom a la fonction syntaxique de sujet de la phrase courante et la phrase précédente ne contient qu'un sujet comme antécédent possible, sélectionner cet antécédent ;
7. Dans tous les autres cas, le pronom est laissé sans antécédent et l'anaphore est non résolue.

---

[170] Texte source : « Evaluation reports 75% accuracy but this has to be given a « bonus » for this results span a very wide coverage : the evaluation was based on a random selection of genres ».
[171] Le système a été étendu par la suite, mais nous nous limitons ici au traitement des pronoms.
[172] Nous avons défini cette mesure de qualité dans le chapitre 3 de la deuxième partie de notre travail.

Dans un contexte assez restreint, Baldwin (1997) donne une évaluation de son système : sur des textes narratifs mettant en jeu deux personnages de même sexe, le système traite les pronoms singuliers dénotant uniquement des êtres humains (*he, she, him...)* et cela pour minimiser l'ambiguïté de résolution des pronoms. La précision (mesure de qualité) est de 92 % et le rappel[173] (mesure de quantité) est de 64 **%**. Baldwin a mentionné que grâce à la troisième règle, 35% des pronoms du corpus sont résolus. On remarquera que la règle 5 est un cas particulier de la règle 3. Nous pouvons supposer que la valeur des règles 3 et 5 vient du fait que les reprises interphrastiques sont en général nettement moins fréquentes que les reprises intraphrastiques. En ajoutant les deux règles ci-dessous à CogNIAC, Baldwin (1997) a amélioré son système et a augmenté le rappel en obtenant un taux de succès de 77,9% sur un corpus comptant 298 pronoms *it* :

7a. *Cb. Picking* : Cette règle repose sur la théorie du centrage : S'il existe plusieurs antécédents possible, sélectionner celui qui est coréférent avec une expression de la proposition ou phrase précédente ;

7b. *Pick Most Recent* : Sélectionner l'antécédent le plus récent dans le texte.

En insistant sur le fait qu'il ne s'engage pas à fournir une réponse pour tout pronom, Baldwin distingue son système de l'algorithme de Hobbs ou du système de Lappin/Leass. Cette caractéristique n'est cependant pas si spécifique. En effet, un système tel que celui de Lappin/Leass serait aisément adaptable pour ne rattacher effectivement un pronom à un antécédent que si celui-ci a une valeur de saillance nettement supérieure à celle des autres. En revanche, ce qui est plus particulier au système de Baldwin est le fait qu'il sélectionne l'antécédent d'un pronom sur un critère absolu. Chaque règle est susceptible soit de sélectionner, parmi l'ensemble des antécédents possibles, un et un seul antécédent, soit de n'avoir aucun effet sur eux. Une telle approche a l'avantage de permettre une évaluation moins diffuse des règles utilisées, le degré de validité de chacune pouvant être mesuré (par exemple la règle 1 a été évaluée comme produisant toujours une réponse correcte dans les deux évaluations, la règle 3 comme ayant une précision de 96 % sur un corpus et de 72 % sur un autre corpus).

[173] *Cf.* chapitre 3 de la deuxième partie.

1.1.2.4. Mitkov (1998)

L'approche localiste était à la base des premiers travaux dans la résolution référentielle. Le principe sur lequel se base cette approche est que l'antécédent du pronom figure explicitement dans l'environnement discursif où il apparaît. Cette approche, qui se caractérise principalement par le peu de connaissances linguistiques qu'elle utilise, peut être associée à l'approche pauvre en connaissances (*Knowledge poor approach*) de Mitkov (1998, 2002). Mitkov (1998), visant à réduire l'utilisation des données syntaxiques et sémantiques, a proposé un algorithme sans analyse syntaxique, sémantique et discursive complexe. L'objectif de son système MARS, applicable à plusieurs langues (anglais, polonais, arabe) était d'assurer un bon taux de réussite de résolution anaphorique dans des manuels techniques informatiques.

MARS nécessite juste en entrée la sortie d'un étiqueteur morpho-syntaxique et pourra se passer d'analyse syntaxique et sémantique : le texte est d'abord soumis à une analyse syntaxique de surface *part-of-speech tagger* et à un extracteur de groupes nominaux (GN). Ensuite, on localise les groupes nominaux identifiés à une distance de deux phrases de l'anaphore à résoudre, pour retenir les GN qui s'accordent avec l'anaphore. On effectue, après, une vérification de l'accord en genre et en nombre. Des heuristiques sont appliquées afin de déterminer le score, qui varie pour chacune d'elles entre (-1, 0, 1, ou 2), des groupes nominaux antécédents potentiels. On choisit comme antécédent le candidat qui a le score le plus élevé. Dans le processus, les heuristiques sont mises en œuvre sur une liste de préférences appelées des indicateurs d'antécédents : la répétition d'expressions, la saillance, la distance référentielle, et la topologie lexicale du texte. La liste des heuristiques de MARS, qui en compte 10, est recensée dans ce qui suit :

- *Definiteness* : on peut obtenir de manière efficiente un antécédent grâce à ce facteur. Les groupes nominaux indéfinis, pénalisés avec score de -1, sont des antécédents moins probables que les définis qui ont un score de 0,

- *Givenness :* si un candidat représente le thème de la phrase précédente (*the given information*), il obtient le score 1 sinon 0,

- *Indicating verbs :* si un candidat suit un verbe « indicateur[174] », il est considéré comme le candidat préféré avec un score de 1 sinon son score est 0,

[174] Liste des verbs "indicators" d'après Mitkov (1998 : 870) : discuss, present, illustrate, identify, summarise, examine, describe, define, show, check, develop, review, report, outline, consider, investigate, explore,

- *Lexical reiteration :* les groupes nominaux qui sont répétés deux fois ou plus dans le même paragraphe ont 2 points, 1 point s'ils sont répétés une fois, 0 points sinon. Cela inclut les synonymes et les têtes lexicales[175] ,

- *Section heading preference* : si un candidat apparaît dans le titre d'une section du document, alors il est favorisé d'1 point, sinon son score est 0,

- *« Non-prepositional » noun phrases :* un candidat qui fait partie d'un groupe prépositionnel (-1 points) est lésé par rapport à un candidat qui ne fait pas partie d'un GP (0 point),

- *Collocation pattern preference :* un candidat qui apparait dans une construction identique au pronom est favorisé par un score 2 sinon 0. La préférence est limitée aux contextes de parallélisme de la forme (verbe, GN) et (verbe, pronom).

- *Immediate reference :* la construction fréquente dans les manuels techniques est la suivante : verbe1 + GN + conjonction (*and, or, before,* …) + verbe2 + anaphore[176] . Si un candidat apparait dans telle construction, son score est 2 sinon 0.

- *Referential distance :* plus un GN est proche d'une anaphore et plus il a de chances d'être considéré comme antécédent :

  - Dans une phrase simple : si le candidat est dans la phrase précédente, son score est 1, s'il est dans 2 phrases avant, son score est 0, s'il est situé plus loin, son score est -1.

  - Dans une phrase complexe : le meilleur candidat est le candidat situé dans la proposition précédente et son score est 2, suivi des candidats dans la phrase précédente (1), ensuite deux phrases avant (0) et finalement 3 phrases avant (-1),

- *Term preference :* le candidat qui relève du domaine couvert par le texte a le score 1 sinon 0.

---

assess, analyse, synthesise, study, survey, deal, cover. Une caractéristique qui leur confère le statut de bons indicateurs est que les noms qui suivent ces verbes sont particulièrement saillants.

[175] Exemple (Mitkov, 1998): The toner bottle, the bottle of toner, the bottle.

[176] Exemple: *Press the key down…press it again.*

Indicateur	Valeur	Score
Information	Premier GN de la phrase	+1
Verbes	GN suivant certains verbes (*analyse, consider, discuss, ...*)	+ 1
Répétition lexicale	GN répété deux fois dans le même paragraphe que le pronom	+2
	GN répété une fois dans le même paragraphe que le pronom	+1
Titre de section	GN dans le titre de la section	+1
Collocation	GN combiné au même verbe	+2
Distance	GN contenu dans la phrase précédente immédiate	+2
	Décrémentation de 1 (-1) par phrase	
Parallélisme	GN contenu dans la phrase précédente avec des connecteurs *(and, or, ...)*	+2
Terminologie	terme du même domaine que celui du texte	+ 1
Définitude	GN indéfini	-1
Expression prépositionnelle	GN tête d'une expression prépositionnelle	-1

Tableau 14 : Résumé des indicateurs de saillance de Mitkov (1998)

MARS[177] fonctionne selon la procédure suivante :

- Rechercher les GN à gauche de l'anaphore et analyser les trois phrases précédant l'anaphore (si disponibles).
- Les candidats dont les traits de genre et nombre sont incohérents avec l'anaphore sont supprimés de la liste.
- Attribuer des points aux heuristiques préférentielles et les appliquer.
- L'antécédent est le candidat avec le score total le plus élevé. Sera privilégié le candidat avec le meilleur score dans « *immediate reference* » si deux candidats ont le même score. Si malgré tout la sélection n'est pas faisable, alors c'est « *collocation pattern*» qui décide ; sinon « *indicating verbs* », sinon choisir le candidat le plus récent.

[177] qui ne traite pas les cataphores.

Mitkov (1998) a évalué son système sur des textes techniques en anglais contenant un total de 294 pronoms anaphoriques ; il a pu résoudre correctement 264 anaphores, soit un taux de réussite assez favorable de 89,7%. Les tests d'autres langues comme le polonais et l'arabe ont donné des résultats meilleurs (93,3% pour le polonais et 95,8% pour l'arabe). Néanmoins, nous pensons que certains choix n'ont pas été suffisamment justifiés et que quelques questions s'imposent. Par exemple, sur quelle base le choix des scores s'est-il opéré ? Ou encore, pourquoi choisir de s'arrêter aux deux phrases précédentes afin d'identifier l'antécédent d'une anaphore ?

Au vue des performances élevées des approches de Mitkov et de Lappin/Leass et du fait qu'elles n'exigent pas d'analyses syntaxique et sémantique complexes nécessitant l'utilisation d'outils qui ne sont pas toujours disponibles, il nous a semblé intéressant de nous en inspirer. L'approche que nous proposons ci-dessous se veut encore moins exigeante en ressources ; néanmoins, nous voulons justifier au moins le choix de la taille de l'espace de recherche des antécédents des anaphores.

1.2. Systèmes Français

Tous les systèmes que nous venons de présenter ci-dessus sont des systèmes anglais[178]. Puisque nous traitons le français dans ce travail, il sera judicieux de présenter quelques approches de résolution de l'anaphore dans des textes français[179]. Cette section sera divisée en deux partie : nous présenterons, en premier lieu, les systèmes à base cognitive en nous appuyant sur l'approche de Popescu-Belis *et al.* (1998), ensuite, nous citerons des systèmes spécialisés dont un type de référence particulier est l'objet d'étude.

---

[178] Une recherche très active dans la résolution d'anaphore est menée par le GPLSI de l'Université d'Alicante. Un système, susceptible de fonctionner sur des textes écrits (Ferrández *et al.,* 1998) ou des dialogues (Martínez Barco *et al.,* 1999), a été implémenté pour l'espagnol . La procédure générale de résolution est similaire à celle utilisée par Mitkov ou Lappin & Leass et consiste en trois modules :

- un premier qui détermine l'espace dans lequel figure l'antécédent et retourne une liste d'antécédents possibles,
- un deuxième qui filtre la liste en éliminant les antécédents possibles qui ne satisfont pas les conditions d'accord et restrictions syntaxiques,
- un dernier qui applique des préférences. Néanmoins, ces dernières sont peu documentées dans les articles que nous avons pu consulter.

Le système a bien été évalué mais sur un ensemble réduit de pronoms (80 pronoms dans *(Martínez Barco et al.,* 1999) et un corpus de 9600 mots dans (Ferrández *et al.,* 1998)).
Ces approches ont rapidement trouvé leurs limites et l'essor de la linguistique de corpus a encouragé l'exploitation de données attestées.

[179] Malheureusement, comme nous l'avons montré dans le deuxième chapitre de la première partie, la rareté de tels systèmes en français est notable.

1.2.1. Systèmes à base cognitive[180]

Au niveau de l'approche cognitiviste, le référent des formes linguistiques est ainsi toujours, dans cette approche, un élément cognitif appartenant à la conscience du lecteur. Le système exploité doit présenter des connaissances relatives à la cognition humaine[181] afin de permettre une facilité et une accessibilité des interprétations mises en place par la machine, en d'autres termes, leur réponse aux attentes des utilisateurs. On utilisera ce que l'on connait du fonctionnement cognitif humain face au langage analogiquement au système lui-même (même si ce n'est pas une obligation, c'est souvent une source d'inspiration extrêmement utile) grâce à l'accessibilité des connaissances. La compréhension est aussi le résultat émanant de processus cognitifs qu'il n'est pas toujours aisé de décrire d'une façon algorithmique. Elle ne repose pas uniquement sur un ensemble de critères logiques d'évaluation.

C'est un mécanisme prédictif techniquement très différent des analyses classiques, réalisé par des processus totalement automatiques (c'est-à-dire non contrôlés ni réflexifs). En effet, le contexte doit diriger le système vers une interprétation résultante, souvent unique, car l'énoncé en processus de traitement peut se prêter à plusieurs interprétations construites en parallèle. L'état du contexte cognitif s'apparente alors à un ensemble d'hypothèses qui permet l'élargissement du nombre des interprétations les plus cohérentes.

Selon Salmon-Alt (2001), afin de garantir la construction de modèles du langage plus abstraits, il serait préférable de fonder le processus d'interprétation sur des principes généraux comme : « l'attachement minimal » (ne pas retenir des nœuds de l'arbre syntaxique potentiellement inutiles) et la « clôture différée » (relier les nouveaux éléments au syntagme en cours de traitement). Par exemple, en analyse syntaxique, les interprétations les plus retenues généralement sont celles qui se conforment à ces principes. Malgré le fait qu'isoler les éléments de jugement pertinents (si l'étude statistique de corpus permet de révéler les règles générales, elle ne donne néanmoins pas le moyen de traiter les cas particuliers) est une tâche assez délicate, il faudrait mettre en place un moyen d'identifier leur exception. Cela expliquerait ainsi le fait que ces régularités ne peuvent pas être utilisées comme des règles formelles d'analyse. En revanche, il est plus facile de les expliquer comme un effet découlant de l'organisation concurrentielle des processus

[180] On peut citer aussi, parmi les travaux de calcul référentiel s'appuyant sur cette approche : le modèle proposé par Salmon-Alt (2001) et l'algorithme Dupont (2003), où l'élément cognitif présent dans la conscience du lecteur est appelé l'entité.

[181] Il est certain que le système doit avoir une bonne représentation de son interlocuteur et de son fonctionnement cognitif.

interprétatifs : les interprétations qui vérifient l'attachement minimal et la clôture différée sont généralement les plus simples à réaliser et donc les premières à être perçues.

Le système de Popescu-Belis *et al.* (1998) a été créé pour la résolution des pronoms dans des textes en français en s'appuyant sur la représentation mentale de la référence[182] qui se compose d'une série d'expressions référentielles et leur référent. Développé dans le contexte plus large d'un système de résolution de la référence, ce système vise à traiter l'ensemble des phénomènes de coréférence. Il manipule, pour ce faire, des objets conceptuels représentant les êtres dénotés par les expressions ou représentations mentales. Le texte est analysé linéairement de gauche à droite. Quand le système rencontre une expression référentielle, il détermine si celle-ci doit donner lieu à la création d'une nouvelle représentation mentale ou être rattachée à une représentation mentale existante (il y a reprise avec coréférence).

L'auteur détaille peu les règles ou facteurs déterminant le rattachement éventuel d'une expression à une représentation mentale particulière (Popescu Belis, 1999 : 210) : d'abord, le système cherche les représentations mentales qui peuvent se rattacher à l'expression référentielle. Pour ce faire, il utilise trois règles : deux morpho-syntaxiques qui vérifient l'accord en genre et en nombre entre deux expressions coréférentes et une règle d'accord sémantique entre l'expression référentielles à interpréter et les expressions référentielles d'une représentation mentale. Ce système utilise un dictionnaire spécifique comme une ressource sémantique lexicale restreinte aux termes du corpus. Après avoir éliminé les représentations mentales qui ne répondent pas aux contraintes citées ci-dessus, le système calcule le taux d'activation (équivaut à la saillance) de chaque représentation restante. Il lui attribue une valeur initiale de 15. En cas de réactivation, un taux d'activation supplémentaire lui est ajouté et ce selon la catégorie de l'expression référentielle (nom propre=40, nom commun=20, pronom=10). Tout au long du texte, la réactivation de la représentation mentale diminue de 2 points dans chaque nouvelle phrase et de 4 points dans chaque nouveau paragraphe.

Popescu-Belis et *al.* (1998) ont évalué leur système sur deux textes narratifs français : *Vittoria Accoramboni* de Stendhal et un extrait du *Père Goriot* de Balzac. Ils ont obtenu les taux de réussite suivants : 65.48% pour le premier texte et 78.4% pour le deuxième. Popescu-Belis ne fournit pas de résultats spécifiques de la tâche d'interprétation des pronoms dans ses publications les plus récentes.

[182] Cf. Reboul et Gaiffe (1999) pour d'amples informations sur la représentation mentale et la référence.

1.2.2. Systèmes spécifiques

La résolution automatique de l'anaphore en français a vu la création de systèmes particularisés dans le cadre de travaux de recherche académique. Certains auteurs ont consacré leurs études à un type bien précis de référence comme les anaphores infidèles (Salmon-Alt 2004), les anaphores événementielles (Bittar 2006), les noms propres (Boudreau et Kittredge 2006) et la coréférence (Longo 2013). D'autres se sont intéressé au traitement des corpus spécialisés comme les constats des accidents de la route (Nouioua 2007) et le discours politique (Adam 2007).

1.2.2.1. Résolution d'un type de référence spécifique

a. Résolution des anaphores infidèles (Salmon-Alt 2004)

Différentes ressources linguistiques ont été utilisées dans ce système de résolution des anaphores appelées *infidèles* (possédant une tête nominale qui n'est pas la même que celle de l'antécédent (Kleiber 1994)). L'algorithme utilisé sert à évaluer différentes configurations en cascade de divers critères de l'antécédent comme les critères d'accord, de distance phrastique et de tête de l'antécédent. Ainsi, le but de ce système est d'identifier, dans les cinq phrases qui précèdent l'anaphore, l'antécédent répondant à ces critères.

En utilisant essentiellement différentes ressources linguistiques, telle que l'analyse syntaxique fine, ce système obtient des performances de 31.9%, néanmoins, sans ces ressources ce chiffre tombe à 15%. Son autrice admet qu'il a besoin d'amélioration mais souligne par ailleurs qu'il est tout aussi performant que le système proposé par Poesio en 2002.

b. Résolution des anaphores événementielles[183] (Bittar 2006)

Le système proposé par Bittar débute par l'annotation du texte. Afin d'éliminer les emplois non anaphoriques des pronoms, il les identifie d'abord en étiquetant morpho-syntaxiquement un texte et le découpant en propositions, grâce à l'établissement de la liste de verbes dont le sujet est un pronom non anaphorique. Ensuite, afin de repérer les événements sous forme phrastique, ce sont les verbes événementiels[184], qui sont identifiés puis regroupés en groupes événementiels qui peuvent être repris par le même pronom. Afin de sélectionner le type d'antécédent d'un pronom, les anaphores pronominales sont

[183] Les anaphores appelées *évènementielles* sont une reprise par un pronom d'un évènement antérieur (ici le système ne prend en compte que les antécédents phrastiques).
[184] Désignant achèvement, activité ou accomplissement (Vendler 1957).

détectées suivant des conteneurs regroupant les verbes nécessitant comme argument événementiels un sujet ou un objet événementiel ; on les appelle des conteneurs événementiels (Vendler 1957). Par la suite, un algorithme décisionnel, inspiré de Lappin/Leass (1994) et Mitkov (2002), sélectionne, suivant différents critères comme la préférence pour un événement déjà anaphorisé, l'antécédent potentiel pour chaque anaphore et choisit celui qui obtient le meilleure score (c'est celui qui est le plus proche de l'anaphore qui est choisi en cas d'égalité).

Nous remarquons que l'algorithme de Bittar (2006) accorde une grande importance au trait de distance au détriment des autres. L'évaluation manuelle de ce modèle sur trois textes de *Frantext*[185] a montré la défaillance de l'importance accordée au critère de distance : en effet, des erreurs sont induites lors de la sélection des antécédents en déséquilibrant le score final.

c. Résolution des noms propres (Boudreau et Kittredge 2006)

Afin d'identifier les chaînes de référence initiées par un nom propre, Boudreau et Kittredge (2006) ont développé un algorithme à base de connaissances linguistiques limitées s'intéressant aux noms propres et à leurs coréférents et en traitant différents textes courts (manuels d'installation informatique, critiques de films et fusions de compagnies). Tout d'abord, le système identifie dans des textes bruts toutes les expressions référentielles (les groupes nominaux simples associés au domaine, les pronoms, etc.). Ensuite, des groupes nominaux complexes, des groupes prépositionnels et des noms propres complets sont créés et le système leurs attribue, suivant la position des expressions référentielles, des fonctions syntaxiques. Enfin, suivant différents critères pondérés (répétition de la tête lexicale, fonction syntaxique, distance, etc.), les paires antécédents-anaphores sont sélectionnées.

Le système de Boudreau et Kittredge n'a pas été évalué, pourtant, les autrices affirment qu'il peut caractériser, dans des textes qui sont courts, des chaînes de référence qui ont été initiées par un nom propre ; ceci en utilisant des ressources linguistiques qui sont limitées. Par ailleurs, elles admettent que leur système requiert des améliorations, notamment concernant la délimitation des frontières des groupes complexes ou encore l'attribution des fonctions syntaxiques qui n'est pas assez précise. Ceci peut engendrer des

[185] http://www.frantext.fr/

erreurs qui ne permettent pas de l'utiliser à grande échelle, malgré l'intérêt que l'on peut porter à ce système.

d. Résolution de la coréférence (Longo 2013)

Dans sa thèse, Longo (2013) s'intéresse à l'étude et la modélisation des chaînes de référence portant sur des référents humains et non humains dans des textes narratifs. Elle a élaboré le module *RefGen* qui est le module linguistique central du module de détection automatique des thèmes, *ATDS-FR*[186]. *RefGen* est divisé en deux sous-modules : *RefAnnot* pour l'annotation des différentes expressions référentielles et *CalcRef* pour le calcul de la référence. Le texte d'entrée est segmenté et annoté automatiquement. Le premier module *RefAnnot* procède à l'identification des expressions référentielles (groupes nominaux complexes et entités nommées) qui peuvent être des candidats potentiels au poste de premier maillon référentiel. Ensuite, le deuxième module *CalcRef* calcule le score de chaque candidat en se basant sur des préférences lexicales, morphosyntaxiques et sémantiques. L'algorithme de ce module est inspiré de l'approche de Mitkov (2002). Chaque antécédent se voit attribuer un score selon son accessibilité globale, son rôle syntaxique, et sa position dans le texte. Les paires antécédents-anaphores validées sont enfin regroupées en chaînes de coréférence lorsqu'elles ont le même référent.

L'évaluation de RefGen a donné un score de performance moyen (entre 63% et 73%). Selon Longo (2013), les cas d'erreurs sont dus aux annotations de l'analyseur syntaxique de départ qui ont servi comme texte d'entrée. Elle propose, comme perspectives, quelques suggestions pour améliorer son système comme : l'utilisation de la dimension temporelle, la constitution d'un corpus annoté au préalable (ce qui manque au français) et de faire appel à des marqueurs linguistiques pour délimiter les frontières de chaque candidat.

[186] ATDS-FR : Automatic Topic Detection System for French.

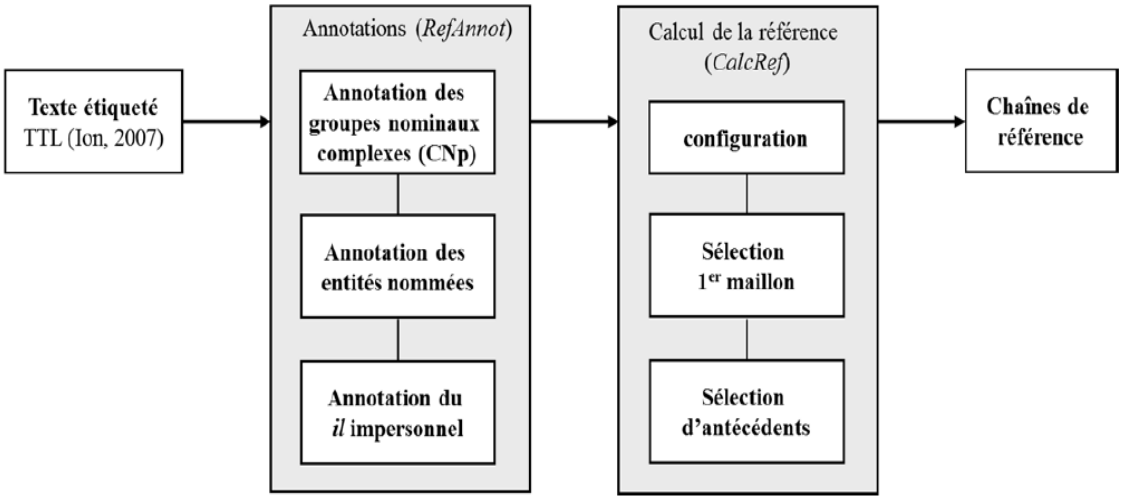


Figure 28 : Architecture du module RefGen de Longo (2013 : 255)

1.2.2.2. Résolution dans un corpus spécifique

a. Résolution des anaphores dans les textes d'accidents de la route (Nouioua 2007)

Dans un corpus regroupant des textes d'accidents routiers, ce système est une heuristique pour la résolution des anaphores où différentes relations sont caractérisées à partir d'une analyse syntaxique de surface, puis traitées pour établir la cause de l'accident. Les personnes et les véhicules en relation avec l'accident sont pris en compte (référents potentiels) dans ce système et sont associés à des informations renvoyant à leur nature. Dans le but de propager la référence et la nature du référent, l'algorithme fonctionne par étape, les anaphores simples sont résolues et remplacées par une constante, alors que pour celles non résolues, il établit une série d'antécédents potentiels et choisit celui qui est le plus rapproché de l'anaphore. La procédure de ce système est itérative, c'est-à-dire que le processus ne s'arrête que lorsqu'il n'y a plus d'anaphore à résoudre.

L'évaluation de ce système montre des performances de 95% (figure 3) ce qui montre son efficacité ; néanmoins, cette approche est dépendante du domaine spécifique des accidents de la route. Il faudrait créer un algorithme plus généralisé, mais dans ce cas la mise au point de nombreuses heuristiques seraient nécessaires.

	Textes d'entraînement	Textes de validation
Nombre d'anaphores	428	592
Nombre de références correctes	424	564
Pourcentage des références correctes	**99 %**	**95 %**

Figure 29 : Résultats d'évaluation du système de Nouioua (2006)

b. Résolution de la coréférence dans un discours politique (Adam 2007)

Adam[187] (2007) a traité automatiquement la coréférence dans un corpus politique d'une application de veille terminologique LexiMedia2007[188], qui a permis de suivre l'actualité des élections présidentielles de 2007 semaine après semaine. Le système d'Adam permet de créer des relations entre des expressions ayant une même référence politique au fil d'un corpus couvrant une période allant du 21 août 2006 au 20 avril 2007, inclus. Son corpus est constitué des textes de trois journaux : Le Monde, Le Figaro et Libération. Adam (2007) a répartit son corpus en 3 sous-corpus : un d'entraînement, un d'évaluation (annoté manuellement) et un d'évaluation non annoté qui a servi à « observer la progression des résultats avec l'augmentation du corpus » (Adam 2007 : 56).

Les textes du corpus d'entraînement subissent un prétraitement au début : une annotation syntaxique au préalable par un analyseur syntaxique Syntex[189] ; extraction semi-automatique des mots clés (selon leurs fréquences d'apparition) et leur filtration manuelle. Par la suite, un module de balisage permet d'annoter les groupes nominaux dont la tête lexicale appartient à la liste des mots-clés ; c'est selon leur type que les noms propres sont annotés et on leur attribue un genre. Ensuite, les expressions potentiellement référentielles à des personnes sont repérées. Ces derniers sont, alors, associés à un identifiant et classés selon leur score d'association. Enfin, en usant d'indices de contexte, un module de projection permet la résolution de groupes nominaux balisés. Ainsi, on classe les antécédents potentiels en fonction de leur score de saillance total : celui qui a le meilleur score est sélectionné pour être l'antécédent.

[187] Clémentine Adam, dans le cadre d'un mémoire de Master en 2007.
[188] http://redac.univ-tlse2.fr/LexiMedia2007/infos/apropos.jsp
[189] Un analyseur syntaxique de corpus développé par Bourigault et Fabre (2000).

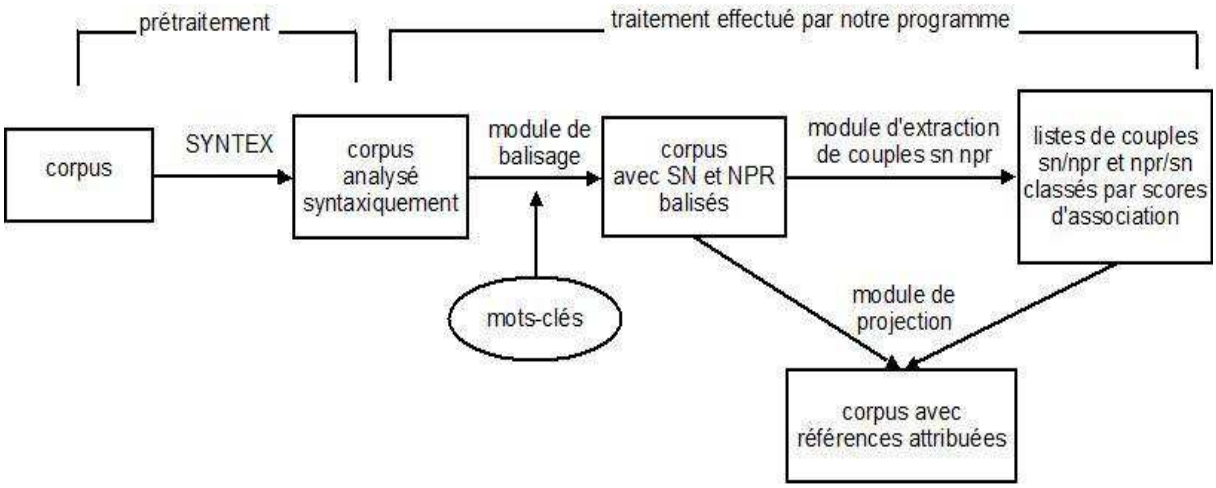


Figure 30 : Traitement de la coréférence selon Adam (2007 : 61)

Lors de son évaluation, ce système a montré des performances moyennes (67,14%) malgré la spécificité et l'homogénéité du corpus. Deux manques de précision ont été relevés par l'autrice, qui pourraient expliquer ce taux de réussite : un manque d'utilisation précise de l'analyseur syntaxique SYNTEX ainsi que pour le traitement des abréviations.

Nous déduisons, à partir de cette vue d'ensemble des systèmes de résolution de la référence en français que $RESUMAN_O$ est 'doublement' spécialisé : d'une part, il traite un type de référence spécifique qui est l'anaphore pronominale et dans un corpus spécifique : les résumés littéraires[190].

Face aux limites des systèmes à base de connaissances linguistiques cités ci-dessus, des systèmes par apprentissage statistique ont été créés. Sans viser à l'exhaustivité, nous évoquerons brièvement quelques systèmes dans la section suivante.

2. Systèmes par apprentissage statistique[191]

Les systèmes par apprentissage se sont développés au cours des deux dernières décennies. Ils sont entrainés sur des corpus annotés en relations coréférentielles mis à disposition par des compagnes d'évaluation[192] comme MUC, SemEval, CoNLL, etc. Ces systèmes procèdent en premier lieu à une classification probabiliste des paires anaphore-antécédent afin de vérifier si elles sont coréférentes ou non. Ensuite, ils regroupent les

[190] Qui pourra se généraliser sur des textes littéraires longs. Nous testerons cette particularité dans le chapitre suivant.
[191] Nous tacherons d'en citer trois seulement : d'une part car nous avons remarqué que ces systèmes traitaient, en majorité, la coréférence et d'autre part, notre outil est à base de connaissances linguistiques. Pour une revue exhaustive, nous orientons le lecteur vers : Trouilleux (2004), Rahman et Ng (2009), Ng (2010), Poesio *et al.* (2010), Zheng *et al.* (2011), Longo (2013).
[192] Voir chapitre 2 de la première partie.

paires coréférentes pour former une chaîne de coréférence (Recasens 2010). La plupart de ces systèmes sont à base de corpus en anglais[193].

2.1. Connolly *et al.* (1994)

Connolly *et al.* (1994) décomposent la résolution de la coréférence en deux étapes : pour une expression anaphorique donnée, d'abord, ils classent successivement des paires de candidats potentiels pour retenir après le meilleur des deux candidats. Connolly *et al.* utilisent les informations suivantes pour classer les deux antécédents de l'anaphore :

- accord morphosyntaxique,
- fonction grammaticale,
- et récence phrastique et distance entre les trois éléments identifiés.

Après l'attribution de ces traits, un classifieur parcourt une paire d'antécédents potentiels en éliminant le candidat perdant et en retenant le meilleur candidat. Une nouvelle paire se forme alors et sera testée à son tour de la même manière. A la fin de ce processus, le meilleur candidat retenu est choisi comme antécédent. Les auteurs, dans Connolly *et al.* (1997), ont évalué leur système sur un corpus d'articles de presse où les liens de coréférence ont été annotés manuellement et où, par des moyens automatiques, les informations indiquées ci-dessus ont été associées aux expressions concernées (l'anaphore et les antécédents potentiels regroupés en paire). Ils ont obtenu un taux de succès de 55,3 %, un score assez nettement inférieur à celui obtenu par les algorithmes décrits plus haut. Aucune explication à cette différence n'est fournie par les auteurs ; Nous pensons que la pauvreté en connaissances linguistiques pourrait bien justifier ce score bas.

2.2. Ge *et al.* (1998)

Ge *et al.* (1998) ont implanté un système par apprentissage à partir d'une partie du corpus annoté *Penn TreeBank*[194] (environ 94 000 mots). Ce système calcule la probabilité qu'un groupe nominal soit l'antécédent d'un pronom *it* anaphorique, *he* et *she* en attribuant des poids à des facteurs similaires à ceux qui sont utilisés en général dans la majorité des algorithmes : distance, accord, tête lexicale de l'antécédent et nombre d'cooccurrence des mentions (*mention-count*). Le système utilise ensuite ces statistiques pour résoudre les anaphores dans un corpus test.

---

[193] Nous avons démontré dans le chapitre 2 de la première partie la rareté des corpus français annotés en chaînes référentielles ce qui engendre la rareté des systèmes par apprentissage en français.
[194] http://www.cis.upenn.edu/~treebank/

Les auteurs ont procédé à plusieurs tests pour l'amélioration de leur système. Ils ont vérifié l'influence de chaque trait de sélection à retenir en implémentant, au départ, un système auquel ils ajoutaient, à chaque fois, un de ces facteurs de classement. Un taux de succès de 43 % serait produit par un système consistant à retenir comme antécédent l'expression la plus proche. L'importance de la syntaxe est indiquée par l'évaluation de l'apport des différentes sources d'information : un taux de succès de 65,3 % est produit par l'utilisation d'informations sur la structure syntaxique. Ce taux est porté à 75,7 % par l'ajout des contraintes d'accord morphologique (genre/nombre). Les auteurs soulignent ainsi l'importance de cette information. Une faible amélioration de 2,2 **%** (77,9 %) est obtenue par l'utilisation du facteur *mention-count*. Enfin, une amélioration de 4,6 % est obtenue par la prise en compte du nombre d'évocations d'un référent. Le système de Ge *et al.* (1998) a interprété correctement 82,5 **%** des pronoms quand il est appliqué à un échantillon du même corpus n'ayant pas servi à l'entraînement.

Nous constatons que Ge *et al.* (1998) ont choisi les poids utilisés dans leur système selon des mesures statistiques, ce qui rend leur utilisation justifiable dès le départ. Cette méthode d'élaboration s'oppose à celle utilisée par Lappin/Leass (1994) ou Mitkov (1998) qui, après avoir déterminé les poids utilisés d'une manière intuitive, ont amélioré leurs systèmes en effectuant des tests successifs. Nous retenons aussi que l'adaptation du système à un corpus donné est aisément envisagée par l'approche par apprentissage statistique.

2.3. BART (2010)

Le système BART[195] a été conçu la première fois par Versley et *al.* (2008) dans le cadre du projet *Exploiting Lexical and Encyclopedic Resources For Entity Disambiguation at the Johns Hopkins Summer Workshop*. Charniak/Elsner (2009) ont essayé d'inclure BART dans l'évaluation de leur propre système par rapport à d'autres, mais ont été incapables de le faire car ils n'ont pas pu faire fonctionner le programme. Broscheit *et al.* (2010) ont réussi à implémenter BART. L'algorithme de BART effectue la résolution automatique de coréférence par une approche statistique par couples de mentions comme celui de Connolly *et al.* (1994). Il peut être utilisé pour plusieurs langues, même s'il a été conçu principalement pour l'anglais. Cependant, en raison de sa spécialisation, la boîte à

[195] *Beautiful Anaphora Resolution Toolkit* : accessible via http://www.bart-coref.org/

outils donne de meilleurs résultats sur les données en anglais que sur les données en d'autres langues (Berndtsson 2014).

BART (Broscheit *et al.* 2010) est constitué de deux modules : le prétraitement, qui consiste à analyser les textes d'entrées, constitue la première étape de l'architecture du système. Son but étant de former des chaînes de coréférence, il est fondé sur des contraintes éliminatoires et des préférences de sélection. Sur la base des résultats de prétraitement, un ensemble de paires composées d'une anaphore et d'un candidat pour un antécédent est ensuite créé. Sur cette base, il utilise l'un des facteurs disponibles afin de déterminer si la paire fait partie de la même chaîne coréférente ou non. Le système est entraîné avec des paires de mentions coréférentes et non coréférentes, pour apprendre une représentation de ces deux types de paires. L'approche utilisée repose sur l'apprentissage par calcul statistique des poids de chaque paire pour identifier les chaînes de coréférences dans un corpus donné.

3. Discussion

La résolution des anaphores ne fait pas encore l'objet d'une réalisation automatique parfaite et demeure une tâche complexe. Soulignons que la réussite des quelques algorithmes précédents est déterminée par plusieurs paramètres que nous avons cités plus haut. Dans cette partie de l'étude, nous analyserons, dans la première sous-section, les lacunes de ses systèmes afin de les éviter lors de l'élaboration de notre algorithme et nous tenterons, dans la deuxième, d'identifier les modules qui doivent être pris en compte pour évaluer les performances de RESUMAN$_A$.

3.1. Limite des systèmes de résolution d'anaphores

En dépit des améliorations apportées à ces systèmes, les résultats sont encore loin d'être prometteurs et cela est dû à plusieurs défaillances dans certaines résolutions. Les systèmes symboliques (reposant sur des connaissances linguistiques) souffrent de l'inexistence d'un grand corpus annoté en chaînes référentielles ce qui limite la nature de l'anaphore traitée : la majorité des algorithmes cherche à résoudre l'anaphore pronominale et ne prennent pas en compte la coréférence.

L'approche de Hobbs (1976) a ses limites : afin établir des liens sémantiques implicites entre les entités du discours, l'auteur suppose qu'un prétraitement révèle les attributs pertinents, qui constituent les entrées d'un lexique qui encode les connaissances sémantiques du monde, des phrases du texte et les traduit en énoncés de premier ordre. Les

faits implicites du texte sont révélés et les entités du discours sont unifiées grâce à une série d'inférences. Néanmoins, il existe encore des insuffisances dans la précision et la teneur des connaissances sémantiques dont dispose son système. Ces insuffisances empêchent l'application efficiente des stratégies automatiques. Par ailleurs, d'autres difficultés ne sont que partiellement résolues. Citons, par exemple, l'élaboration d'une analyse syntaxique complète d'une phrase, le choix des attributs pertinents et la démarche nécessaire à la résolution ou encore le calcul relatif à la cohérence d'un contexte local permettant de déterminer l'élément occurrent de ce contexte. Cette approche était non performante à cause soit de l'impossibilité de traiter d'une manière automatique des connaissances sémantiques soit parce qu'elles étaient trop peu fiables pour être utilisables. Le recours à une annotation manuelle était donc omniprésent lors de la résolution, ce qui nécessitait un important travail d'analyse textuelle. Les systèmes symboliques, à l'instar de celui de Hobbs, ne pouvaient pas alors traiter automatiquement des corpus volumineux. L'utilisation automatique de ces systèmes sur les longs corpus est perturbée car les connaissances linguistiques sont souvent traitées manuellement.

Une autre lacune dépend étroitement du type de corpus. Nous avons remarqué[196] que la plupart des textes sont journalistiques ce qui implique l'étude d'un matériau linguistique très spécialisé et freine alors la performance de l'algorithme qui 'a appris' les chaînes référentielles de son corpus d'apprentissage. L'application des systèmes par apprentissage sur d'autres textes s'avèrerait alors non réussie car ils éprouveraient des difficultés lors de l'identification des différents termes de la référence. A notre connaissance, il n'existe pas d'évaluation qui tente d'adopter ce genre de système sur des textes 'inconnus'.

Concernant le français écrit, c'est l'absence d'un corpus large, annoté en coréférence et en accès libre qui bloque la résolution automatique de l'anaphore. Nous avons étudié ces limites dans la première partie et nous pouvons ajouter que, malgré les efforts des chercheurs francophones ces dernières années, une résolution totalement automatique de la référence en français n'est pas disponible.

Afin de surmonter ces limites, des chercheurs ont opté pour une approche hybride qui combine l'approche symbolique et l'approche par apprentissage. Nous citons comme exemple Weissenbacher (2008), qui a proposé un système reposant sur la classification

[196] Nous avons déjà discuté cette limite, de manière plus systématique dans ce qui précède, notamment dans le Chapitre Analyse de corpus de la première partie. Nous allons ici l'évoquer brièvement, pour éviter de faire quelques redites.

bayésienne[197]. Sa modélisation permet de combiner des connaissances linguistiques avec des indices de surface : un premier module sélectionne un antécédent candidat parmi d'autres en appliquant les facteurs avancés par Mitkov (2002). Le deuxième module ajoute des indices à l'antécédent depuis les annotations utilisées dans le corpus d'entraînement. Ce système hybride a eu 61% comme taux de réussite mais a rencontré plusieurs erreurs liées au calcul de la saillance. Ainsi, l'approche hybride semble être prometteuse pour améliorer les systèmes de résolution de l'anaphore malgré qu'elle reste étroitement liée à l'existence d'un grand corpus annoté en référence au préalable. Cette condition freine déjà l'application de telles approches en français.

A la lumière des systèmes cités ci-dessus et de leurs lacunes, nous avancerons dans ce qui suivra notre positionnement et justifierons notre choix de RESUMAN_A.

3.2. Bilan : Algorithme général de RESUMAN_A

Etant donné que RESUMAN_C n'est pas un corpus annoté en coréférence et qu'il n'appartient pas à la catégorie fréquente des textes journalistiques, nous ne pouvons pas, à l'heure actuelle, utiliser l'approche par apprentissage. Néanmoins, et dans le cadre de notre travail, non inscrit à un projet ni financé[198], nous ne pouvons pas annoter RESUMAN_C pour adopter une technique hybride. Cela dit, nous allons utiliser l'approche symbolique reposant sur des connaissances linguistiques. Dans la littérature, les systèmes proposés recourent à des connaissances linguistiques complexes, syntaxiques et sémantiques, déterminant l'anaphore au niveau de trois étapes[199] : l'identification des pronoms anaphoriques est la première étape dans leur résolution. Ensuite, pour chaque pronom, l'algorithme doit lui attribuer un antécédent. Lorsque la substitution entre l'anaphore et son antécédent n'entraîne pas une déviation sémantique, alors une anaphore est correctement résolue. Nous avons choisi d'utiliser cet algorithme, parce qu'il peut s'adapter facilement à notre problématique de résolution de pronoms de troisième personne. De plus, cet algorithme est toujours au cœur de des systèmes au niveau de l'état de l'art et a le mérite d'être simple à implémenter.

---

[197] Pour plus d'informations, *Cf.* Weissenbacher (2008).

[198] Le système RefGen (Longo, 2013), qui a bénéficié d'un financement total de 168 K€, s'inscrit dans le cadre de l'approche pauvre en connaissances.

[199] Ces étapes ne sont pas respectées par certains algorithmes de résolution automatique des anaphores, par exemple celui de Dagan & Itai (1990). En effet, il se base sur une sélection de l'antécédent en même temps qu'il met en place une liste de termes susceptibles d'être des antécédents à l'antécédent lui-même. D'après Kehler et *al.* (2001), cette technique « n'améliore [ent] pas les performances d'un système qui exploite déjà des informations morphosyntaxiques ».

Une stratégie similaire à celle utilisée dans les systèmes développés par Lappin/Leass (1994) et Mitkov (1998) est mise en œuvre dans notre travail pour l'interprétation des expressions pronominales dans $RESUMAN_c$. La structure syntaxique du texte, les informations de nature morphologique (genre et nombre) et une quantité réduite d'information de sémantique lexicale (noms propres des personnes ou bien entités nommées) constituent essentiellement l'information que nous utiliserons. C'est la précision de l'interprétation d'un sous-ensemble des expressions pronominales en français qui est visée par notre système de résolution des pronoms RESUMAN (*il, ils, elle, elles*)[200]. Les étapes suivantes constituent le processus de la résolution des pronoms :

1) **Distinction des pronoms anaphoriques :** tous les pronoms de $RESUMAN_C$ doivent être répertoriés, dans cette première étape, en pronoms anaphoriques et en pronoms impersonnels. Nous aurons besoin d'un classifieur des pronoms dans notre corpus. Nous utiliserons l'analyseur syntaxique Fips qui a été élaboré par Laenzlinger et Wehrli en 1991 à l'ATALA[201] et qui a réussi à résoudre, en recourant à un système de classification automatique, le problème de la distinction des pronoms anaphoriques des pronoms impersonnels. Ces derniers ne constituant aucune entité sémantiquement et ne renvoyant à aucun élément du discours, cela évite alors de rechercher inutilement leur antécédents.

2) **Sélection des candidats :** dans cette deuxième étape, afin de pouvoir établir des possibles antécédences entre les pronoms anaphoriques, une liste des candidats possibles est mise en place. L'algorithme se limite à la détermination des syntagmes nominaux de la phrase où apparaît le pronom ainsi que ceux des phrases précédentes pour les anaphores pronominales.

3) **Choix de l'antécédent :** au niveau de cette dernière étape, pour résoudre le pronom, l'algorithme détermine le candidat qui sera l'antécédent proposé par le système. Deux possibilités, souvent exploitées simultanément, sont envisageables :

---

[200] Pour la première version de $RESUMAN_O$, nous avons choisi de mettre l'accent sur un petit ensemble d'expressions pronominales, nous espérons dans l'avenir le généraliser à d'autres expressions pronominales.
[201] Association pour le Traitement Automatique des Langues : http://www.atala.org/

- C’est selon des contraintes que l’algorithme doit éliminer les candidats non potentiels. Les contraints les plus utilisées sont généralement l’accord de genre et de nombre entre le pronom et le candidat.

- Les candidats restant seront privilégiés selon des préférences qui sont des critères tirés de Lappin/Leass (1994) et Mitkov (2002) : si pour un couple pronom-antécédent, il reste plus d'un antécédent possible, l’algorithme réduit l’antécédent à un seul élément basé sur ces préférences ordonnées.

En ce qui concerne notre stratégie générale adoptée et les informations utilisées, nous faisons remarquer que notre approche ne diffère pas de façon significative d'autres systèmes symboliques reposant sur des connaissances linguistiques mis en œuvre à ce jour. Cependant, notre système a quelques caractéristiques intéressantes : Le fait par exemple qu’il soit mis en œuvre en utilisant un seul outil unique pour l'analyse syntaxique et morphologique et l’élimination des pronoms impersonnels (cela démontre par ailleurs la puissance expressive de Fips) et le fait qu'il utilise un ensemble de préférences ordonnées pour choisir le meilleur antécédent parmi un ensemble des candidats. Cette dernière propriété permet une évaluation indépendante de chaque préférence, offrant ainsi une meilleure compréhension du processus de résolution de pronom.

En résumé, le bref inventaire que nous avons présenté dans ce chapitre nous a permis d’étudier l’évolution des principaux algorithmes disponibles en comparant leurs performances. Nous en avons ensuite tiré notre propre algorithme en prenant en compte les points faibles des travaux précédents pour les éviter. Dans le chapitre suivant, nous détaillerons l’approche que nous proposons pour la résolution automatique de l’anaphore pronominale dans RESUMAN$_C$.

Chapitre 3 : RESUMAN$_O$, un outil de résolution automatique de l'anaphore pronominale

Nous présentons dans ce chapitre l'outil RESUMAN$_O$ que nous avons créé pour résoudre automatiquement les anaphores pronominales de notre corpus. Après description de l'architecture générale de RESUMAN$_O$, nous en exposons en détail ses étapes. Notre approche se sert des concepts issus des travaux linguistiques ainsi que d'un certain nombre de contraintes et d'hypothèses de travail exposées dans le précédent chapitre.

1. Architecture générale de l'approche proposée

RESUMAN$_O$ est un outil de résolution automatique des anaphores pronominales dans un corpus textuel brut. Cet outil combine deux modules principaux : un module de prétraitement de corpus et un module de résolution des anaphores pronominales trouvées dedans (*cf.* Figure 31).

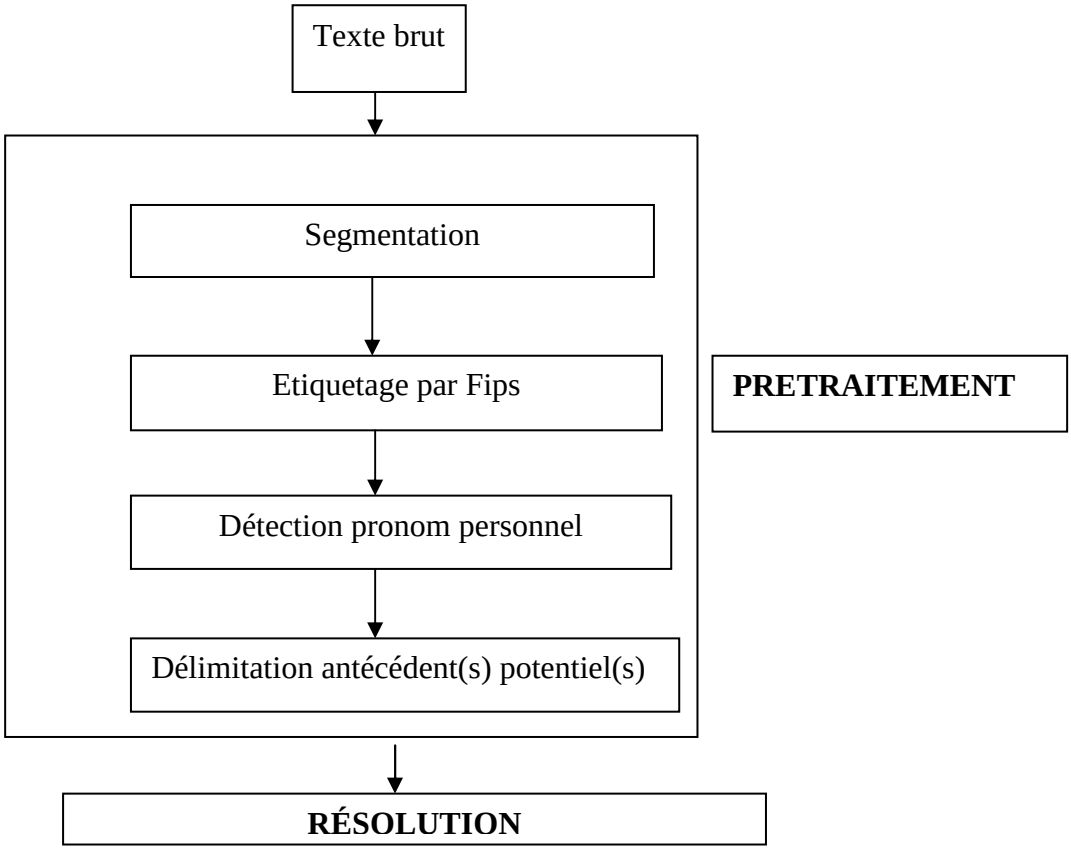


Figure 31 : Architecture globale de RESUMAN$_O$

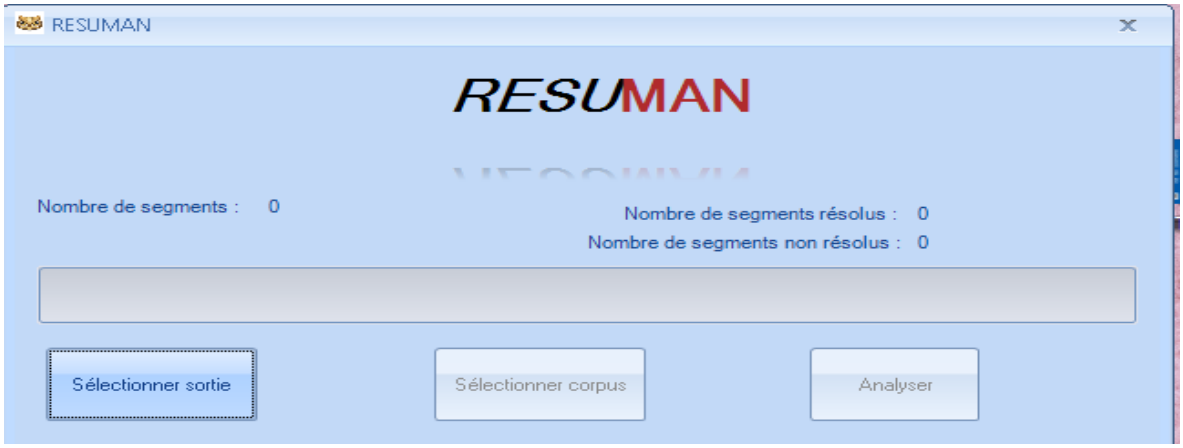
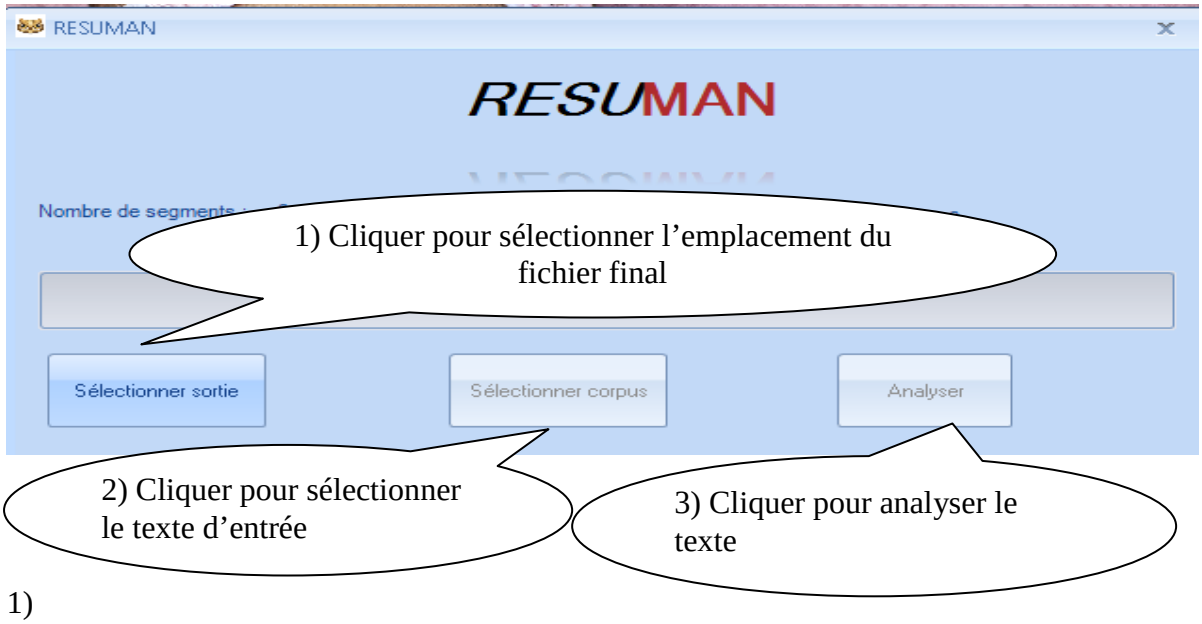


Figure 32 : Interface de RESUMAN$_O$

Au départ, le texte brut est extrait à partir du corpus RESUMAN$_C$. A l'issue de l'extraction, il est découpé en des segments[202] selon deux critères : la ponctuation finale et la présence d'un pronom personnel (*il*, *ils*, *elle*, *elles*). Ensuite intervient le module d'étiquetage morphosyntaxique, pour lequel nous avons utilisé une approche basée, d'une part, sur l'analyseur syntaxique Fips et, d'autre part, sur un extracteur des entités nommées que nous avons développé avec l'aide de l'informaticienne en collaboration ; les segments textuels sont annotés par Fips : l'appel à Fips se faisant directement en ligne, le poste de travail doit être impérativement connecté à un réseau en libre accès. La tâche de détection des pronoms et des groupes nominaux se réalise par synchronisation avec Fips. Le module de la résolution est la dernière étape du processus.

Avant de détailler toutes ces étapes, en les illustrant avec des exemples, dans les sections suivantes, nous tenons à présenter et décrire l'interface de RESUMAN$_O$[203]. Elle est composée de trois boutons : Sélectionner sortie, Sélectionner corpus et Analyser. Leurs fonctions sont expliquées ci-dessous.

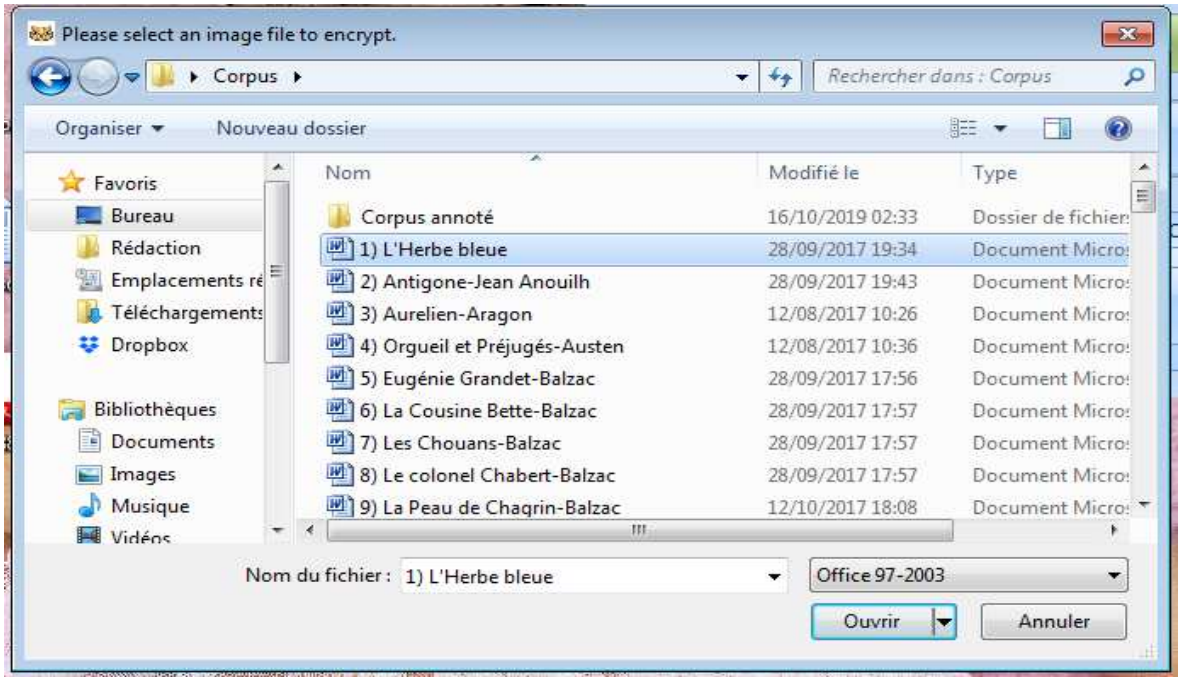
[202] Nous optons pour ce terme neutre afin de dénommer une réalité topologique.
[203] Nous avons choisi comme logo la chouette, symbole de Dijon.


RESUMAN
RESUMAN
Nombre de segments
1) Cliquer pour sélectionner l'emplacement du fichier final
Sélectionner sortie
Sélectionner corpus
Analyser
2) Cliquer pour sélectionner le texte d'entrée
3) Cliquer pour analyser le texte

1)


Enregistrer un document de resultat
Corpus
Rechercher dans : Corpus
Nom du fichier : Test
Type : Microsoft Word 97-2003
Parcourir les dossiers
Enregistrer
Annuler

2)


Please select an image file to encrypt.
Corpus
Rechercher dans : Corpus
Organiser
Nouveau dossier
Favoris
Bureau
Rédaction
Emplacements ré
Téléchargements
Dropbox
Bibliothèques
Documents
Images
Musique
Vidéos
Nom
Modifié le
Type
Corpus annoté 16/10/2019 02:33 Dossier de fichier
1) L'Herbe bleue 28/09/2017 19:34 Document Micro
2) Antigone-Jean Anouilh 28/09/2017 19:43 Document Micro
3) Aurelien-Aragon 12/08/2017 10:26 Document Micro
4) Orgueil et Préjugés-Austen 12/08/2017 10:36 Document Micro
5) Eugénie Grandet-Balzac 28/09/2017 17:56 Document Micro
6) La Cousine Bette-Balzac 28/09/2017 17:57 Document Micro
7) Les Chouans-Balzac 28/09/2017 17:57 Document Micro
8) Le colonel Chabert-Balzac 28/09/2017 17:57 Document Micro
9) La Peau de Chagrin-Balzac 12/10/2017 18:08 Document Micro
Nom du fichier : 1) L'Herbe bleue
Office 97-2003
Ouvrir
Annuler

3)

Figure 33 : Description de l'interface de RESUMAN$_O$

2. Prétraitement de RESUMAN$_C$

Chaque outil nécessite une entrée, brute ou annotée au préalable, qui subit les différentes étapes du traitement informatique pour parvenir à l'objectif recherché.

2.1. Segmentation du corpus

La première étape de notre algorithme est le découpage du résumé en des phrases. Nous considérons que chaque phrase se termine par un signe de ponctuation (nous avons exclu la virgule).

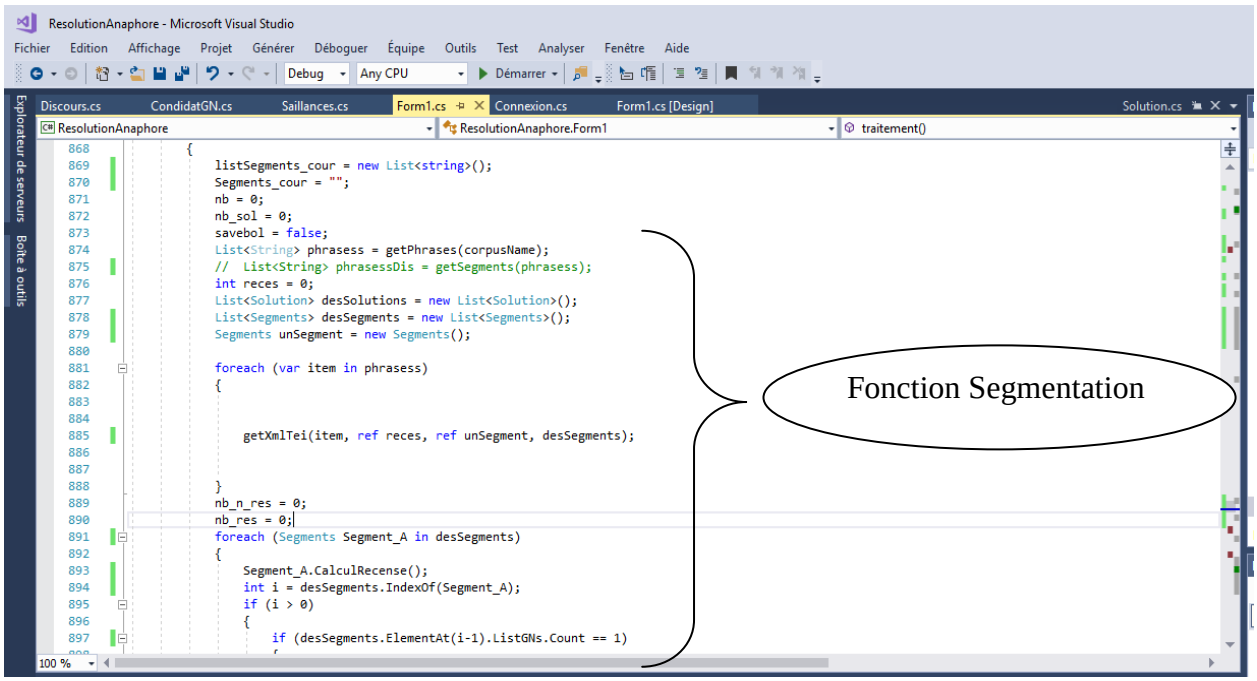


Figure 34 : La fonction Segmentation

Nous prenons l’extrait suivant comme exemple tout au long de notre explication et démonstration :

> [1] Antigone rentre chez elle, à l'aube, après une escapade nocturne. Elle est surprise par sa nourrice qui lui adresse des reproches. L'héroïne doit affronter les questions de sa nounou. Le dialogue donne lieu à un quiproquo. La nourrice prodigue des conseils domestiques (" il va falloir te laver les pieds avant de te remettre au lit") tandis qu'Antigone évoque son escapade avec beaucoup de mystère (" oui j'avais un rendez-vous"). Mais elle n'en dira pas plus. (Résumé Antigone)

Antigone rentre chez elle, à l'aube, après une escapade nocturne.
Elle est surprise par sa nourrice qui lui adresse des reproches.
L'héroïne doit affronter les questions de sa nounou.
Le dialogue donne lieu à un quiproquo.
La nourrice prodigue des conseils domestiques il va falloir te laver les pieds avant de te remettre au lit tandis qu'Antigone évoque son escapade avec beaucoup de mystère oui j'avais un rendez-vous.
Mais elle n'en dira pas plus.

Figure 35 : Exemple de corpus en phrases

Après avoir découpé les phrases du corpus, nous allons créer des segments : un segment est un ensemble de phrases jusqu’à atteindre une phrase qui contient un pronom personnel (*il, ils, elle, elles*)

[Antigone rentre chez elle, à l'aube, après une escapade nocturne.]

[Elle est surprise par sa nourrice qui lui adresse des reproches.]

[L'héroïne doit affronter les questions de sa nounou.
Le dialogue donne lieu à un quiproquo.
La nourrice prodigue des conseils domestiques il va falloir te laver les pieds avant de te remettre au lit tandis qu'Antigone évoque son escapade avec beaucoup de mystère oui j'avais un rendez-vous.]
[Mais elle n'en dira pas plus.]

Figure 36 : Exemple de corpus en segment

2.2. Étiquetage morphosyntaxique

L'étiquetage morphosyntaxique du corpus est l'étape qui suit celle de segmentation. L'étiquetage morphosyntaxique d'un texte est une étape fondamentale de son analyse, et un préliminaire à tout traitement de plus haut niveau. L'objectif de l'étiquetage est d'attribuer à chacun des mots d'un corpus une étiquette qui récapitule ses informations morphosyntaxiques. Le processus d'étiquetage peut accompagner la lemmatisation dont l'objectif est de ramener l'occurrence d'un mot donné à sa forme de base ou « lemme ». L'étiquetage morphosyntaxique permet d'envisager des recherches non plus sur des formes particulières telles qu'elles se rencontrent dans les textes (chaînes de caractères) mais aussi sur des lemmes (formes canoniques) ou encore sur des catégories syntaxiques. L'entrée de notre outil est un résumé brut, transformé en une représentation morphosyntaxique fournie par l'analyseur Fips[204], ce dernier prodiguant des informations sur les caractéristiques lexicales, morphologiques et syntaxiques pour tous les éléments d'une phrase dans leur contexte. L'exemple suivant, extrait de notre corpus $RESUMAN_C$, permet de mieux comprendre le fonctionnement de Fips :

[204] Pour accéder à Fips : http://latlapps.unige.ch/Parser
Pour une description détaillée de Fips : http://alpage.inria.fr/iwpt09/atala/fips.pdf

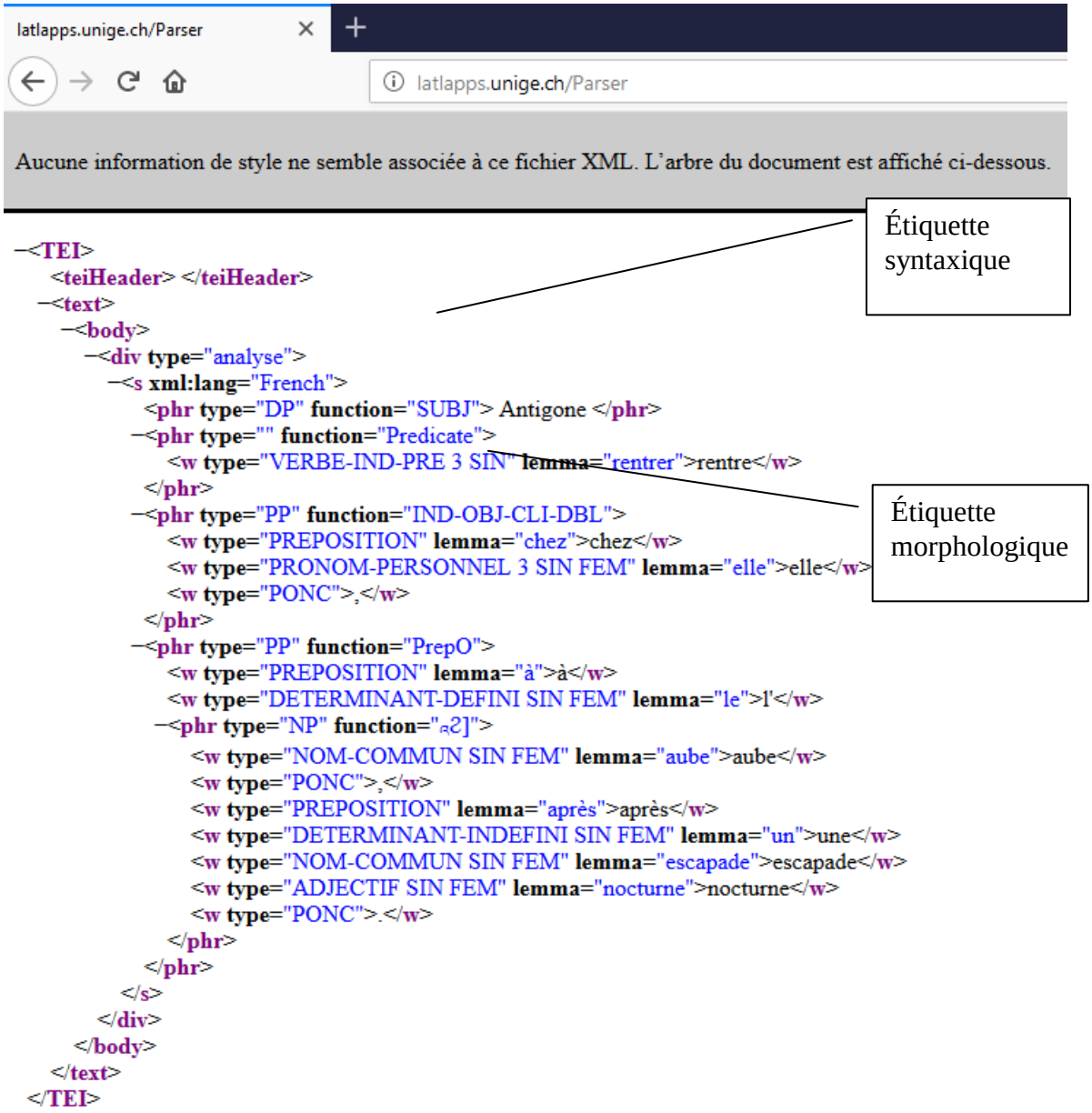
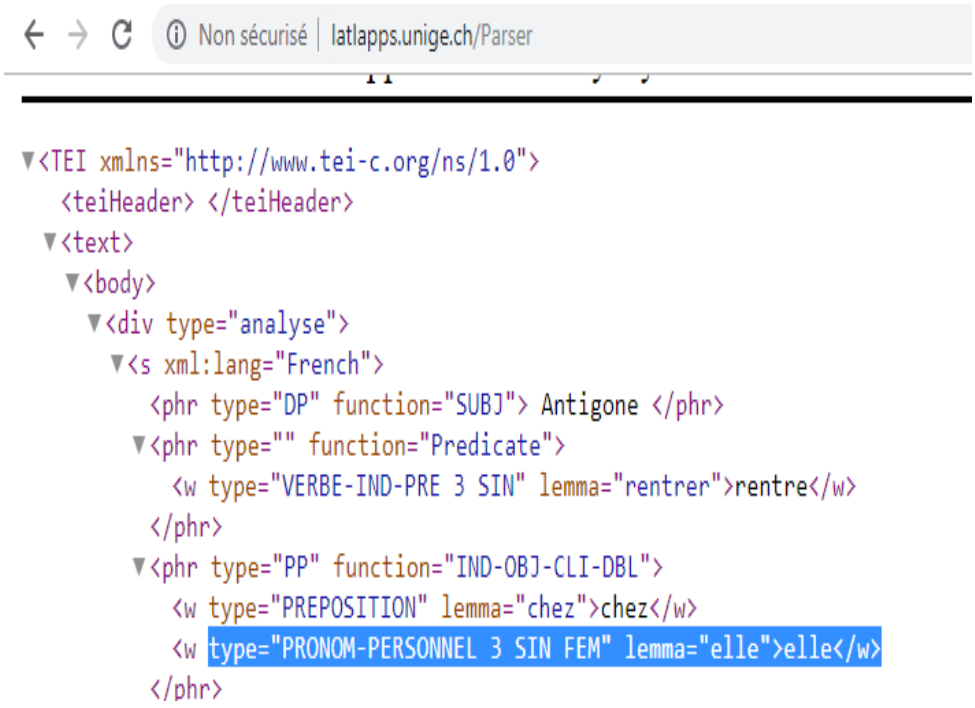


Figure 37 : Exemple de l'étiquetage morphosyntaxique par Fips

C'est dans la représentation syntaxique d'entrée que le processus recherche les pronoms. Quand un pronom est détecté, le texte qui le précède est délimité par le processus : lors de l'étiquetage morphosyntaxique du corpus, à chaque fois que `type="PRONOM-PERSONNEL"` apparaît, le pronom est stocké dans une liste de pronoms[205].

[205] Nous avons deux listes : une contenant les pronoms personnels et l'autre les candidats potentiels.

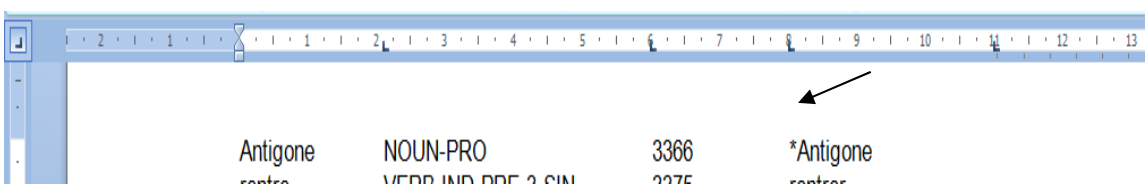


L'élimination des pronoms impersonnels est une étape importante pour éviter toute confusion avec le pronom personnel *il*. Nous rejoignons Danlos (2005 : 390) dans le sens où :

> Un système de résolution des anaphores doit être capable de repérer les occurrences des pronoms impersonnels avant de s'attaquer aux pronoms anaphoriques et aux autres anaphores.

Grâce à Fips, cette tâche nous ne pose pas de problème : RESUMAN$_O$ ignore les pronoms impersonnels `type="PRON-IMPER-SIN-MAS"`.

Après l'identification des pronoms personnels du corpus, une étape d'extraction des antécédents potentiels intervient. Il y a examen du corpus dans le but d'identifier les différents groupes nominaux candidats, suite à l'examen des pronoms et des segments précédents : chaque segment du texte est parcouru et chaque GN (nom commun (NOUN-COM) ou propre (NOUN-PRO)) est ajouté à une liste de candidats, même dans le cas où ce dernier se trouve dans un groupe verbal. Puisque Fips n'attribue pas d'étiquette morphologique aux noms propres (ils sont précédés d'un astérisque),



nous avons créé une base de données semi-automatique[206] contenant tous les noms propres de notre corpus : dans un premier temps, nous avons extrait les tokens avec type=" NOUN-PRO". Ensuite, nous les avons annotés manuellement en rajoutant le genre et le nombre et en respectant les normes d'écriture de Fips pour conserver l'homogénéité des étiquettes du corpus.

L'outil attribue les paramètres Genre, Nombre et Fonction à chaque nom comme suit :

a. Identifier les groupes nominaux.
b. Attribuer pour chaque groupe nominal les paramètres :
    - Genre (masculin/féminin) à l'aide de l'analyseur Fips et la base de donnée crée manuellement pour les noms propres ;
    - Nombre (singulier/pluriel) à l'aide de l'analyseur Fips et la base de donnée crée manuellement pour les noms propres ;
    - Fonction (COD, COI, sujet, attribut, subordonné) à l'aide de l'analyseur Fips.

Ci-dessous une démonstration de l'étiquetage du nom propre *Antigone* :

[Antigone rentre chez elle, à l'aube, après une escapade nocturne.]

➢ Résultat Fips :

<phr type="DP" function="SUBJ"> Antigone</phr>

➢ Selon l'analyseur Fips, la fonction d'*Antigone* est sujet function="SUBJ" ;

➢ Pour le genre et le nombre, on voit que Fips n'a pas attribué une valeur à *Antigone* type="DP".

➢ L'outil fait appel à la base des noms propres pour compléter le champ morphologique :

[206] Nous en avons parlé dans la section 1.1 du chapitre 3 de la deuxième partie (*Cf.* Figure 13 et Annexe).

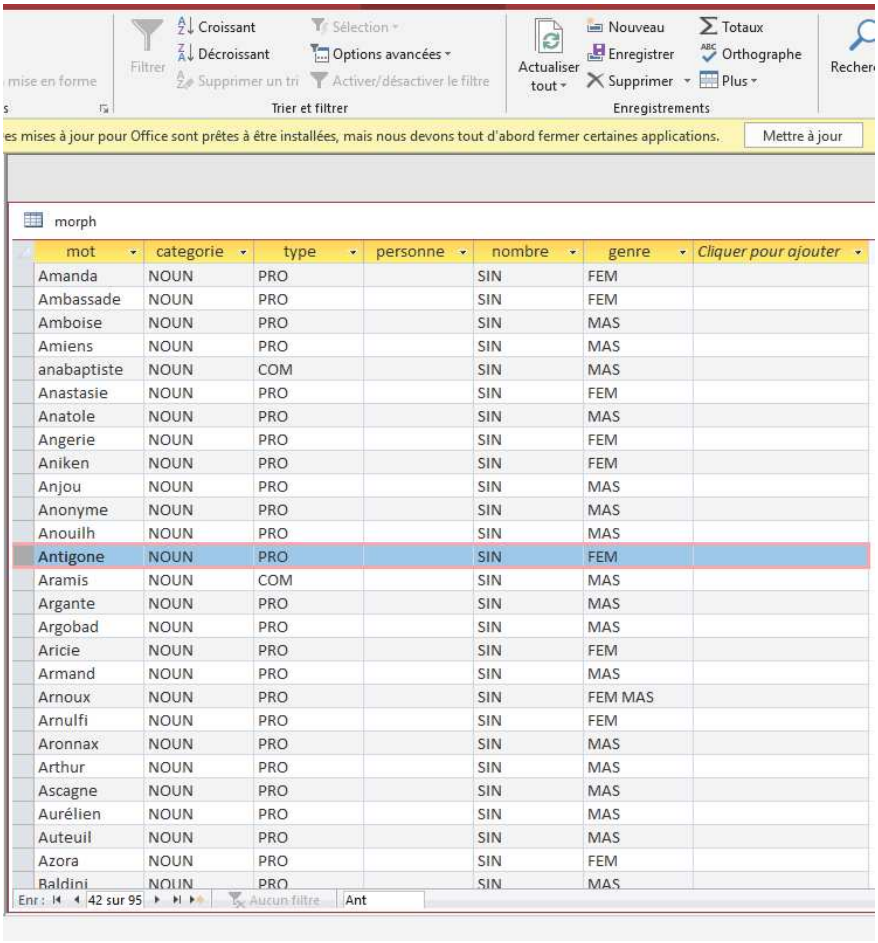


Figure 38 : Exemple de la base de données des noms propres de RESUMAN$_C$

Cette étape permet d'empiler les GN et les pronoms dans une pile au fur et à mesure de la lecture du texte tout en mémorisant leur ordre d'apparition dans le texte. Cela servira par la suite pour décrire la récence des GN. Après avoir traité le corpus (segmentation, étiquetage et délimitation des deux listes de pronoms et de candidat(s) potentiel(s)), les segments subissent les étapes de la résolution pour attribuer à chaque pronom son bon antécédent. Nous détaillons ces étapes dans la section suivante.

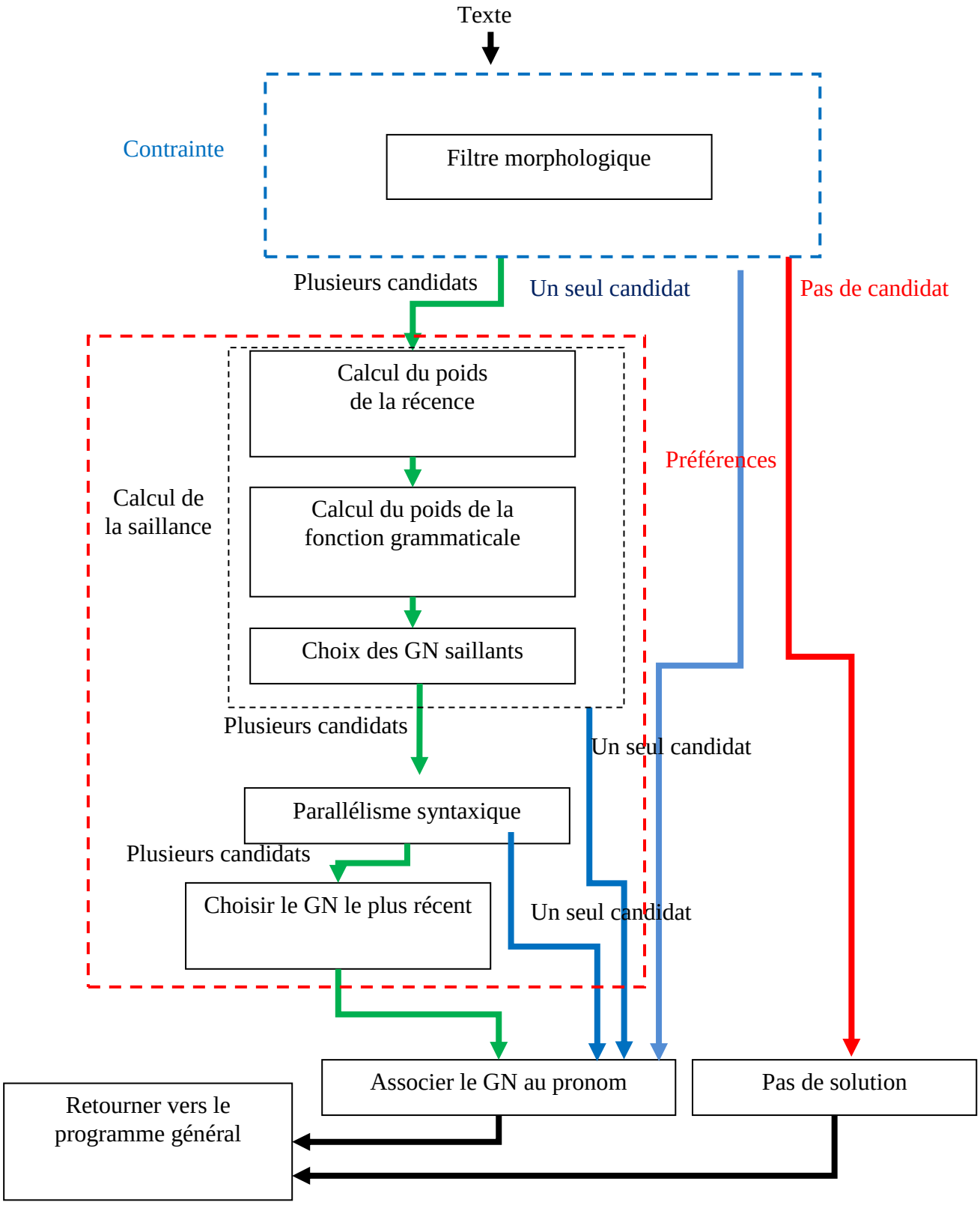


Figure 39 : Schéma de l'algorithme RESUMAN$_O$

3. Module Résolution de RESUMAN$_O$

L'approche que nous exposons permet de résoudre les anaphores pronominales dans un texte. Nous nous sommes servie de Fips, une base de données de noms propres que nous avons développé et des concepts issus de travaux linguistiques. Les règles de résolution d'anaphores pronominales basées sur les différentes sources de connaissances et utilisées durant le processus de résolution sont désignées sous le terme de facteurs. Ces facteurs peuvent être divisés en deux catégories : les contraintes et les préférences.

3.1. Les contraintes

Les contraintes représentent la première catégorie des facteurs de la résolution des anaphores. Elles se réfèrent par exemple aux accords en genre et en nombre, aux contraintes de commandement *(C-command)*, etc. Selon une contrainte, parmi la totalité des candidats potentiels, certains GN peuvent être supprimés ou, au contraire, d'autres peuvent être privilégiés. Nous adoptons seulement le filtre morphologique. Les autres contraintes ne correspondent pas à notre approche.

Une contrainte assez forte est celle se rapportant à l'accord en genre et en nombre de l'antécédent avec le pronom. Cette contrainte permet d'éliminer un certain nombre de candidats qui ne respectent pas les conditions mentionnées. Néanmoins, cette seule contrainte est insuffisante, puisqu'un candidat s'accordant en genre et en nombre pourrait être associé de manière erronée à un pronom, ce après l'exclusion des antécédents qui ne respectent pas cette contrainte.

Nous avons ainsi complété cette contrainte par des hypothèses formulées par Dimitrov *et al.* (2005) à partir d'analyses de corpus de textes et étayées par des études statistiques :

- Il est très probable qu'un antécédent se trouve dans la même phrase que le pronom auquel il se rapporte.
- Il est important de prendre en considération la récence (*recency factor*) qui stipule qu'en cas de doute sur l'antécédent, le plus récemment identifié est probablement le bon.
- Même si les liens anaphoriques sont généralement intraphrasiques, un antécédent peut-être identifié dans une phrase précédente. Selon Dimitrov *et al.* (2005), la recherche d'antécédents devrait s'effectuer dans les 3 à 4 phrases précédentes, car l'augmentation de la taille de l'espace de recherche implique une augmentation du coût de traitement. Il est ainsi important de limiter ce coût sans

pour autant compromettre le taux de réussite. Néanmoins Mitkov (1999) rapporte dans son étude que l'antécédent a été retrouvé dans la 17ème phrase précédant celle où se trouve l'anaphore. De ce fait, nous sommes consciente qu'il est difficile de définir un espace de recherche qui donne une solution définitive et nous prenons le parti de fournir une solution acceptable. Ainsi, nous avons choisi un contexte médiane, c'est-à-dire, l'outil cherche jusqu'à la 10ème phrase précédente le pronom.

Reprenons alors notre exemple :

[2] Antigone rentre chez elle, à l'aube, après une escapade nocturne. Elle est surprise par sa nourrice qui lui adresse des reproches. L'héroïne doit affronter les questions de sa nounou. Le dialogue donne lieu à un quiproquo. La nourrice prodigue des conseils domestiques (" il va falloir te laver les pieds avant de te remettre au lit") tandis qu'Antigone évoque son escapade avec beaucoup de mystère (" oui j'avais un rendez-vous"). Mais elle n'en dira pas plus. (Résumé Antigone)

a. Si le nombre de groupes nominaux est égal à 1, l'anaphore est résolue, et le pronom se réfère à ce groupe nominal.

   - Antigone rentre chez elle (Antigone), à l'aube, après une escapade nocturne

b. Si le nombre de groupes nominaux est supérieur à 1, l'outil élimine des antécédents en se reposant sur la contrainte morphologique :

   - [L'**héroïne** doit affronter les **questions** de sa **nounou**.
     Le **dialogue** donne lieu à un **quiproquo**.
     La **nourrice** prodigue des **conseils** domestiques il va falloir te laver les **pieds** avant de te remettre au **lit** tandis qu'**Antigone** évoque son **escapade** avec beaucoup de **mystère** oui j'avais un **rendez-vous**.]
     [Mais **elle** n'en dira pas plus.]

   - Avant le pronom *elle*, on compte 13 groupes nominaux.
   - le filtrage morphologique permet de réduire cet ensemble en 3 groupes nominaux (*héroïne*, *nourrice*[207], *Antigone*)

[207] Il faut noter que la préposition est toujours éliminée de la segmentation si le GN se trouve dans un groupe prépositionnel (GP). Par exemple, dans :
L'héroïne doit affronter les questions de sa nounou.
Ce que l'on garde ce n'est pas *de sa nounou* ou encore *de nounou* mais bien *sa nounou*.

- Si, après le filtrage morphologique, il ne reste qu'un seul groupe nominal, l'anaphore est résolue. Si non, on passe au calcul de la saillance (section suivante).

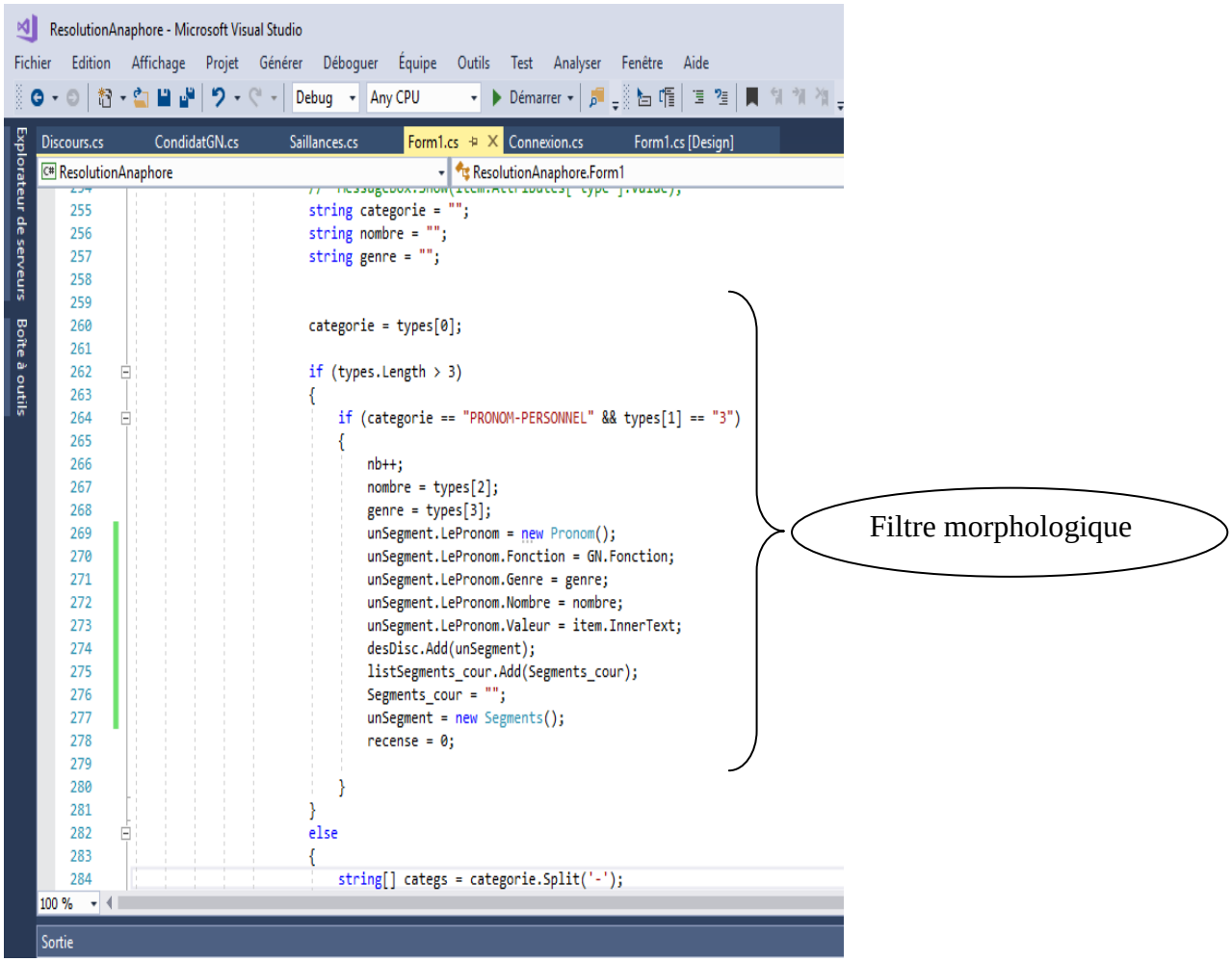


Figure 40 : Filtre morphologique de RESUMAN$_O$

3.2. Les préférences

À l'inverse des contraintes, les préférences ne sont pas nécessairement prises en compte par tous les chercheurs dans le domaine de la résolution de l'anaphore. Leur choix, justifiable dans tous les cas, reste arbitraire. Deux types de préférences de choix sont considérés dans notre approche:

- Préférence du sujet

  Quand un GN possède la fonction syntaxique de sujet, on va le **préférer** dans le processus de résolution d'anaphore puisque souvent le sujet est le centre de la phrase (SVO).

- Préférence du candidat le plus récent : l'algorithme consiste en un parcours en largeur de gauche à droite avec une **préférence** pour l'antécédent le plus proche du pronom.

- Parallélisme syntaxique

  Quand l'antécédent demeure ambigu, le parallélisme syntaxique entre en jeu. Celui-ci donne la **préférence** aux GN possédant la même fonction syntaxique que le pronom.

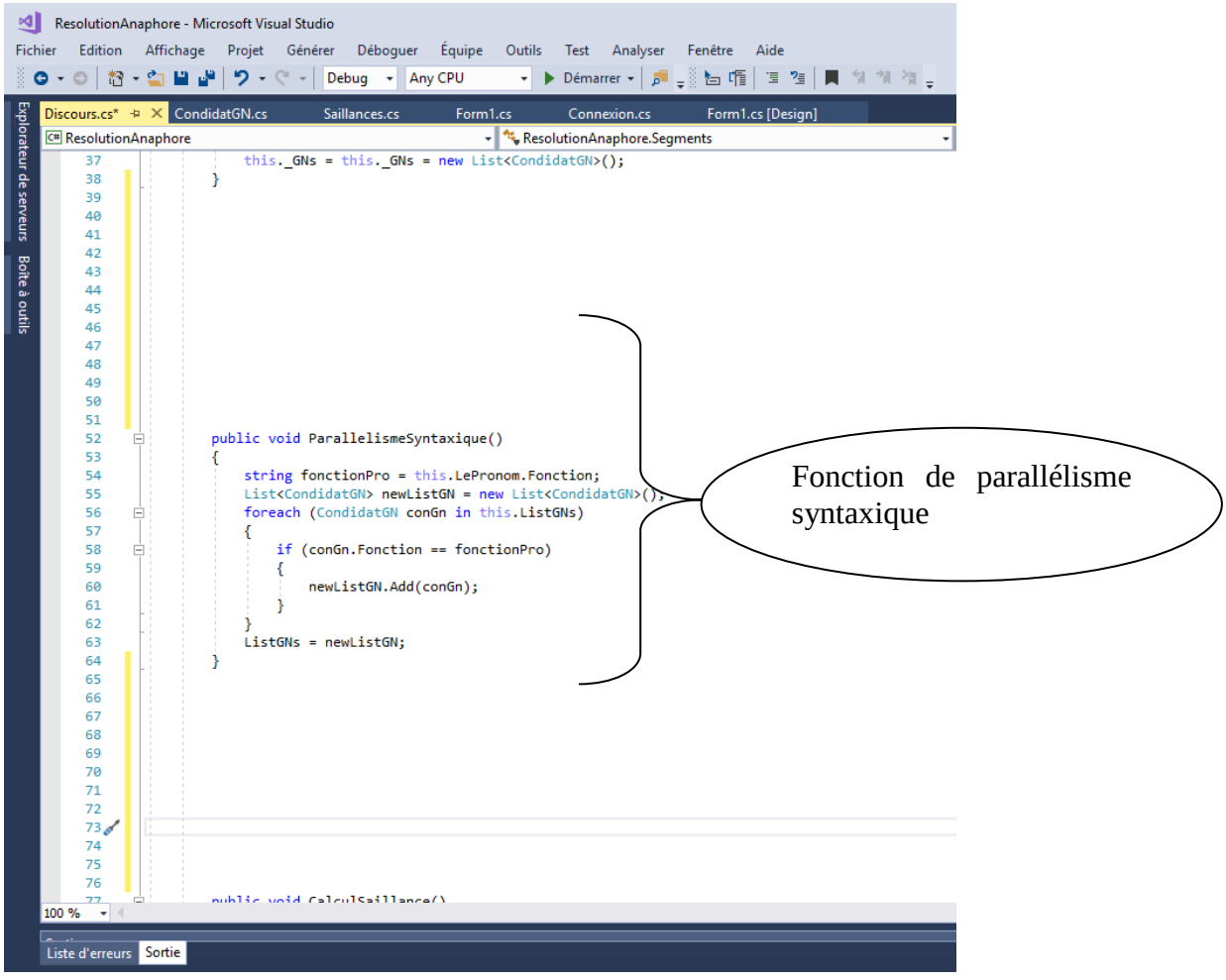


Figure 41 : Fonction de parallélisme syntaxique

La résolution de l'anaphore ne pose pas de problème si au final le processus sélectionne un seul élément. Dans le cas où plusieurs éléments sont sélectionnés, alors la saillance est prise en compte. Si la saillance ne résout pas le problème alors le parallélisme syntaxique entre en jeu, et si le conflit n'est toujours pas résolu alors c'est l'élément le plus proche du pronom qui sera sélectionné.

3.3. Calcul de la saillance

Ce sont les facteurs de saillance que possède un référent, avec différents poids, qui permettent le calcul de la saillance. Ce calcul est simplement la somme de ces différents facteurs : la récence et le poids grammatical de chaque GN.

3.3.1. La récence de phrase

La portée du facteur de récence est maximale pour les référents de la phrase courante et cette portée diminue si c'est la phrase précédente du texte qui est considérée. Ariel (1990 : 20), Dimitrov et al. (2005) et Demol (2007) remarquent que la plupart des anaphores possèdent leur référent dans la phrase courante et qu'en moyenne dans un dixième des cas, le référent n'est ni dans la phrase courant ni dans la précédente. Nous pouvons récapituler le pourcentage de la distance des anaphores et de leurs antécédents dans le tableau suivant.

Position du groupe nominal et du pronom	Valeur de la récence
Dans la même phrase	100
Dans la phrase précédente	Diminution de 10 à chaque retour en arrière
Dans la 10ème phrase précédent l'anaphore	0 (valeur de récence supprimée)

Tableau 15 : Poids de la valeur de la récence

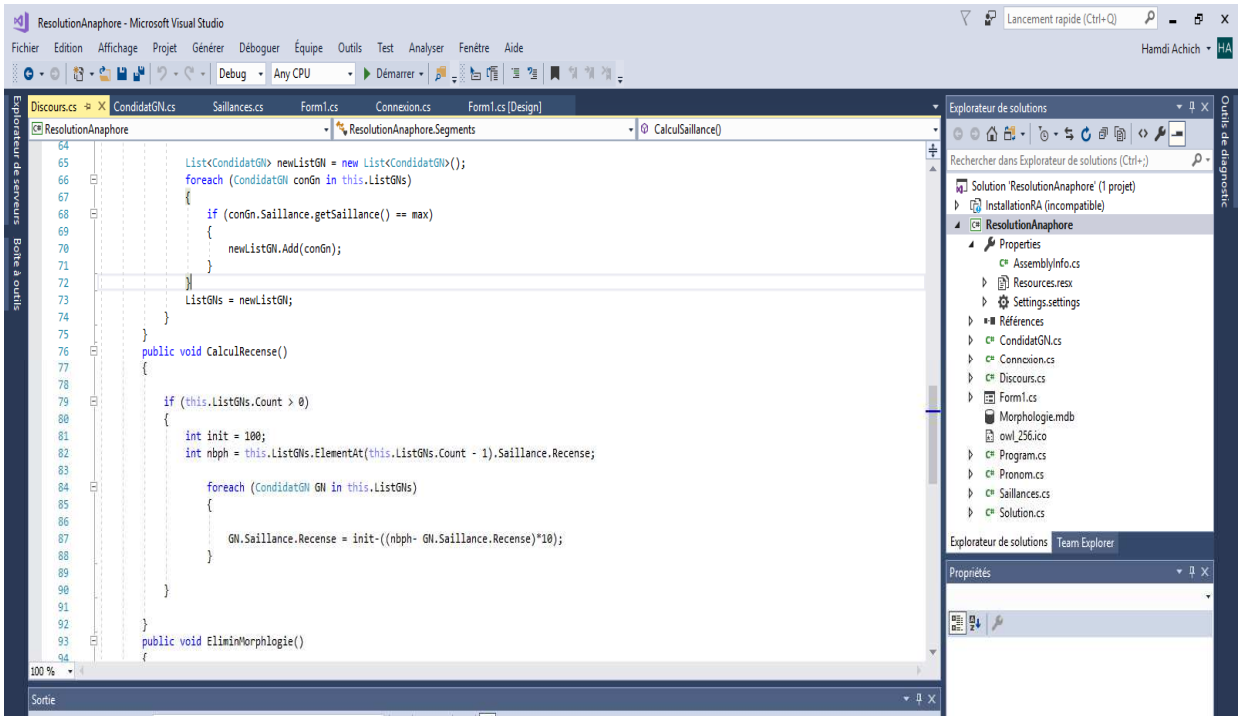


Figure 42 : Calcul du poids de la récence

3.3.2. Poids grammatical[208]

Les règles de la hiérarchie grammaticale prodiguent un poids de saillance plus importants à la fonction *Sujet*. Nous détaillons les différentes valeurs attribuées à chaque fonction dans ce qui suit :

a. Si le GN fait fonction de Sujet, il aura une valeur de 80.

b. Si le GN fait fonction d'objet direct, il aura une valeur de 50. Nous remarquons qu'il est saillant mais pas autant que celui dans la position sujet.

c. Si le GN fait fonction d'objet indirect, il aura une valeur de 40, il est alors moins saillant qu'un GN faisant fonction d'objet direct.

d. Si le GN fait fonction d'attribut, il aura une valeur de 30.

e. Si le GN appartient à une subordonnée, il aura la valeur de 10.

[208] Les valeurs sont arbitraires et reposent sur les principes de la saillance.

Fonction	Poids grammatical
Sujet	80
COD	50
COI	40
Attribut	30
Subordonnée	10

Tableau 16 : Poids des fonctions syntaxiques

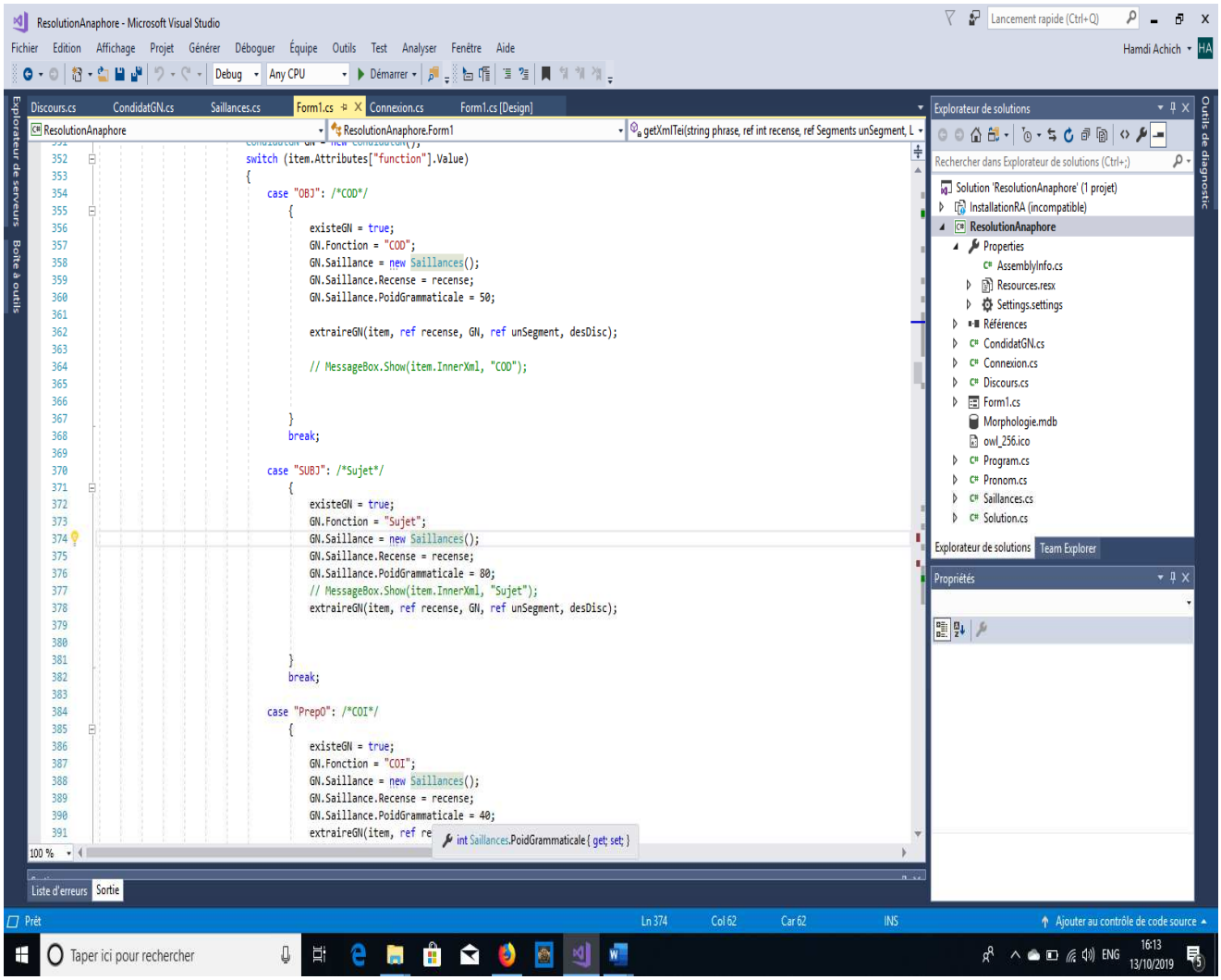


Figure 43 : Calcul du poids de la fonction syntaxique

Nous montrons comment RESUMAN$_O$ calcule la saillance des GN antécédents :

- Calcul de la Saillance (récence + poids grammatical)

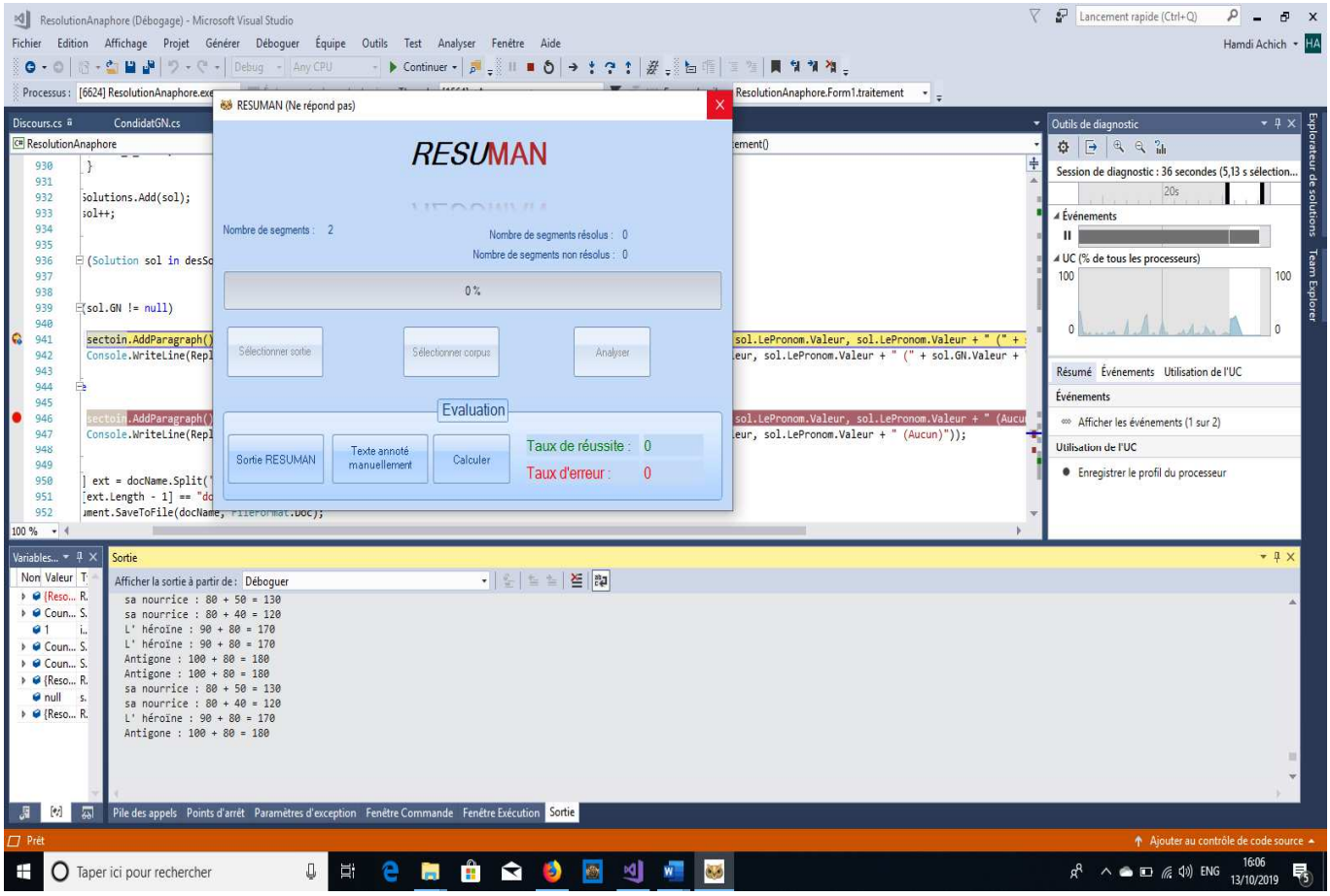


Figure 44 : Calcul de la saillance

Selon ces résultats, *Antigone* a le score le plus élevé (180 = 100 récence + 80 poids grammatical) ; vient après *L'héroïne* avec un score très proche (170 = 90 récence + 80 poids grammatical). Nous expliquons cela par le fait que les deux noms sont coréférents.

Comme résultat, on ne prend que les groupes nominaux qui ont une valeur maximale de saillance.

- Puisque *Antigone* a une valeur plus grande (saillance = 180), le nom est choisi et l'anaphore est résolue.

  Elle -> Antigone

4. Résultats et évaluation

Pour résumer les étapes explicitées dans ce chapitre, nous partons de l'extrait brut de démonstration de notre corpus :

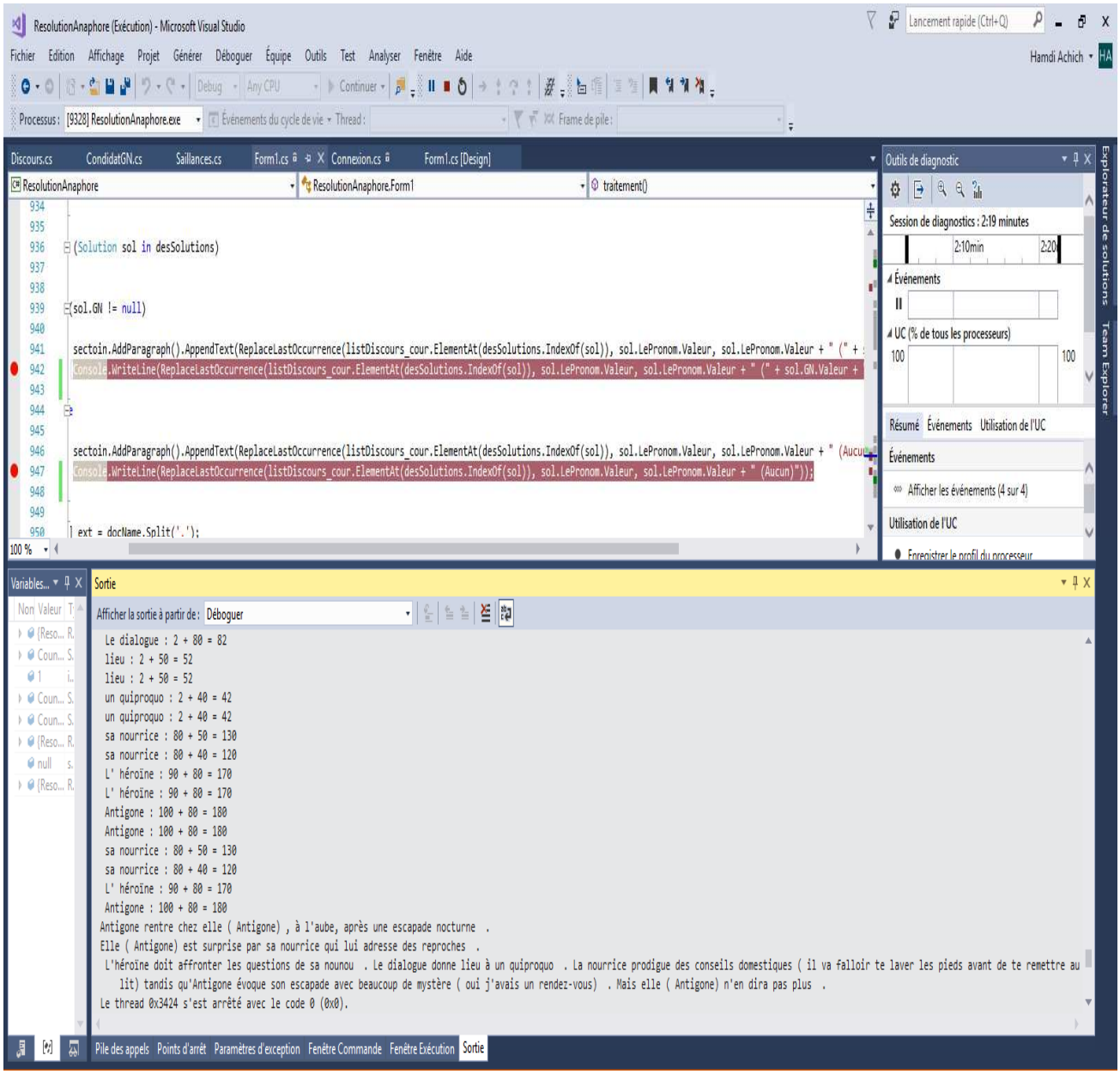
[3] Antigone rentre chez elle, à l'aube, après une escapade nocturne. Elle est surprise par sa nourrice qui lui adresse des reproches. L'héroïne doit affronter les questions de sa nounou. Le dialogue donne lieu à un quiproquo. La nourrice prodigue des conseils domestiques (" il va falloir te laver les pieds avant de te remettre au lit") tandis qu'Antigone évoque son escapade avec beaucoup de mystère (" oui j'avais un rendez-vous"). Mais elle n'en dira pas plus.

que nous avons annoté manuellement afin de le comparer avec le résultat final de $RESUMAN_O$:

[4] Antigone rentre chez elle(Antigone), à l'aube, après une escapade nocturne. Elle(Antigone) est surprise par sa nourrice qui lui (Antigone) adresse des reproches. L'héroïne doit affronter les questions de sa nounou. Le dialogue donne lieu à un quiproquo. La nourrice prodigue des conseils domestiques (" il va falloir te laver les pieds avant de te remettre au lit") tandis qu'Antigone évoque son escapade avec beaucoup de mystère (" oui j'avais un rendez-vous"). Mais elle(Antigone) n'en dira pas plus.

Ci-dessous le texte annoté par $RESUMAN_O$:

[5] Antigone rentre chez elle (Antigone), à l'aube, après une escapade nocturne. Elle (Antigone) est surprise par sa nourrice qui lui adresse des reproches. L'héroïne doit affronter les questions de sa nounou. Le dialogue donne lieu à un quiproquo. La nourrice prodigue des conseils domestiques (il va falloir te laver les pieds avant de te remettre au lit) tandis qu'Antigone évoque son escapade avec beaucoup de mystère (oui j'avais un rendez-vous). Mais elle (Antigone) n'en dira pas plus.



Nous remarquons que tous les pronoms personnels sont résolus et que les paires antécédent/pronom sont équivalentes aux paires annotées manuellement. Pour évaluer notre travail, nous avons choisi de segmenter notre corpus en 3 sous-corpus (Annexe 1, sur CD) :

- Un corpus d'entraînement, ou corpus de travail
- Un corpus de test
- Un corpus d'évaluation que nous avons annoté manuellement en attribuant à chaque pronom personnel son antécédent.

Le premier critère, selon Mitkov (2002), consiste à délimiter ce qui est à évaluer : l'algorithme de résolution seul ou le système dans son ensemble. Dans le cas de l'évaluation de l'algorithme seul, il est nécessaire d'évaluer le nombre d'anaphores qu'il résout correctement lorsque les informations d'entrée de l'algorithme ne contiennent aucune erreur étant préalablement révisée, autrement dit lorsqu'il se trouve dans des conditions idéales. Dans le cas de l'évaluation du système de résolution, c'est en étant placé dans les conditions normales d'utilisation que l'algorithme est évalué. Les outils d'annotation que comporte le système (segmenteur, analyseur syntaxique...) calculent automatiquement les informations d'entrée de l'algorithme sans avoir été révisé.

La puissance du système est le second critère relatif à l'évaluation. Quand ils sont considérés puissants, les systèmes suggèrent un antécédent pour toutes les anaphores à résoudre. La mesure du rappel d'un système puissant est donc analogue à la mesure de précision. L'antécédent ne peut être suggéré que lorsque la pertinence du candidat est certaine. Par ailleurs, lorsque l'ambiguïté est importante, d'autres systèmes ne suggèrent aucun antécédent. Pour les systèmes robustes, les mesures classiques de rappel et de précision sont inadaptées :

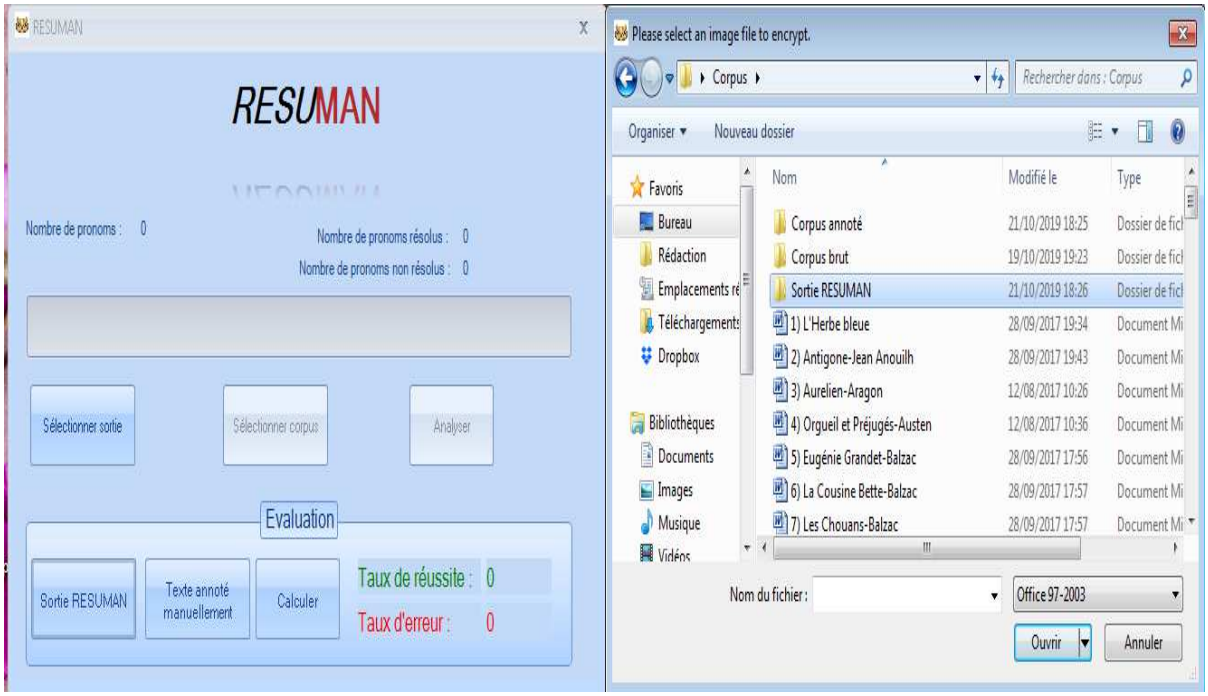
Nombre d'anaphores correctement résolues = Nombre d'anaphores identités par le système

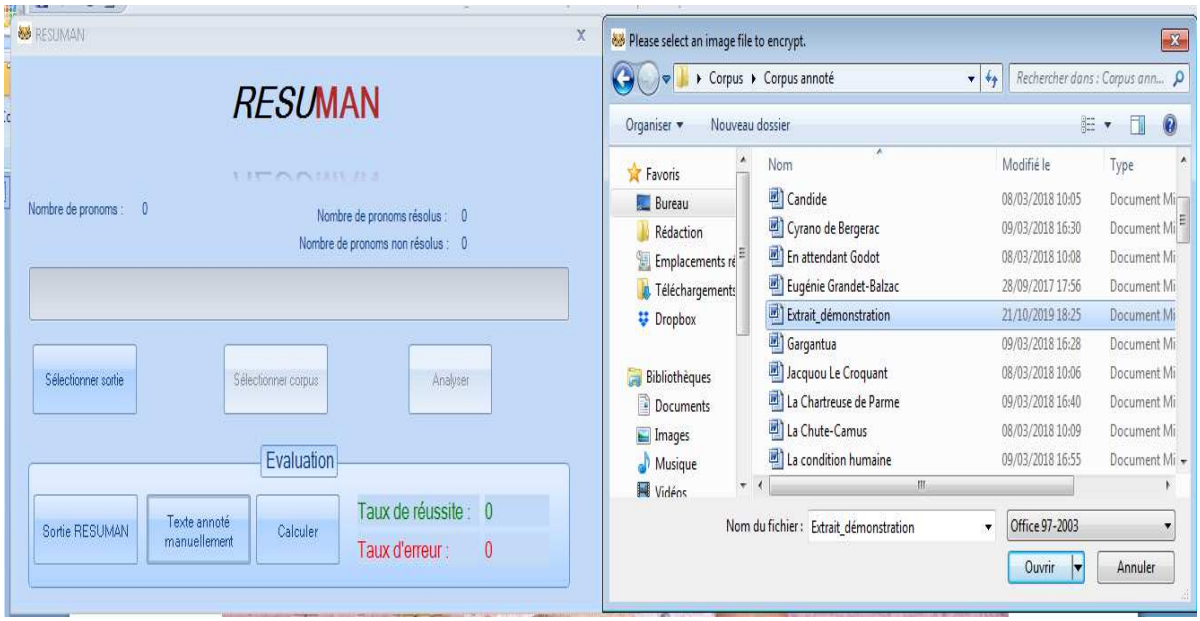
Afin d'évaluer un système, il faut comparer ses performances avec celles résultant d'autres systèmes sur un corpus de référence, tout en prenant en compte les informations d'entrée exploitées par les systèmes. Pour cela, ces informations doivent être identiques, or ce qui n'est pas toujours le cas car ce sont souvent des outils différents qui composent les systèmes. De plus, il faut préciser que les outils qui les composent influencent les performances générales des systèmes. Ainsi, la comparaison s'avère impossible lorsque les performances des outils varient d'un système à l'autre. Parfois, certains auteurs se limitent à quelques informations en omettant de toutes les corriger. Nous pouvons citer par exemple la segmentation des phrases et des mots uniquement, en laissant telles quelles d'autres informations.

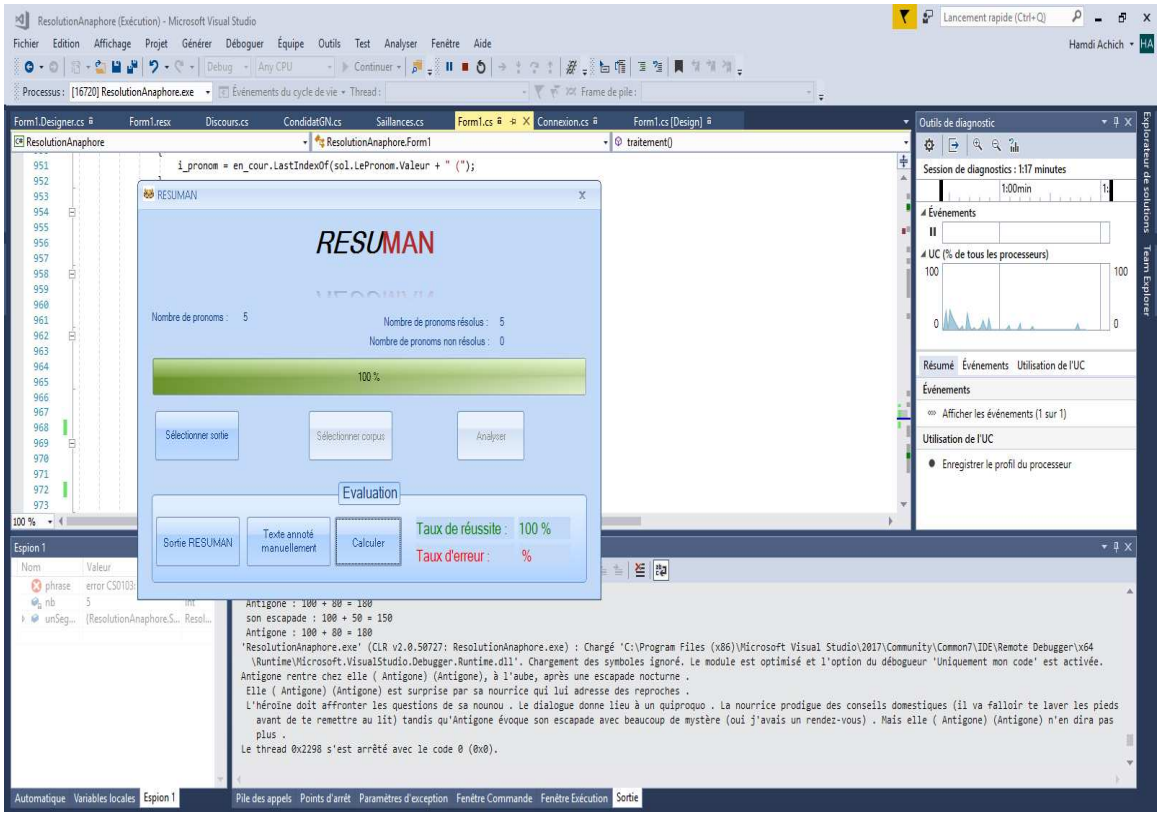
Afin de mesurer l'efficacité de $\text{RESUMAN}_{\text{O}}$ au moyen de l'exactitude, nous avons choisi de nous conformer à cette même stratégie dans notre travail. Les évaluations réalisées manuellement sont aujourd'hui considérées comme des références. Elles servent à

évaluer la qualité des outils produits pour la résolution des anaphores et pour évaluer leur performance. Aucune procédure générale pour évaluer un système de résolution d'anaphores n'existe jusqu'à présent. Par ailleurs, avant de choisir une procédure spécifique, plusieurs critères doivent être observés. En effet, pour évaluer $RESUMAN_O$, nous avons créé une fonction intégrée dans l'interface de la résolution : la rubrique Evaluation contient trois boutons comme suit : le premier (Sortie RESUMAN) à gauche sert à sélectionner le texte sortie de RESUMAN, le deuxième (Texte annoté manuellement) le même texte annoté manuellement et le dernier (Calculer) à calculer le taux de réussite et le taux d'erreur selon les fonctions d'évaluation citées dans notre travail.[209].

[209] A la section 1 du 3ème chapitre de la 2ème partie.


RESUMAN
RESUMAN
Nombre de pronoms : 0
Nombre de pronoms résolus : 0
Nombre de pronoms non résolus : 0
Sélectionner sortie
Sélectionner corpus
Analyser
Evaluation
Sortie RESUMAN
Texte annoté manuellement
Calculer
Taux de réussite : 0
Taux d'erreur : 0
Please select an image file to encrypt.
Corpus
Rechercher dans : Corpus
Organiser
Nouveau dossier
Favoris
Bureau
Rédaction
Emplacements ré
Téléchargements
Dropbox
Bibliothèques
Documents
Images
Musique
Vidéos
Nom
Modifié le
Type
Corpus annoté 21/10/2019 18:25 Dossier de fich
Corpus brut 19/10/2019 19:23 Dossier de fich
Sortie RESUMAN 21/10/2019 18:26 Dossier de fich
1) L'Herbe bleue 28/09/2017 19:34 Document Mi
2) Antigone-Jean Anouilh 28/09/2017 19:43 Document Mi
3) Aurelien-Aragon 12/08/2017 10:26 Document Mi
4) Orgueil et Préjugés-Austen 12/08/2017 10:36 Document Mi
5) Eugénie Grandet-Balzac 28/09/2017 17:56 Document Mi
6) La Cousine Bette-Balzac 28/09/2017 17:57 Document Mi
7) Les Chouans-Balzac 28/09/2017 17:57 Document Mi
Nom du fichier :
Office 97-2003
Ouvrir
Annuler


RESUMAN
RESUMAN
Nombre de pronoms : 0
Nombre de pronoms résolus : 0
Nombre de pronoms non résolus : 0
Sélectionner sortie
Sélectionner corpus
Analyser
Evaluation
Sortie RESUMAN
Texte annoté manuellement
Calculer
Taux de réussite : 0
Taux d'erreur : 0
Please select an image file to encrypt.
Corpus
Corpus annoté
Rechercher dans : Corpus ann...
Organiser
Nouveau dossier
Favoris
Bureau
Rédaction
Emplacements ré
Téléchargements
Dropbox
Bibliothèques
Documents
Images
Musique
Vidéos
Nom
Modifié le
Type
Candide 08/03/2018 10:05 Document Mi
Cyrano de Bergerac 09/03/2018 16:30 Document Mi
En attendant Godot 08/03/2018 10:08 Document Mi
Eugénie Grandet-Balzac 28/09/2017 17:56 Document Mi
Extrait_démonstration 21/10/2019 18:25 Document Mi
Gargantua 09/03/2018 16:28 Document Mi
Jacquou Le Croquant 08/03/2018 10:06 Document Mi
La Chartreuse de Parme 09/03/2018 16:40 Document Mi
La Chute-Camus 08/03/2018 10:09 Document Mi
La condition humaine 09/03/2018 16:55 Document Mi
Nom du fichier : Extrait_démonstration
Office 97-2003
Ouvrir
Annuler



Nous avons vérifié les cas d'erreur et avons constaté qu'il y a des GN non étiqueté par Fips (Annexe 2). Cela engendre leur élimination de la liste des candidats potentiels et en résulte soit l'attribution d'un référent erroné soit la non résolution du pronom détecté en laissant la place de l'antécédent vide (Aucun) (Annexe 3). Néanmoins, nous avons apporté manuellement quelques corrections à l'annotation par Fips et nous avons remarqué une nette amélioration au niveau de la résolution et le taux de précision (Annexe 4). Nous ne nions pas que notre choix d'une approche linéaire (la sortie d'un module est l'entrée du suivant) nous a nécessairement menée à récolter des erreurs issues de chaque étape du processus. En revanche, et malgré toutes les contraintes citées, les taux de réussites actuels sont classés entre 65,15% et 85% : des valeurs encourageantes pour de nouvelles améliorations et perspectives.

Conclusion

Depuis quelques décennies, la recherche linguistique ne peut ignorer les avancées du TAL et toutes deux évoluent ensemble. La capacité de traitement et de résolution des problèmes d'anaphores est un aspect très important pour toutes les applications de TAL telle que l'extraction d'information, le résumé automatique, la traduction automatique ou de manière générale la compréhension automatique. Elle consiste à trouver, dans un texte et pour chaque anaphore, son antécédent. Afin d'implémenter un outil pour résoudre automatiquement les anaphores pronominales dans un corpus textuel créé dans le cadre de ce travail, nous avons développé un algorithme à base de connaissances linguistiques. Les résultats obtenus sont encourageants puisqu'ils atteignent parfois ceux des meilleurs résultats actuellement obtenus avec des moyens dont nous ne disposons pas.

La question principale de notre étude concernait l'anaphore et le fonctionnement des chaînes référentielles. Un des principaux objectifs de notre recherche consistait à décrire et expliquer le comportement ambigu des anaphores pronominales dans des textes français. Apothéloz (1995 : 313) a ainsi remarqué qu' :

> On trouve dans la littérature sur l'anaphore un certain nombre d'hypothèses et d'affirmations sur les conditions d'emploi des diverses formes d'anaphoriques (…) [qui] sont la plupart du temps fondées sur quelques exemples isolés de leur contexte et souvent construits pour les besoins de la démonstration : elles sont donc des plus fragiles.

Nous nous distinguons de cette littérature par le fait que nous avons proposé de traiter l'anaphore dans un contexte textuel non isolé, comme un procédé actif de gestion cognitive de la dynamique textuelle et sur un corpus fermé et homogène : en effet, de notre côté, il ne s'agit pas d'un travail réalisé dans une optique représentative mais plutôt d'une étude d'une haute densité des phénomènes analysés menée dans une perspective informatique, partant d'une analyse sur corpus. Le corpus ainsi constitué, RESUMAN, vise à interroger le fonctionnement de l'anaphore pronominale ambiguë dans les textes retenus en vue de mettre en évidence des caractéristiques syntaxico-cognitives propres aux chaînes anaphoriques. Notre objectif n'étant pas de classer les antécédents potentiels, nous avons cherché s'il y a des préférences entre eux, qui véhiculent des dimensions linguistiques spécifiques dans des textes pouvant être qualifiés de brefs. Ce faisant, nous nous sommes interrogée sur les mécanismes exploités dans la levée des ambiguïtés par le sujet humain peut offrir des indications les plus pertinentes possibles quant au type de techniques informatiques qui pourraient faciliter l'intégration de ce phénomène dans le processus de compréhension automatique. En effet, nous prenons ici en compte les effets cognitifs et

textuels de l'emploi de *il*. Ce sont ces questions qui ont suscité le parcours de recherche. Les résultats auxquels nous sommes parvenus sont de plusieurs ordres.

En nous appuyant sur les travaux existant (malgré que la description reste parfois lacunaire), nous avons d'une part mis en évidence les points de vue existants, et d'autre part établi des critères pour identifier le bon antécédent d'une anaphore pronominale. Dans un premier temps, un éclairage du terme anaphore s'est avéré nécessaire. Ont été abordées ensuite l'anaphore coréférentielle et celle non-coréférentielle et enfin, nous nous sommes arrêtée sur l'évolution de la notion de la référence par l'intermédiaire du changement de ses perspectives notamment avec le tournant pragmatique et l'intégration de la cognition dans la réflexion linguistique. Nous avons montré, à travers un choix de références représentatives, comment la thématique de l'anaphore peut être incluse dans la thématique de la référence, les deux étant indissociables. Cette idée apparaît dans la conception de l'anaphore même qui est en relation de dépendance avec la référence. Nous avons ainsi retenu la position de Milner concernant la thématique de la référence contemporaine sans éluder quelques problèmes que cette théorie peut poser : la saturation référentielle n'est pas un simple problème de distribution automatique, pas plus un problème de stratégie de l'interprétation cognitive ; le modèle du code et le modèle inférentiel interviennent chacun pour sa part et dans des moments différents dans la stratégie consistant à trouver le bon référent. La plupart des expressions référentielles ne sont pas saturées sémantiquement et référentiellement : le sémantisme de l'expression référentielle, le cotexte linguistique, les informations extralinguistiques et la situation de communication sont des facteurs que l'interlocuteur utilise pour déterminer une classe de référents possibles et un référent à l'intérieur de cette classe.

Notre étude parallèle de la littérature a souligné l'importance de la mise en place d'une approche pluridimensionnelle. Les approches traitant la notion de l'anaphore se basent sur des modalités complémentaires dans le but d'expliquer le phénomène. L'essence de notre approche réside dans la prise en considération de facteurs cognitifs qui ont amené certains linguistes à rendre compte des anaphoriques. En effet, l'introduction de ces derniers permet de mieux caractériser l'anaphore, puisque l'approche syntaxique, à elle seule, entrave l'étude complète du phénomène anaphorique. La structure informationnelle du modèle contextuel est balisée de manière prégnante par un fait discursif réel qui est l'introduction explicite d'une entité par le texte. En revanche, une introduction implicite non textuelle n'a pas de poids contraignant. Elle ne reçoit en effet sa légitimé de saillance

que plus tard, étant donné que ce n'est qu'un fait potentiel. Par ailleurs, nous avons montré que les trois approches sont complémentaires et notre étude de l'approche textuelle n'enlève rien de la pertinence de l'approche mémorielle et de l'approche substitutive. Dans l'approche cognitiviste, la résolution référentielle des pronoms est conçue comme un processus consistant à les rattacher non pas à des constituants linguistiques présents dans le co-texte, mais à des entités de nature cognitive construites par l'interprétant au cours du traitement du discours. Dans cette approche, le réfèrent des formes linguistiques est toujours un élément cognitif appartenant à la conscience du lecteur.

Afin de dépasser les relations superficielles entre les énoncés, la recherche de la cohérence implique l'utilisation de procédures inférentielles chez l'interprétant. De la même façon, un travail de résolution est souvent nécessaire pour les marques de relations impliquées dans la cohésion. En ce sens, nous avons remarqué qu'il n'y a pas autant de différences entre cohérence et cohésion que la littérature présentée le laissait supposer. Pour la cohésion, des indications explicites sont fournies par le locuteur sur la mise en relation des énoncés par l'interprétant, alors que l'établissement de la cohérence implique, en plus, la constitution de relations non explicitement marquées dans le texte. Nous préférons dire que les deux notions sont complémentaires plutôt que distinctes. Ces unités linguistiques, comme l'anaphore, qui sont organisées de manières précises, établissent des unités spécifiques qui permettent aux locuteurs de dépasser le cadre d'une phrase et pouvoir comprendre un texte.

Les recherches consacrées à l'étude de l'acquisition des possibilités de traitement du système anaphorique ont le plus souvent utilisé des anaphores de type pronominal, en particulier des pronoms personnels. Parmi les procédures les plus souvent citées, nous avons retenu les procédures *d'analyse morphologique,* la procédure *des fonctions syntaxiques parallèles,* la procédure *de distance minimale* et la procédure *du sujet* pour concevoir l'algorithme de RESUMAN. Nous avons essayé d'analyser les différentes situations impliquant un antécédent flou : cela nous a effectivement aidée dans la tâche de la résolution puisque celle-ci passe par l'identification des expressions anaphoriques, celle de leurs antécédents potentiels, et l'attribution de relations entre les entités des segments textuels retenus.

Dans ce travail, nous nous sommes intéressée aux ajustements propres à l'anaphore ainsi qu'aux liens anaphoriques unissant les syntagmes d'une même chaîne. Nous avons analysé un corpus simple, utilisant des connaissances linguistiques simples, qui partitionne

les expressions référentielles d'un texte en chaînes de coréférences distinctes. Dans cette étude, nous avons constaté que même avec des ressources linguistiques limitées, il est possible de construire un algorithme simple qui permet d'établir les paires pronom/antécédent dans de courts textes. Pour ce faire, nous avons utilisé en grande partie les éléments linguistiques présents à la surface du texte. Nous retiendrons donc qu'une analyse linguistique est un mécanisme plus ou moins complexe suivant les besoins, basée généralement sur un processus à trois étapes nécessitant des connaissances préalables du domaine étudié, ces connaissances étant d'ordre lexical, syntaxique et sémantique. Compte tenu des meilleures performances de l'approche de Mitkov et surtout de la non-nécessité des analyses syntaxique et sémantique complexes exigeant des outils qui ne sont pas toujours disponibles, nous avons trouvé intéressant de nous inspirer de cette approche. Cependant, un certain nombre de choix faits dans cette approche n'ont pas toujours été suffisamment justifiés à nos yeux. Aussi un certain nombre d'autres questions relatives à ces choix méritent-elles d'être posées. Notamment : sur quelle base s'est opéré le choix des scores ? Pourquoi le choix de s'arrêter aux deux phrases précédentes pour chercher l'antécédent d'une anaphore ?

L'un des points intéressants du processus de compréhension automatique de texte réside dans le fait que lors de la construction de la représentation sémantique, on ne dispose d'aucune information sur les types de questions pouvant être formulées sur un document. En fin de compte, on peut avancer que la compréhension automatique de textes devrait être capable de donner une description du contenu d'un document d'une manière la plus complète possible afin de pouvoir fournir le maximum de renseignements possibles.

Pour cela, il est nécessaire de rappeler que les dispositifs anaphoriques dépassent les limites de la proposition et de la phrase puisque antécédent et anaphorique peuvent appartenir à des phrases différentes. En résumé, la façon dont le processus de sélection du référent des anaphoriques est envisagé permet de cerner les différentes sources potentielles d'erreurs d'interprétation. Le modèle minimaliste semble plus adapté que les modèles antérieurs pour décrire le processus de traitement des pronoms. L'hypothèse d'un processus automatique permet de comprendre en quoi l'utilisation d'un pronom est plus économique en ressources cognitives qu'une simple reprise nominale. Elle autorise aussi l'identification d'une grande variété d'erreurs d'interprétation potentielles.

L'approche que nous avons proposée se veut encore moins exigeante en ressources, mais pour autant nous avons tenu au moins à justifier le choix de la taille de l'espace de recherche des antécédents des anaphores. Cette approche entre dans le cadre des méthodes

de résolution d'anaphores exploitant des connaissances linguistiques. Durant cette étape, nous avons été amenée à proposer des heuristiques pour inférer certaines informations syntaxiques et cognitives (saillance) nécessaires pour l'approche. Ceci nous a permis de prendre conscience du fait qu'avec peu de ressources, nous pouvons arriver à des résultats acceptables. L'un des aspects sur lesquels nous avons insisté au cours de ce travail a été de consacrer une grande attention à la formalisation. En effet, nous pensons que cette façon de procéder contribue à améliorer facilement notre approche. Aussi, nous avons mis l'accent sur la modularité. Les caractéristiques principales que nous avons retenues sont une prise en compte des pronoms impersonnels, un calcul de la saillance basé sur d'autres critères syntaxiques que la seule linéarité, le choix sur des critères de proximité en cas d'ambiguïté sur la saillance et la préférence des anaphores intra-phrastiques sur les anaphores interphrastiques.

La prise en compte des relations entre les différentes entités du texte et du problème de résolution des anaphores est motivée par le fait que la sémantique d'un texte se construit petit à petit. Il s'agit du principe de compositionnalité qui stipule que le sens d'une expression composée dépend essentiellement du sens de ses composants et des règles syntaxiques par lesquelles ils sont combinés. Tenir compte du contexte où se situe le texte implique la disponibilité d'un certain nombre de connaissances générales et spécifiques sur les conditions d'énonciation des faits évoqués dans ce texte. La compréhension va ainsi de pair avec l'acquisition des connaissances et l'exploitation de ces connaissances.

Un certain nombre de limites sont à noter pour notre approche. Elles sont dues aux choix que nous avons effectués, ce qui implique un certain nombre de contraintes.

Tout d'abord, Fips, l'analyseur morphosyntaxique utilisé pour le prétraitement de notre corpus, n'annote pas les noms propres, ce qui nous a conduite à accomplir cette tâche manuellement. Nous avons annoté aussi les deux tiers de notre corpus manuellement. Bien que ces tâches ne soient pas visibles dans notre outil, elles se sont révélées chronophages. Nous avons constaté que les erreurs de résolution viennent en partie des erreurs d'étiquetage de Fips. Une amélioration de Fips conduira certainement à une nette amélioration de $\text{RESUMAN}_{\text{O}}$. Ensuite, l'analyse permet d'émettre des hypothèses sur certaines sources potentielles de difficultés d'interprétation. L'absence de formes de pronoms personnels spécifiques à chacune des fonctions qu'ils doivent assurer peut être

une des causes provoquant des erreurs d'attribution de la référence. L'interprétation des pronoms personnels semble pouvoir être affectée par leur caractère "coréférentiel".

Le but ultime de ce travail n'était pas de fournir un système « clé en main » pour résoudre les anaphores pronominales dans des textes français, mais d'apporter notre contribution à l'élaboration d'un certain nombre d'outils permettant d'atteindre cet objectif. Notre objectif était aussi de participer à enrichir les ressources de corpus de référence en français : nous avons vu au chapitre dédié au corpus que le nombre des corpus français annotés en anaphore/référence est très réduite par apport aux autres langues. Nous voulons rajouter un corpus annoté automatiquement en anaphore pronominale et qui sera libre d'accès. Nous pourrions (enfin) dire que notre outil permettrait de réduire le temps d'annotation manuelle des textes français et le prétraitement des relations référentielles.

La réalisation d'un système intégrant une approche sémantique est sans aucun doute un prolongement intéressant de ce travail. Nous avons abordé cette thématique au troisième chapitre de la deuxième partie de notre travail. Nous voulions à ce stade, intégrer un module sémantique à notre outil. Néanmoins, nous nous sommes rendue compte que cette tentative dépassait l'étendue de notre thèse. Les interprétations que nous avons pu retenir à la fin dudit chapitre nous invitent vivement à aller en ce sens comme perspectives.

Pour ce qui est des règles d'identification des rôles sémantiques, on pourra imaginer une façon de présenter ces règles en les stockant par exemple dans une base de données accessible à partir de la construction d'un dictionnaire sémantique. L'idée de l'approche pourra se décliner ainsi : proposer une méthode de représentation sémantique des phrases d'un texte. Ainsi, en se basant sur l'hypothèse que la sémantique d'un terme s'obtient par une mise en relation entre le matériau linguistique et un ensemble de connaissance sur ce matériau, nous pourrions nous affronter au problème d'acquisition des connaissances. En se basant sur une autre hypothèse, la composionnalité qui veut que la construction de la sémantique d'une expression soit la combinaison des sémantiques des termes qui la composent, nous proposerions une méthode de résolution des liens entre les différentes entités d'une chaîne anaphorique. Le point de départ pourrait être la grammaire de Fillmore qui se définit par un certain nombre de rôles sémantiques. Le fonctionnement de ces règles est fondé principalement sur l'identification du verbe principal de la phrase et des fonctions syntaxiques des différents termes constituant cette phrase y compris les termes anaphoriques. C'est ainsi que nous proposerions la construction d'un dictionnaire

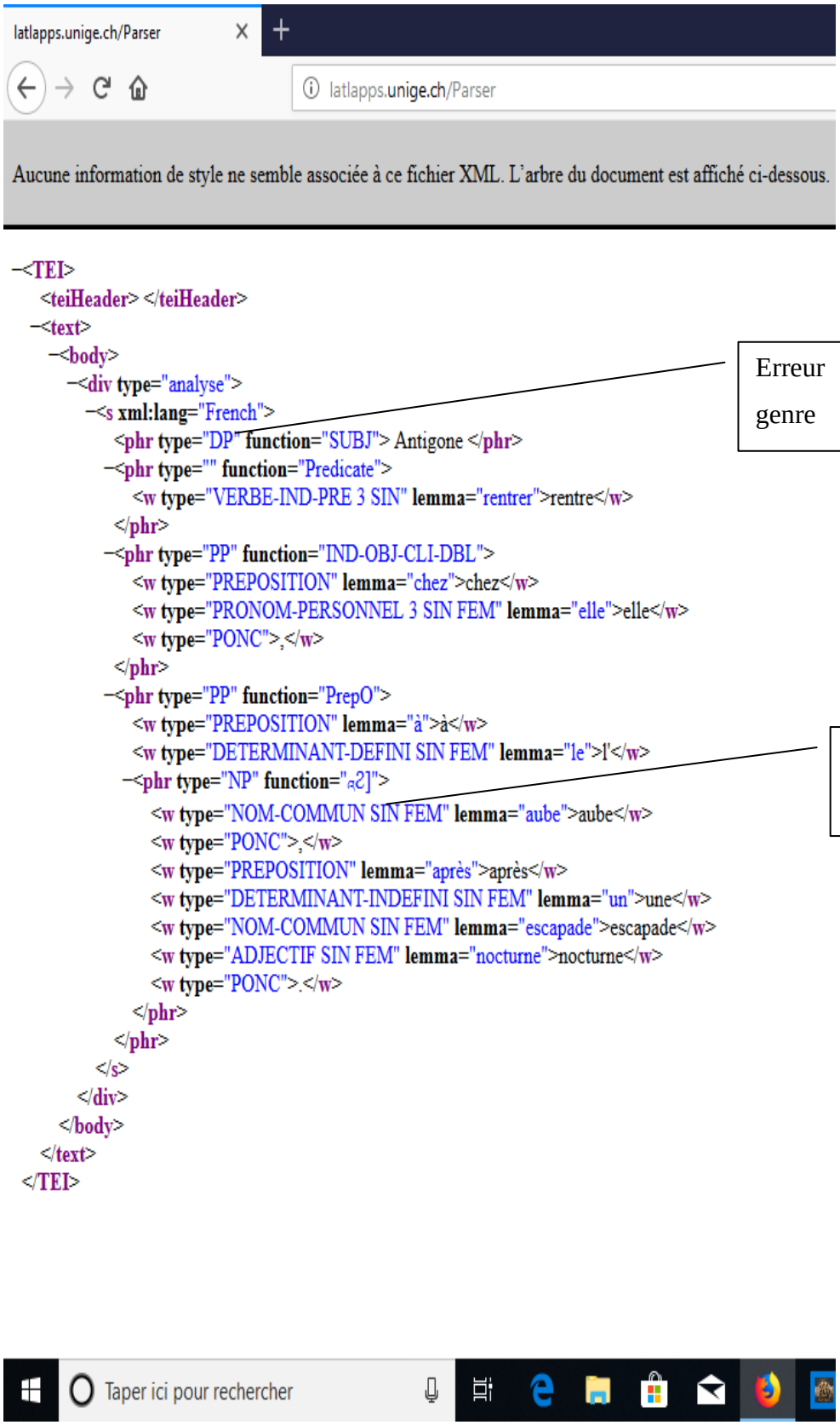
sémantique à partir du corpus RESUMAN. En effet, nous pensons que cette façon de procéder contribuerait facilement à améliorer notre outil.

Annexe

Annexe 1 : Corpus RESUMAN (sur CD[210])

Annexe 2 : Erreurs d'étiquettes produites par Fips

[210] Vu que notre corpus est volumineux, nous le joignons à ce manuscrit sur CD.


latlapps.unige.ch/Parser
latlapps.unige.ch/Parser
Aucune information de style ne semble associée à ce fichier XML. L'arbre du document est affiché ci-dessous.
−<TEI>
<teiHeader> </teiHeader>
−<text>
−<body>
−
−<s xml:lang="French">
<phr type="DP" function="SUBJ"> Antigone </phr>
−<phr type="" function="Predicate">
<w type="VERBE-IND-PRE 3 SIN" lemma="rentrer">rentre</w>
</phr>
−<phr type="PP" function="IND-OBJ-CLI-DBL">
<w type="PREPOSITION" lemma="chez">chez</w>
<w type="PRONOM-PERSONNEL 3 SIN FEM" lemma="elle">elle</w>
<w type="PONC">,</w>
</phr>
−<phr type="PP" function="PrepO">
<w type="PREPOSITION" lemma="à">à</w>
<w type="DETERMINANT-DEFINI SIN FEM" lemma="le">l'</w>
−<phr type="NP" function="[illegible]2]">
<w type="NOM-COMMUN SIN FEM" lemma="aube">aube</w>
<w type="PONC">,</w>
<w type="PREPOSITION" lemma="après">après</w>
<w type="DETERMINANT-INDEFINI SIN FEM" lemma="un">une</w>
<w type="NOM-COMMUN SIN FEM" lemma="escapade">escapade</w>
<w type="ADJECTIF SIN FEM" lemma="nocturne">nocturne</w>
<w type="PONC">.</w>
</phr>
</phr>
</s>

</body>
</text>
</TEI>
Erreur de genre
Erreur de fonction
Taper ici pour rechercher

Annexe 3 : Résumé Antigone

a) Résumé brut[211] d'Antigone : http://www.alalettre.com/anouilh-oeuvres-antigone.php

<Œuvre>Antigone
<Auteur> Jean Anouilh
<Genre> Théâtre
<Résumé> Le personnage baptisé le Prologue présente les différents protagonistes et résume la légende de Thèbes (Anouilh reprend cette tradition grecque qui consiste à confier à un personnage particulier un monologue permettant aux spectateurs de se rafraîchir la mémoire. Le Prologue replace la pièce dans son contexte mythique). Toute la troupe des comédiens est en scène. Si certains personnages semblent ignorer le drame qui se noue, d'autres songent déjà au désastre annoncé.
Antigone rentre chez elle , à l'aube, après une escapade nocturne. Elle est surprise par sa nourrice qui lui adresse des reproches. L'héroïne doit affronter les questions de sa nounou. Le dialogue donne lieu à un quiproquo . La nourrice prodigue des conseils domestiques (" il va falloir te laver les pieds avant de te remettre au lit") tandis qu'Antigone évoque son escapade avec beaucoup de mystère (" oui j'avais un rendez-vous") . Mais elle n'en dira pas plus.
La nourrice sort et Ismène, la sœur d'Antigone, dissuade cette dernière d'enfreindre l'ordre de Créon et d'ensevelir le corps de Polynice. Ismène exhorte sa sœur à la prudence ("Il est plus fort que nous, Antigone, il est le roi") . Antigone refuse ces conseils de sagesse . Elle n'entend pas devenir raisonnable.
Antigone se retrouve à nouveau seule avec sa nourrice. Elle cherche à surmonter ses doutes et demande à sa nourrice de la rassurer. Elle tient aussi des propos ambigus pour ceux (et c'est le cas de la nourrice) qui ne connaissent pas son dessein . Elle semble décidée à mourir et évoque sa disparition à mots couverts " Si, moi , pour une raison ou pour une autre, je ne pouvais plus lui parler...".
Antigone souhaite également s'expliquer avec son fiancé Hémon. Elle lui demande de le pardonner pour leur dispute de la veille. Les deux amoureux rêvent alors d'un bonheur improbable. Sûre d'être aimée , Antigone est rassurée. Elle demande cependant à Hémon de garder le silence et lui annonce qu'elle ne pourra jamais l'épouser. Là encore , la scène prête au quiproquo : le spectateur comprend qu'Antigone pense à sa mort prochaine, tandis qu'Hémon , qui lui n'a pas percé le dessein d'Antigone, est attristé de ce qu'il prend pour un refus.
Ismène revient en scène et conjure sa sœur de renoncer à son projet. Elle affirme même que Polynice, le "frère banni", n'aimait pas cette sœur qui aujourd'hui est prête à se sacrifier pour lui.
Antigone avoue alors avec un sentiment de triomphe, qu'il est trop tard, car elle a déjà , dans la nuit, bravé l'ordre de Créon et accompli son geste " C'est trop tard. Ce matin , quand tu m'as rencontrée , j'en venais."
Jonas, un des gardes chargés de surveiller le corps de Polynice, vient révéler à Créon, qu'on a transgressé ses ordres et recouvert le corps de terre. Le roi veut croire à un complot dirigé contre lui et fait prendre des mesures pour renforcer la surveillance du corps de

[211] Nous voulons dire par brut le fait que nous utilisons le résumé tel qu'il est en ligne. Nous ne le modifions pas ni le corrigeons.

Polynice. Il semble également vouloir garder le secret sur cet incident : " Va vite. Si personne ne sait, tu vivras."
Le chœur s'adresse directement au public et vient clore la première partie de la pièce. Il commente les événements en exposant sa conception de la tragédie qu'il oppose au genre littéraire du drame. Le chœur affiche également une certaine ironie et dévoile les recettes de l'auteur : "c'est cela qui est commode dans la tragédie. On donne un petit coup de pouce pour que cela démarre... C'est tout. Après on n'a plus qu'à laisser faire. On est tranquille. Cela roule tout seul."
Antigone est traînée sur scène par les gardes qui l'ont trouvée près du cadavre de son frère. Ils ne veulent pas croire qu'elle est la nièce du roi , et la traitent avec brutalité. Ils se réjouissent de cette capture et des récompenses et distinctions qu'elle leur vaudra.
Créon les rejoint. Les gardes font leur rapport . Le roi ne veut pas les croire. Il interroge sa nièce qui avoue aussitôt. Il fait alors mettre les gardes au secret, avant que le scandale ne s'ébruite.
Créon et Antigone restent seuls sur scène. C'est la grande confrontation entre le roi et Antigone. Le roi souhaite étouffer le scandale et ramener la jeune fille à la raison. Dans un premier temps , Antigone affronte Créon qui tente de la dominer de son autorité.
Les deux protagonistes dévoilent leur personnalité et leurs motivations inconciliables. Créon justifie les obligations liées à son rôle d'homme d'état . Antigone semble sourde à ses arguments : (Créon : Est ce que tu le comprends cela ? Antigone : " Je ne veux pas le comprendre.") . A court d'arguments Créon révèle les véritables visages de Polynice et d'Etéocle et les raisons de leur ignoble conflit. Cet éclairage révolte Antigone qui semble prête à renoncer et à se soumettre. Mais c'est en lui promettant un bonheur ordinaire avec Hémon, que Créon ravive son amour-propre et provoque chez elle un ultime sursaut. Elle rejette ce futur inodore et se rebelle à nouveau. Elle choisit une nouvelle fois la révolte et la mort.
Ismène , la sœur d'Antigone entre en scène alors que cette dernière s'apprêtait à sortir et à commettre un esclandre , ce qui aurait obligé le roi à l'emprisonner. Ismène se range aux côtés d'Antigone et est prête à mettre elle aussi sa vie en jeu. Mais Antigone refuse , prétextant qu'il est trop facile de jouer les héroïnes maintenant que les dés ont été jetés. Créon appelle la garde , Antigone clôt la scène en appelant la mort de ses cris et en avouant son soulagement (Enfin Créon !)
Le chœur entre en scène. Les personnages semblent avoir perdu la raison, ils se bousculent. Le chœur essaye d'intercéder en faveur d'Antigone et tente de convaincre Créon d'empêcher la condamnation à mort d'Antigone. Mais le roi refuse , prétextant qu'Antigone a choisi elle-même son destin, et qu'il ne peut la forcer à vivre malgré elle.
Hémon vient lui aussi, ivre de douleur, supplier son père d'épargner Antigone, puis il s'enfuit.
Antigone reste seule avec un garde. Elle rencontre là le "dernier visage d'homme". Il se révèle bien mesquin, et ne sait parler que de grade et de promotion. Il est incapable d'offrir le moindre réconfort à Antigone. Cette scène contraste, par son calme, avec le violent tumulte des scènes précédentes. Apprenant qu'elle va être enterrée vivante, éprouvant de profonds doutes (" Et Créon avait raison, c'est terrible maintenant, à côté de cet homme, je ne sais plus pourquoi je meurs." , Antigone souhaite dicter au garde une lettre pour Hémon dans laquelle elle exprime ses dernières pensées. Puis elle se reprend et corrige ce dernier message ("Il vaut mieux que jamais personne ne sache"). C'est la dernière apparition d'Antigone.
Le messager entre en scène et annonce à Créon et au public la mort d'Antigone et la mort de son fils Hémon. Tous les efforts de Créon pour le sauver ont été vains. C'est alors le chœur qui annonce le suicide d'Eurydice, la femme de Créon : elle n'a pas supporté la mort

de ce fils qu'elle aimait tant. Créon garde un calme étonnant . Il indique son désir de poursuivre " la salle besogne " sans faillir. Il sort en compagnie de son page.
Tous les personnages sont sortis. Le chœur entre en scène et s'adresse au public : Il constate avec une certaine ironie la mort de nombreux personnages de cette tragédie : "Morts pareils, tous, bien raides, bien inutiles, bien pourris." La mort a triomphé de presque tous . Il ne reste plus que Créon dans son palais vide . Les gardes , eux continuent de jouer aux cartes , comme ils l'avaient fait lors du Prologue. Ils semblent les seuls épargnés par la tragédie. Ultime dérision.

b) Sortie RESUMAN sans correction manuelle de l'étiquetage par Fips

<Œuvre>Antigone .<Auteur> Jean Anouilh .<Genre> Théâtre .<Résumé> Le personnage baptisé le Prologue présente les différents protagonistes et résume la légende de Thèbes (Anouilh reprend cette tradition grecque qui consiste à confier à un personnage particulier un monologue permettant aux spectateurs de se rafraîchir la mémoire . Le Prologue replace la pièce dans son contexte mythique) . Toute la troupe des comédiens est en scène . Si certains personnages semblent ignorer le drame qui se noue, d'autres songent déjà au désastre annoncé .Antigone rentre chez elle (Antigone) , à l'aube, après une escapade nocturne .
Elle (Antigone) est surprise par sa nourrice qui lui (Antigone) adresse des reproches . L'héroïne doit affronter les questions de sa nounou . Le dialogue donne lieu à un quiproquo . La nourrice prodigue des conseils domestiques (il va falloir te laver les pieds avant de te remettre au lit) tandis qu'Antigone évoque son escapade avec beaucoup de mystère (oui j'avais un rendez-vous) . Mais elle (Antigone) n'en dira pas plus .
La nourrice sort et Ismène, la sœur d'Antigone, dissuade cette dernière d'enfreindre l'ordre de Créon et d'ensevelir le corps de Polynice . Ismène exhorte sa sœur à la prudence (Il est plus fort que nous, Antigone, il (le roi) est le roi) . Antigone refuse ces conseils de sagesse . Elle (Antigone) n'entend pas devenir raisonnable .
Antigone se retrouve à nouveau seule avec sa nourrice . Elle (Antigone) cherche à surmonter ses doutes et demande à sa nourrice de la rassurer .
Elle (Antigone) tient aussi des propos ambigus pour ceux (et c'est le cas de la nourrice) qui ne connaissent pas son dessein . Elle (Aucun) semble décidée à mourir et évoque sa disparition à mots couverts Si, moi, pour une raison ou pour une autre, je ne pouvais plus lui parler .
Antigone souhaite également s'expliquer avec son fiancé Hémon . Elle (Antigone) lui demande de la pardonner pour leur dispute de la veille .
Les deux amoureux rêvent alors d'un bonheur improbable . Sûre d'être aimée, Antigone est rassurée . Elle (Antigone) demande cependant à Hémon de garder le silence et lui annonce qu'elle (Antigone) ne pourra jamais l'épouser .
Là encore, la scène prête au quiproquo . le spectateur comprend qu'Antigone pense à sa mort prochaine, tandis qu'Hémon, qui lui n'a pas percé le dessein d'Antigone, est attristé de ce qu'il (le spectateur) prend pour un refus .
Ismène revient en scène et conjure sa sœur de renoncer à son projet . Elle affirme même que Polynice, le frère banni, n'aimait pas cette sœur qui aujourd'hui est prête à se sacrifier pour lui (le spectateur) .
Antigone avoue alors avec un sentiment de triomphe, qu'il (le spectateur) est trop tard, car elle a déjà, dans la nuit, bravé l'ordre de Créon et accompli son geste C'est trop tard .

e matin, quand tu m'as rencontrée, j'en venais .Jonas, un des gardes chargés de surveiller le corps de Polynice, vient révéler à Créon, qu'on a transgressé ses ordres et recouvert le corps de terre . Le roi veut croire à un complot dirigé contre lui (Le roi) et fait prendre des mesures pour renforcer la surveillance du corps de Polynice .

Il (Le roi) semble également vouloir garder le secret sur cet incident .

Va vite . Si personne ne sait, tu vivras .Le chœur s'adresse directement au public et vient clore la première partie de la pièce . Il (Le chœur) commente les événements en exposant sa conception de la tragédie qu'il (Le roi) oppose au genre littéraire du drame .

Le chœur affiche également une certaine ironie et dévoile les recettes de l'auteur . c'est cela qui est commode dans la tragédie . On donne un petit coup de pouce pour que cela démarre . C'est tout . Après on n'a plus qu'à laisser faire . On est tranquille . Cela roule tout seul .Antigone est traînée sur scène par les gardes qui l'ont trouvée près du cadavre de son frère . Ils (les gardes) ne veulent pas croire qu'elle (Antigone) est la nièce du roi, et la traitent avec brutalité . Ils (les gardes) se réjouissent de cette capture et des récompenses et distinctions qu'elle (Antigone) leur vaudra .

Créon les rejoint . Les gardes font leur rapport . Le roi ne veut pas les croire . Il (Le roi) interroge sa nièce qui avoue aussitôt .

Il (le roi) fait alors mettre les gardes au secret, avant que le scandale ne s'ébruite .

Créon et Antigone restent seuls sur scène . C'est la grande confrontation entre le roi et Antigone . Le roi souhaite étouffer le scandale et ramener la jeune fille à la raison . Dans un premier temps, Antigone affronte Créon qui tente de la dominer de son autorité .Les deux protagonistes dévoilent leur personnalité et leurs motivations inconciliables . Créon justifie les obligations liées à son rôle d'homme d'état . Antigone semble sourde à ses arguments . (Créon . Est ce que tu le comprends cela . Antigone . Je ne veux pas le comprendre .) . A court d'arguments Créon révèle les véritables visages de Polynice et d'Etéocle et les raisons de leur ignoble conflit . Cet éclairage révolte Antigone qui semble prête à renoncer et à se soumettre . Mais c'est en lui promettant un bonheur ordinaire avec Hémon, que Créon ravive son amour-propre et provoque chez elle (Antigone) un ultime sursaut . Elle rejette ce futur inodore et se rebelle à nouveau . Elle (Antigone) (Antigone) choisit une nouvelle fois la révolte et la mort .

Ismène, la sœur d'Antigone entre en scène alors que cette dernière s'apprêtait à sortir et à commettre un esclandre, ce qui aurait obligé le roi à l'emprisonner . Ismène se range aux côtés d'Antigone et est prête à mettre elle (Ismène) aussi sa vie en jeu . Mais Antigone refuse, prétextant qu'il est trop facile de jouer les héroïnes maintenant que les dés ont été jetés . Créon appelle la garde, Antigone clôt la scène en appelant la mort de ses cris et en avouant son soulagement (Enfin Créon .) .Le chœur entre en scène . Les personnages semblent avoir perdu la raison, ils (les personnages) se bousculent .

Le chœur essaye d'intercéder en faveur d'Antigone et tente de convaincre Créon d'empêcher la condamnation à mort d'Antigone . Mais le roi refuse, prétextant qu'Antigone a choisi elle-même (Antigone) son destin, et qu'il (le roi) ne peut la forcer à vivre malgré elle (Antigone) .Hémon vient lui aussi, ivre de douleur, supplier son père d'épargner Antigone, puis il (Hémon) s'enfuit .

Antigone reste seule avec un garde . Elle (Antigone) rencontre là le dernier visage d'homme . Il (un garde) se révèle bien mesquin, et ne sait parler que de grade et de promotion .

Il (un garde) est incapable d'offrir le moindre réconfort à Antigone . Cette scène contraste, par son calme, avec le violent tumulte des scènes précédentes . Apprenant qu'elle (Antigone) va être enterrée vivante, éprouvant de profonds doutes (Et Créon avait raison, c'est terrible maintenant, à côté de cet homme, je ne sais plus pourquoi je meurs .

Antigone souhaite dicter au garde une lettre pour Hémon dans laquelle elle (Antigone) exprime ses dernières pensées .

Puis elle (Antigone) se reprend et corrige ce dernier message (Il vaut mieux que jamais personne ne sache) . C'est la dernière apparition d'Antigone .Le messager entre en scène et annonce à Créon et au public la mort d'Antigone et la mort de son fils Hémon . Tous les efforts de Créon pour le sauver ont été vains . C'est alors le chœur qui annonce le suicide d'Eurydice, la femme de Créon . elle (Eurydice) n'a pas supporté la mort de ce fils qu'elle (Eurydice) aimait tant .

Créon garde un calme étonnant . Il (Créon) indique son désir de poursuivre la salle besogne sans faillir . Il (Créon) sort en compagnie de son page .

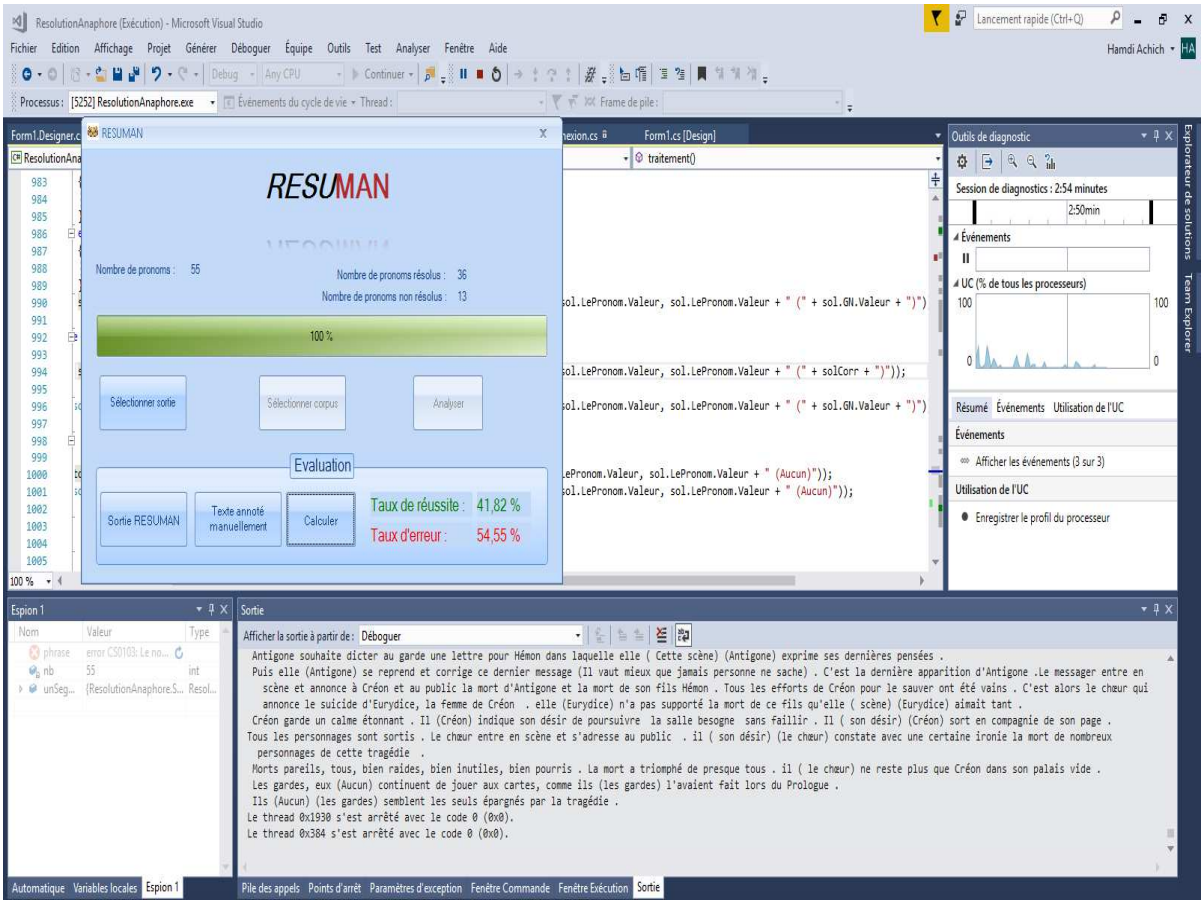
Tous les personnages sont sortis . Le chœur entre en scène et s'adresse au public . il (le chœur) constate avec une certaine ironie la mort de nombreux personnages de cette tragédie .

orts pareils, tous, bien raides, bien inutiles, bien pourris . La mort a triomphé de presque tous . il () ne reste plus que Créon dans son palais vide .

Les gardes, eux (Aucun) continuent de jouer aux cartes, comme ils l'avaient fait lors du Prologue .

Ils (Aucun) semblent les seuls épargnés par la tragédie .

Ultime dérision .



Annexe 4 : Résumé L'Herbe Bleue

a) Résumé brut de L'Herbe Bleu : http://www.alalettre.com/anonyme-oeuvres-herbe-bleue.php

<Œuvre> L'Herbe bleue
<Auteur> Anonyme
<Genre> Journal
<Résumé> L'Herbe bleue est le journal intime d'une jeune fille de quinze ans qui va sombrer dans la drogue.
«Hier je me croyais la personne la plus heureuse de la terre, de toute la galaxie, de toute la création. Etait-ce seulement hier ou bien à des millions d'années-lumière ? Je pensais que l'herbe n'avait jamais eu d'odeur aussi verte, que le ciel n'avait jamais été aussi haut … »
Dés les premières lignes de ce journal, on pénètre dans l'intimité et les sentiments de cette jeune fille complexée et fragile qui semble pourtant avoir une vie de famille agréable. Elle confie ses joies et ses peines à son journal comme à un véritable confident. En mal de vivre et sans amis, elle se rend à une soirée délirante qui va bouleverser son existence. L'un des organisateurs de la soirée annonce un nouveau jeu : « ce soir nous allons jouer au furet, tu sais : il court , il court le furet ». Quelqu'un met de la drogue dans des verres, ils sont distribués et on ne sait pas qui va le boire. Elle y goûte, et brusquement tout bascule. Les couleurs se mélangent, elle rit, elle a chaud, elle plane « je me suis sentie toute drôle , comme s'il y avait une tempête en moi » ... S'apercevant avec horreur de ce qu'elle vient de faire, elle décide de ne plus jamais en reprendre. Malheureusement, elle ne parviendra pas à tenir son engagement. Elle en reprendra, découvrira de nouvelles sensations, en revendra et fuguera. Quand elle revient chez elle, ses parents l'emmènent dans un centre psychiatrique et de désintoxication.
Ce journal décrit le monde isolé et désespéré des jeunes toxicomanes. Non pas de l'extérieur : cette descente aux enfers est vécue de l'intérieur par une adolescente qui lutte en vain. Le journal donne un aperçu réel de la spirale incontrôlable dans laquelle elle se trouve : un parcours sans issue, où de vraies rechutes succèdent à de faux espoirs
« Triste et poignant, ce journal nous retrace la descente aux enfers d'une génération entière face à laquelle les adultes sont impuissants. Style simple et parlé qui restitue authentiquement l'univers intérieur de l'adolescente ».
À la fin du journal, un épilogue : « L'auteur de ce journal est morte trois semaines après avoir pris la décision de ne plus en tenir un. Ses parents sont rentrés un soir du cinéma et l'ont trouvée morte. Ils ont appelé la police , une ambulance, mais il n'y avait plus rien à faire. Etait-ce une dose trop forte ? Accidentelle ? Préméditée ? Personne ne le sait et cela n'a que peu d'importance , dans le fond. Ce qui importe , c'est que cette enfant est morte. Et qu'elle n'est qu'une des cinquante mille victimes de la drogue qui succombèrent cette année-là. »

b) Sortie RESUMAN sans correction manuelle de l'étiquetage de Fips

<numéro 1> .<Œuvre> L'Herbe bleue .<Auteur> Anonyme .<Genre> Journal .<Résumé> L'Herbe bleue est le journal intime d'une jeune fille de quinze ans qui va sombrer dans la drogue .Hier je me croyais la personne la plus heureuse de la terre, de toute la galaxie, de toute la création . Etait-ce seulement hier ou bien à des millions d'années-lumière . Je pensais que l'herbe n'avait jamais eu d'odeur aussi verte, que le ciel n'avait jamais été aussi haut … .Dés les premières lignes de ce journal, on pénètre dans l'intimité et les sentiments de cette jeune fille complexée et fragile qui semble pourtant avoir une vie de famille agréable . Elle (Aucun) confie ses joies et ses peines à son journal comme à un véritable confident .

En mal de vivre et sans amis, elle (Aucun) se rend à une soirée délirante qui va bouleverser son existence .

L'un des organisateurs de la soirée annonce un nouveau jeu . ce soir nous allons jouer au furet, tu sais . il court , il (un jeu) court le furet .

Quelqu'un met de la drogue dans des verres, ils (des verres) sont distribués et on ne sait pas qui va le boire .

Elle (Aucun) y goûte, et brusquement tout bascule .

Les couleurs se mélangent, elle rit, elle a chaud, elle (Aucun) plane je me suis sentie toute drôle , comme s'il y avait une tempête en moi .

S'apercevant avec horreur de ce qu'elle vient de faire, elle (une tempête) décide de ne plus jamais en reprendre .

Malheureusement, elle (Aucun) ne parviendra pas à tenir son engagement .

Elle (Aucun) en reprendra, découvrira de nouvelles sensations, en revendra et fuguera .

Quand elle revient chez elle, ses parents l'emmènent dans un centre psychiatrique et de désintoxication .Ce journal décrit le monde isolé et désespéré des jeunes toxicomanes . Non pas de l'extérieur . cette descente aux enfers est vécue de l'intérieur par une adolescente qui lutte en vain . Le journal donne un aperçu réel de la spirale incontrôlable dans laquelle elle (cette descente) se trouve .

un parcours sans issue, où de vraies rechutes succèdent à de faux espoirs . Triste et poignant, ce journal nous retrace la descente aux enfers d'une génération entière face à laquelle les adultes sont impuissants . Style simple et parlé qui restitue authentiquement l'univers intérieur de l'adolescente . À la fin du journal, un épilogue . L'auteur de ce journal est morte trois semaines après avoir pris la décision de ne plus en tenir un . Ses parents sont rentrés un soir du cinéma et l'ont trouvée morte . Ils (Ses parents) ont appelé la police , une ambulance, mais il n'y avait plus rien à faire .

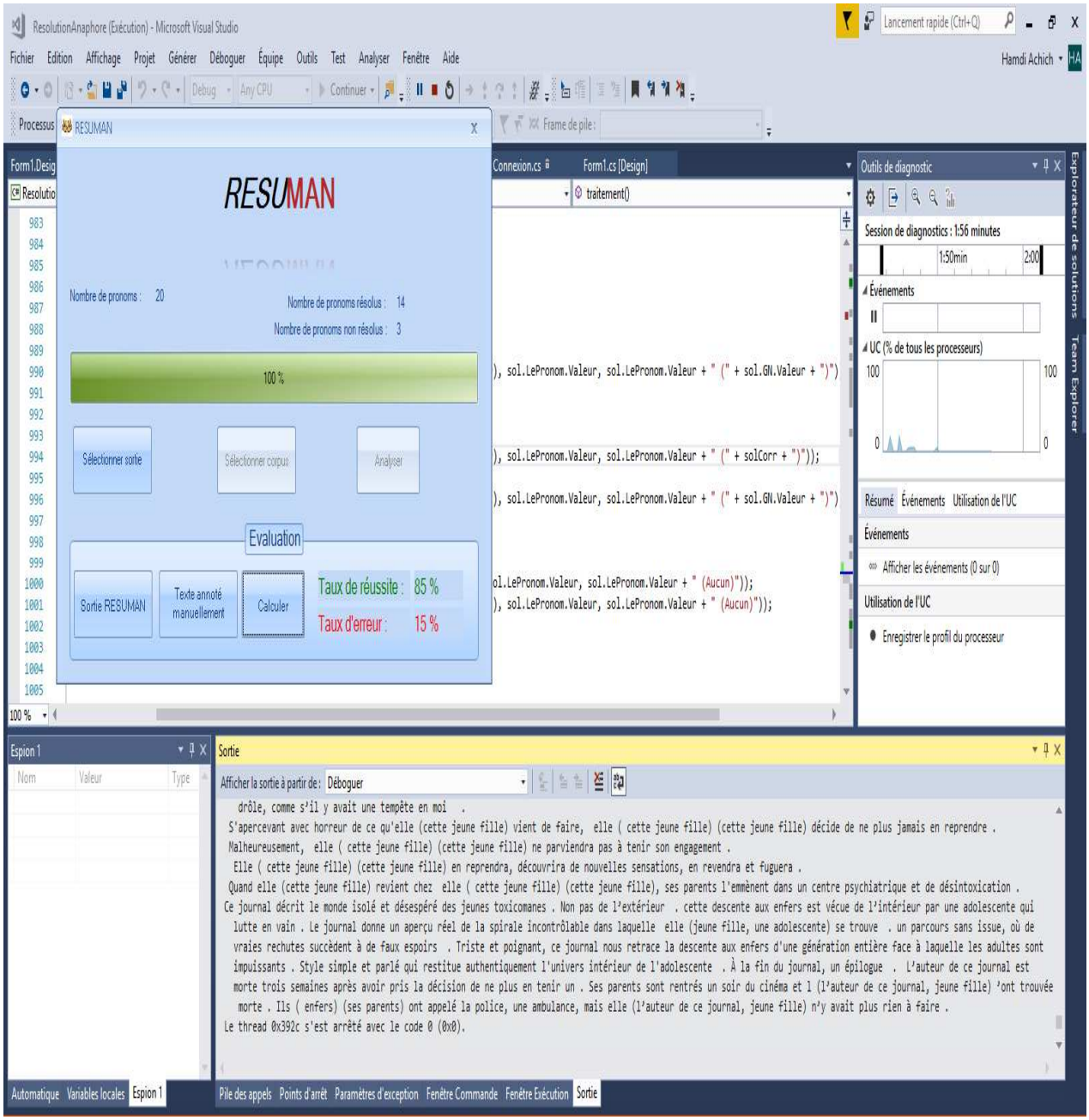
c) Sortie RESUMAN après correction manuelle de l'étiquetage de Fips

<Œuvre> L'Herbe bleue .<Auteur> Anonyme .<Genre> Journal .<Résumé> L'Herbe bleue est le journal intime d'une jeune fille de quinze ans qui va sombrer dans la drogue .Hier je me croyais la personne la plus heureuse de la terre, de toute la galaxie, de toute la création . Etait-ce seulement hier ou bien à des millions d'années-lumière . Je pensais que l'herbe n'avait jamais eu d'odeur aussi verte, que le ciel n'avait jamais été aussi haut … .Dés les premières lignes de ce journal, on pénètre dans l'intimité et les sentiments de cette jeune fille complexée et fragile qui semble pourtant avoir une vie de famille agréable . Elle (cette jeune fille) confie ses joies et ses peines à son journal comme à un véritable confident .
En mal de vivre et sans amis, elle (cette jeune fille) se rend à une soirée délirante qui va bouleverser son existence .
L'un des organisateurs de la soirée annonce un nouveau jeu . ce soir nous allons jouer au furet, tu sais . elle (cette jeune fille) court, elle (cette jeune fille) court le furet .

Quelqu'un met de la drogue dans des verres, elles (des verres) sont distribuées et on ne sait pas qui va le boire .
Elle (Aucun) y goûte, et brusquement tout bascule .
Les couleurs se mélangent, elle (cette jeune fille) rit, elle (cette jeune fille) a chaud, elle (cette jeune fille) plane je me suis sentie toute drôle, comme s'il y avait une tempête en moi .

S'apercevant avec horreur de ce qu'elle (cette jeune fille) vient de faire, elle (cette jeune fille) décide de ne plus jamais en reprendre .
Malheureusement, elle (cette jeune fille) ne parviendra pas à tenir son engagement .
Elle (cette jeune fille) en reprendra, découvrira de nouvelles sensations, en revendra et fuguera .
Quand elle (cette jeune fille) revient chez elle, ses parents l'emmènent dans un centre psychiatrique et de désintoxication.

Ce journal décrit le monde isolé et désespéré des jeunes toxicomanes. Non pas de l'extérieur . cette descente aux enfers est vécue de l'intérieur par une adolescente qui lutte en vain . Le journal donne un aperçu réel de la spirale incontrôlable dans laquelle elle (jeune fille, une adolescente) se trouve. un parcours sans issue, où de vraies rechutes succèdent à de faux espoirs. Triste et poignant, ce journal nous retrace la descente aux enfers d'une génération entière face à laquelle les adultes sont impuissants . Style simple et parlé qui restitue authentiquement l'univers intérieur de l'adolescente . À la fin du journal, un épilogue . L'auteur de ce journal est morte trois semaines après avoir pris la décision de ne plus en tenir un . Ses parents sont rentrés un soir du cinéma et l (l'auteur de ce journal) 'ont trouvée morte . Ils (enfers) (ses parents) ont appelé la police, une ambulance, mais elle (l'auteur de ce journal) n'y avait plus rien à faire .


ResolutionAnaphore (Exécution) - Microsoft Visual Studio
RESUMAN
Nombre de pronoms : 20
Nombre de pronoms résolus : 14
Nombre de pronoms non résolus : 3
100 %
Sélectionner sortie
Sélectionner corpus
Analyser
Evaluation
Sortie RESUMAN
Texte annoté manuellement
Calculer
Taux de réussite : 85 %
Taux d'erreur : 15 %
Outils de diagnostic
Session de diagnostics : 1:56 minutes
Événements
UC (% de tous les processeurs)
Sortie
drôle, comme s'il y avait une tempête en moi .
S'apercevant avec horreur de ce qu'elle (cette jeune fille) vient de faire, elle (cette jeune fille) (cette jeune fille) décide de ne plus jamais en reprendre .
Malheureusement, elle (cette jeune fille) (cette jeune fille) ne parviendra pas à tenir son engagement .
Elle (cette jeune fille) (cette jeune fille) en reprendra, découvrira de nouvelles sensations, en revendra et fuguera .
Quand elle (cette jeune fille) revient chez elle (cette jeune fille) (cette jeune fille), ses parents l'emmènent dans un centre psychiatrique et de désintoxication .
Ce journal décrit le monde isolé et désespéré des jeunes toxicomanes . Non pas de l'extérieur . cette descente aux enfers est vécue de l'intérieur par une adolescente qui lutte en vain . Le journal donne un aperçu réel de la spirale incontrôlable dans laquelle elle (jeune fille, une adolescente) se trouve . un parcours sans issue, où de vraies rechutes succèdent à de faux espoirs . Triste et poignant, ce journal nous retrace la descente aux enfers d'une génération entière face à laquelle les adultes sont impuissants . Style simple et parlé qui restitue authentiquement l'univers intérieur de l'adolescente . À la fin du journal, un épilogue . L'auteur de ce journal est morte trois semaines après avoir pris la décision de ne plus en tenir un . Ses parents sont rentrés un soir du cinéma et l (l'auteur de ce journal, jeune fille) 'ont trouvée morte . Ils (enfers) (ses parents) ont appelé la police, une ambulance, mais elle (l'auteur de ce journal, jeune fille) n'y avait plus rien à faire .
Le thread 0x392c s'est arrêté avec le code 0 (0x0).

Annexe 5 : Résumé Candide

a) Résumé brut de Candide : http://www.alalettre.com/voltaire-oeuvres-candide.php

<Œuvre> Candide
<Auteur> Voltaire
<Genre> conte philosophique
<Résumé> Le jeune Candide, dont le nom traduit à la fois la naïveté et la crédulité vit dans le "meilleur des mondes possibles" chez son oncle, le baron de Thunder-ten-Tronckh.
Enfant naturel, Candide mène une existence heureuse dans cet univers idyllique : Le baron et la baronne de Thunder-ten-Tronckh possèdent en effet "le plus beau des châteaux". Candide est ébloui par la puissance de son oncle, et par les sophismes lénifiants du docteur Pangloss, le précepteur. Il admire également Cunégonde, la fille du baron. Tout bascule le jour des premiers ébats de Candide et de Cunégonde. La réaction du baron est brutale, Candide est banni et chassé de cet Eden. Il se retrouve dans "le vaste monde".
Candide est pris dans une tempête de neige et connaît la faim et le froid. Il est enrôlé de force comme soldat de l'armée bulgare. I prend la fuite. Capturé, il est condamné à recevoir quatre mille coups de bâton. Il échappe de justesse à la mort. Il assiste alors à la guerre et à ses massacres : c'est "une boucherie héroïque". Candide déserte et fuit jusqu'en Hollande. Il y découvre l'intolérance, et notamment l'hypocrisie sectaire d'un prédicateur huguenot. Il retrouve Pangloss rongé par la vérole. Son ancien précepteur a des allures de gueux. Il lui apprend que le beau château du baron Thunder-ten-Tronckh a été détruit et que Cunégonde a été violée et éventrée par les soldats bulgares. L'armée bulgare a également tué le baron, la baronne et leur fils. Candide et Pangloss sont recueillis et embauchés par Jacques, un bon anabaptiste qui les emmène au Portugal où le réclame son commerce. Hélas, au large de Lisbonne, leur navire connaît une horrible tempête. Le bateau du généreux négociant est englouti et ce dernier périt dans le naufrage. Candide et Pangloss en réchappent par miracle. Dès leur arrivée à Lisbonne, se produit un épouvantable tremblement de terre. Candide et Pangloss participent eux opérations de sauvetage, mais nos deux héros sont arrêtés pour propos subversifs et déférés à l'Inquisition. Pangloss est pendu et Candide flagellé. Une vieille dame le soigne et le mène de nuit dans une maison isolée. Il est présenté à une superbe femme : Cunégonde. Elle lui confirme qu'elle a été violée et éventrée, et que c'est par miracle qu'elle est encore en vie : "on ne meurt pas toujours de ces deux accidents". Cunégonde est devenue à la fois la maîtresse de Don Issachar, un banquier juif et du grand inquisiteur de Lisbonne. Menacé par ses deux rivaux, "le doux Candide", parvient à les tuer. Candide, Cunégonde et la vieille dame s'enfuient alors en direction de Cadix. Ils arrivent à Cadix au moment où un bateau s'apprête à partir en Amérique latine. Son équipage est chargé d'aller y combattre la rébellion qui règne contre les rois d'Espagne et du Portugal. Candide parvient à se faire engager. Il embarque avec Cunégonde, la vieille dame et deux valets. Lors de la traversée, la vieille dame raconte son aventure. Fille d'un pape et d'une princesse, elle a grandi " en beauté, en grâces, en talents, au milieu des plaisirs, des respects et des espérances..." Puis elle a connu une suite épouvantable de malheurs : l'empoisonnement de son fiancé, l'enlèvement de sa mère, sa vente à des marchands d'esclaves. Elle s'est retrouvée prisonnière dans un fort, puis elle est devenue l'esclave d'un seigneur moscovite qui l'a batttue. Elle finira par devenir la servante de Don Issachar qui la met à disposition de Cunégonde à qui elle se lie.

Suite à ce récit, la vieille dame demande aux autres passagers de raconter leur histoire. Les récits s'enchaînent, plus noirs les uns que les autres. Candide commence à prendre conscience que le mal existe sur cette terre.
A peine arrivés à Buenos Aires, Candide et Cunégonde sont à nouveau séparés. La vielle dame conseille en effet à Cunégonde de rester auprès du gouverneur qui s'est épris d'elle et à Candide de fuir l'Inquisition qui a retrouvé sa trace. Candide part avec son valet Cacambo se réfugier chez les jésuites du Paraguay. Ils y retrouvent le frère de Cunégonde, lui aussi miraculeusement rescapé. Le baron évoque son miracle : Alors qu'on allait l'enterrer, le battement de sa paupière l'a sauvé. On l'a soigné et guéri. Sa beauté, fort appréciée, lui a valu une grande fortune. Mais le jeune baron refuse qu'un bâtard puisse épouser sa sœur et frappe Candide du plat de son épée. Celui-ci se défend et le tue d'un coup d'épée.
Candide et Cacambo reprennent la fuite et se retrouvent dans un pays inconnu. Il sont faits prisonniers par les indigènes et sont à deux doigts d'être mangés. Ils ne doivent leur salut qu'à la verve et à l'habileté de Cacambo. Ils sont graciés.
Ils se dirigent alors vers Cayenne, à la recherche de la colonie française. Ils souffrent de la faim. Un jour, ils découvrent un canot sur une rivière. Ils montent à bord et se laissent porter par le courant. Le canot emprunte une voûte secrète. Candide et Cacambo se retrouvent sous terre, dans une magnifique contrée, l'Eldorado, "le pays où tout va bien" : un pays où les repas sont délicieux, les mœurs pacifiques, la population heureuse , la religion tolérante et le souverain humaniste. Mais nos héros sont trop vaniteux pour se satisfaire de cet univers idéal. Ils souhaitent revenir en Europe avec l'espoir d'éblouir Cunégonde et le monde entier de leur récit et de leur richesse. Le souverain du royaume en effet les laisse partir avec cent moutons chargés de nourriture, de pierres précieuses et d'or. Il les met aussi en garde : le bonheur ne se trouve ni dans les pierres précieuses ni dans l'or.
Candide et Cacambo retrouvent le monde. Pendant plus de trois mois, ils marchent dans les marais, les déserts et au bord des précipices. Leurs moutons meurent les uns après les autres. Lorsqu'ils arrivent à Surinam, ils n'ont plus que deux moutons. Ils rencontrent alors un esclave noir atrocement mutilé. Ceci révolte Candide et l'amène à donner une autre définition de l'optimisme : " la rage de soutenir que tout est bien quand on est mal".
Nos deux héros se séparent : Candide envoie Cacambo racheter Cunégonde au gouverneur de Buenos Aires , tandis qu'il ira l'attendre à Venise.
Mais Candide se fait duper et voler par un marchand qui lui prend ses deux derniers moutons et s'embarque pour Venise sans l'attendre. Il parvient finalement à trouver un vaisseau en partance pour Bordeaux et s'embarque en compagnie d'un pauvre savant persécuté à qui il paye son voyage. Il a l'espoir que ce compagnon puisse le "désennuyer" durant le traversée.
Sur le bateau qui les emmène à Bordeaux Candide et Martin, le savant discutent du bien et du mal et de la nature de l'homme. Martin lui indique qu'il est convaincu de la prédominance du Mal sur le Bien . Et comme pour illustrer son propos, ils assistent un combat entre un navire espagnol et un vaisseau hollandais . Ce dernier coule et une centaine d'hommes se noient. Ce combat est pour Martin l'illustration des rapports humains de la façon dont " les hommes se traitent les uns les autres."
Après son arrivée à Bordeaux, Candide préfère se rendre à Paris qu'à Venise. Il n'y connaît qu'amertume et déception : un abbé retors et de fausses marquises et une fausse Cunégonde qui se révèlent être de vraies voleuses . Il se fait même injustement arrêter et ne parvient à s'enfuir qu'en soudoyant un officier de police.
Il embarque alors en compagnie de Martin pour l'Angleterre. Il assiste à l'exécution d'un amiral condamné pour " n'avoir pas fait tuer assez de monde." Finalement, il refuse de débarquer en Angleterre et demande au capitaine du bateau de l'emmener directement à Venise.

A Venise, il ne retrouve ni Cacambo, ni Cunégonde mais tombe sur Paquette, l'ancienne suivante de la Baronne de Thunder-ten-Tronckh. Elle vit en compagnie d'un moine, Giroflée. Ses confidences et celles du moine font apparaître à Candide des misères cachées. Candide décide alors de rendre visite au seigneur Pococurante qui a la réputation de n'avoir jamais eu de chagrin.
Le jeune héros s'émerveille de l'univers et de la personnalité de son hôte. Pourtant celui-ci évoque a demi-mot le dégoût et la lassitude du blasé. Candide ressort pourtant de cet entretien avec l'impression que le seigneur Pococurante est "le plus heureux de tous les hommes", car affranchi des biens matériels. Martin, lui, est plus pessimiste, il estime que ce seigneur est écœuré de tout ce qu'il possède.
Au milieu d'un souper de carnaval, alors que Candide dîne avec six malheureux anciens rois qui ont perdu leur royaume, il retrouve Cacambo qui est devenu esclave. Il lui apprend que Cunégonde l'attend sur les bords de la Propontide, près de Constantinople. Elle aussi est devenue esclave et est devenue très laide.
Candide se rend à Constantinople . Sur la galère, il croit reconnaître parmi les galériens le docteur Pangloss et le jeune baron (tous deux mal tués). Il les rachète au capitaine du navire.
Les deux anciens galériens racontent leurs aventures, mais le récit de leur malheurs ne perturbe pas Candide qui est toujours convaincu que " tout est pour le mieux dans le meilleur des mondes."
Candide retrouve Cunégonde, et il est saisi d'horreur à la vue de cette femme hideuse et défigurée. Il la rachète ainsi que la vieille femme. Il ne l'aime plus, mais l'épouse " par bonté" malgré le refus répété de son frère.
Candide se débarrasse du jeune baron en le renvoyant aux galères . il achète avec ses derniers diamants une modeste métairie où viennent se réfugier Paquette , le frère Giroflée, Pangloss, Martin, Cunégonde et Candide. Un sage vieillard leur conseille le travail qui "éloigne de nous trois grand maux, l'ennui , le vice et le besoin".
Candide en arrive à cette conclusion qui recueille l'assentiment de tous ses compagnons : " il faut cultiver son jardin."
Quelques Citations de Candide
Pangloss enseignait la métaphysico-théologo-cosmolonigologie. Il prouvait admirablement qu'il n'y a point d'effet sans cause, et que, dans ce meilleur des mondes possibles, le château de monseigneur le baron était le plus beau des châteaux et madame la meilleure des baronnes possibles.
Les malheurs particuliers font le bien général; de sorte que plus il y a de malheurs particuliers et plus tout est bien.
Tout est bien, tout va bien, tout va le mieux qu'il soit possible
Je n'ai que vingt arpents, répondit le Turc ; je les cultive avec mes enfants ; le travail éloigne de nous trois grands maux : l'ennui, le vice, et le besoin. " Travaillons sans raisonner, dit Martin ; c'est le seul moyen de rendre la vie supportable.
Toute la petite société entra dans ce louable dessein ; chacun se mit à exercer ses talents. La petite terre rapporta beaucoup. Cunégonde était à la vérité bien laide ; mais elle devint une excellente pâtissière ; Paquette broda; la vieille eut soin du linge. Il n'y eut pas jusqu'à frère Giroflée qui ne rendît service ; il fut un très bon menuisier, et même devint honnête homme ; et Pangloss disait quelquefois à Candide : " Tous les événements sont enchaînés dans le meilleur des mondes possibles ; car enfin, si vous n'aviez pas été chassé d'un beau château à grands coups de pied dans le derrière pour l'amour de Mlle Cunégonde, si vous n'aviez pas été mis à l'Inquisition, si vous n'aviez pas couru l'Amérique à pied, si vous n'aviez pas donné un bon coup d'épée au baron, si vous n'aviez pas perdu tous vos moutons

du bon pays d'Eldorado, vous ne mangeriez pas ici des cédrats confits et des pistaches. -- Cela est bien dit, répondit Candide, mais il faut cultiver notre jardin.

b) Sortie RESUMAN$_O$

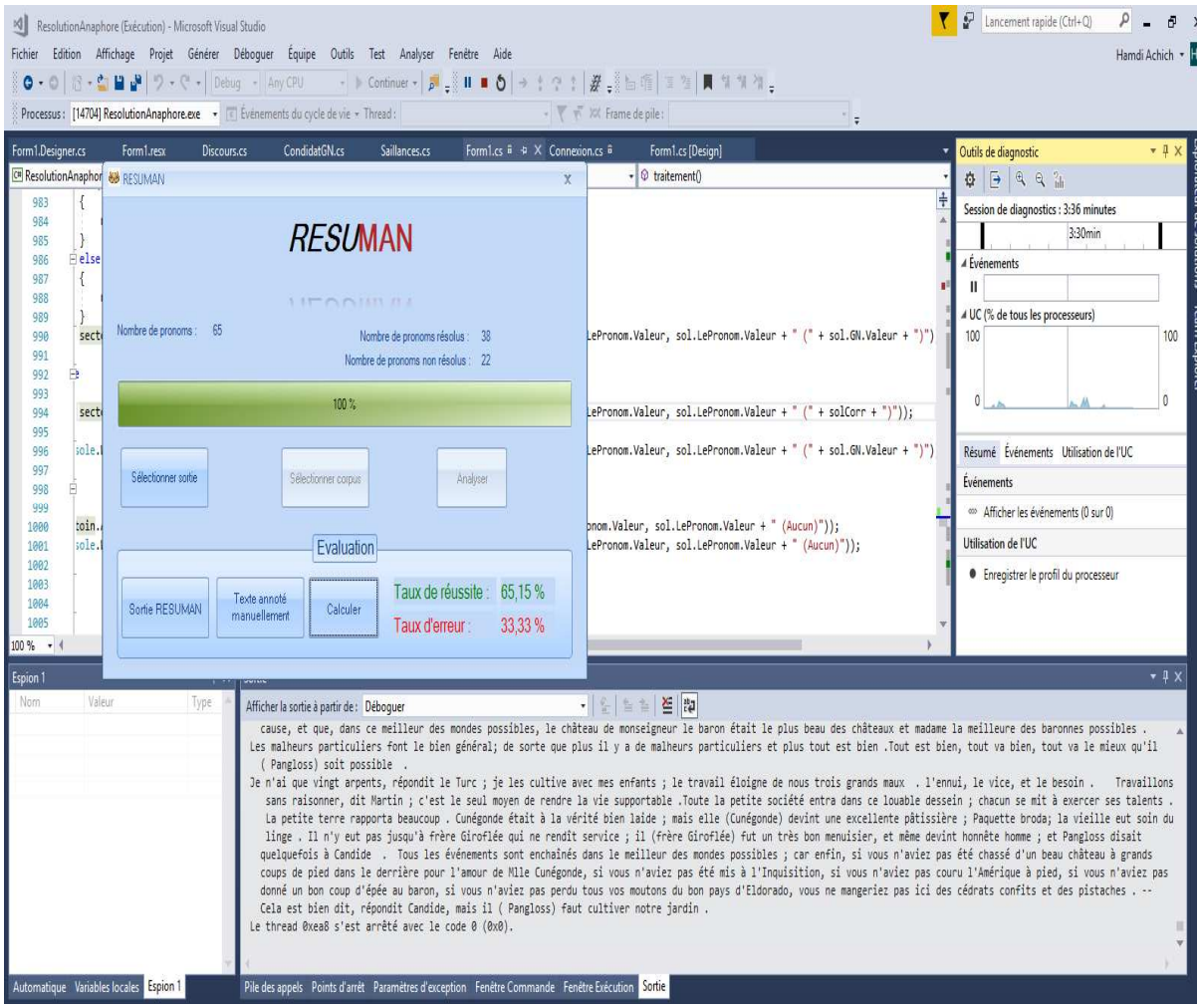
<numéro 100> .<Œuvre> Candide .<Auteur> Voltaire .<Genre> conte philosophique .<Résumé> Le jeune Candide, dont le nom traduit à la fois la naïveté et la crédulité vit dans le meilleur des mondes possibles chez son oncle, le baron de Thunder-ten-Tronckh .Enfant naturel, Candide mène une existence heureuse dans cet univers idyllique . Le baron et la baronne de Thunder-ten-Tronckh possèdent en effet le plus beau des châteaux . Candide est ébloui par la puissance de son oncle, et par les sophismes lénifiants du docteur Pangloss, le précepteur . Il (son oncle) admire également Cunégonde, la fille du baron .
Tout bascule le jour des premiers ébats de Candide et de Cunégonde . La réaction du baron est brutale, Candide est banni et chassé de cet Eden . Il (baron) se retrouve dans le vaste monde .
Candide est pris dans une tempête de neige et connaît la faim et le froid . Il (baron) est enrôlé de force comme soldat de l'armée bulgare .
Il (baron) prend la fuite .
Capturé, il est condamné à recevoir quatre mil (baron)le coups de bâton .
Il (baron) échappe de justesse à la mort .
Il (baron) assiste alors à la guerre et à ses massacres .
c'est une boucherie héroïque . Candide déserte et fuit jusqu'en Hollande . Il (baron) y découvre l'intolérance, et notamment l'hypocrisie sectaire d'un prédicateur huguenot .
Il (baron) retrouve Pangloss rongé par la vérole .
Son ancien précepteur a des allures de gueux . Il (baron) lui apprend que le beau château du baron Thunder-ten-Tronckh a été détruit et que Cunégonde a été violée et éventrée par les soldats bulgares .
L'armée bulgare a également tué le baron, la baronne et leur fils . Candide et Pangloss sont recueillis et embauchés par Jacques, un bon anabaptiste qui les emmène au Portugal où le réclame son commerce . Hélas, au large de Lisbonne, leur navire connaît une horrible tempête . Le bateau du généreux négociant est englouti et ce dernier périt dans le naufrage . Candide et Pangloss en réchappent par miracle . Dès leur arrivée à Lisbonne, se produit un épouvantable tremblement de terre . Candide et Pangloss participent aux opérations de sauvetage, mais nos deux héros sont arrêtés pour propos subversifs et déférés à l'Inquisition . Pangloss est pendu et Candide flagellé . Une vieille dame le soigne et le mène de nuit dans une maison isolée . Il (Pangloss) est présenté à une superbe femme .
Cunégonde . Elle (une femme) lui confirme qu'elle (une femme Cunégonde) a été violée et éventrée, et que c'est par miracle qu'elle (une femme Cunégonde) est encore en vie .
on ne meurt pas toujours de ces deux accidents . Cunégonde est devenue à la fois la maîtresse de Don Issachar, un banquier juif et du grand inquisiteur de Lisbonne . Menacé par ses deux rivaux, le doux Candide, parvient à les tuer . Candide, Cunégonde et la vieille dame s'enfuient alors en direction de Cadix . Ils (Candide, Cunégonde et la vieille dame) arrivent à Cadix au moment où un bateau s'apprête à partir en Amérique latine . Son équipage est chargé d'aller y combattre la rébellion qui règne contre les rois d'Espagne et du Portugal . Candide parvient à se faire engager . Il (Aucun) embarque avec Cunégonde, la vieille dame et deux valets .
Lors de la traversée, la vieille dame raconte son aventure . Fille d'un pape et d'une princesse, elle (la dame) a grandi en beauté, en grâces, en talents, au milieu des plaisirs, des respects et des espérances .

Puis elle (la dame) a connu une suite épouvantable de malheurs .
l'empoisonnement de son fiancé, l'enlèvement de sa mère, sa vente à des marchands d'esclaves . Elle (la dame) s'est retrouvée prisonnière dans un fort, puis elle (la vieille dame) est devenue l'esclave d'un seigneur moscovite qui l'a batttue .
Elle (la vieille dame) finira par devenir la servante de Don Issachar qui la met à disposition de Cunégonde à qui elle (la vieille dame) se lie .Suite à ce récit, la vieille dame demande aux autres passagers de raconter leur histoire . Les récits s'enchaînent, plus noirs les uns que les autres . Candide commence à prendre conscience que le mal existe sur cette terre .A peine arrivés à Buenos Aires, Candide et Cunégonde sont à nouveau séparés . La vielle dame conseille en effet à Cunégonde de rester auprès du gouverneur qui s'est épris d'elle (Cunégonde) et à Candide de fuir l'Inquisition qui a retrouvé sa trace . Candide part avec son valet Cacambo se réfugier chez les jésuites du Paraguay . Ils (Candide et Cacambo) y retrouvent le frère de Cunégonde, lui aussi miraculeusement rescapé . Le baron évoque son miracle . Alors qu'on allait l'enterrer, le battement de sa paupière l'a sauvé . On l'a soigné et guéri . Sa beauté, fort appréciée, lui a valu une grande fortune . Mais le jeune baron refuse qu'un bâtard puisse épouser sa sœur et frappe Candide du plat de son épée . Celui-ci se défend et le tue d'un coup d'épée .Candide et Cacambo reprennent la fuite et se retrouvent dans un pays inconnu . Ils (Candide et Cacambo) sont faits prisonniers par les indigènes et sont à deux doigts d'être mangés .
Ils (Candide et Cacambo) ne doivent leur salut qu'à la verve et à l'habileté de Cacambo . Ils (Candide et Cacambo) sont graciés .
Ils (Candide et Cacambo) se dirigent alors vers Cayenne, à la recherche de la colonie française . Ils (Candide et Cacambo) souffrent de la faim .
Un jour, ils (Candide et Cacambo) découvrent un canot sur une rivière .
Ils (Candide et Cacambo) montent à bord et se laissent porter par le courant .
Le canot emprunte une voûte secrète . Candide et Cacambo se retrouvent sous terre, dans une magnifique contrée, l'Eldorado, le pays où tout va bien . un pays où les repas sont délicieux, les mœurs pacifiques, la population heureuse , la religion tolérante et le souverain humaniste . Mais nos héros sont trop vaniteux pour se satisfaire de cet univers idéal . Ils (Candide et Cacambo) souhaitent revenir en Europe avec l'espoir d'éblouir Cunégonde et le monde entier de leur récit et de leur richesse . Le souverain du royaume en effet les laisse partir avec cent moutons chargés de nourriture, de pierres précieuses et d'or . Il (le souverain) les met aussi en garde . le bonheur ne se trouve ni dans les pierres précieuses ni dans l'or .Candide et Cacambo retrouvent le monde . Pendant plus de trois mois, ils (Candide et Cacambo) marchent dans les marais, les déserts et au bord des précipices .
Leurs moutons meurent les uns après les autres . Lorsqu'ils (Candide et Cacambo) arrivent à Surinam, ils (Candide et Cacambo) n'ont plus que deux moutons .
Ils (Candide et Cacambo) rencontrent alors un esclave noir atrocement mutilé . Ceci révolte Candide et l'amène à donner une autre définition de l'optimisme . la rage de soutenir que tout est bien quand on est mal .Nos deux héros se séparent . Candide envoie Cacambo racheter Cunégonde au gouverneur de Buenos Aires , tandis qu'il (Candide) ira l'attendre à Venise .
Mais Candide se fait duper et voler par un marchand qui lui prend ses deux derniers moutons et s'embarque pour Venise sans l'attendre . Il (Candide) parvient finalement à trouver un vaisseau en partance pour Bordeaux et s'embarque en compagnie d'un pauvre savant persécuté à qui il (Candide) paye son voyage . Il (Candide) a l'espoir que ce compagnon puisse le désennuyer durant le traversée .Sur le bateau qui les emmène à Bordeaux Candide et Martin, le savant discutent du bien et du mal et de la nature de

l'homme . Martin lui indique qu'il (Martin) est convaincu de la prédominance du Mal sur le Bien .
Et comme pour illustrer son propos, ils (Aucun) assistent un combat entre un navire espagnol et un vaisseau hollandais .
Ce dernier coule et une centaine d'hommes se noient . Ce combat est pour Martin l'illustration des rapports humains de la façon dont les hommes se traitent les uns les autres .Après son arrivée à Bordeaux, Candide préfère se rendre à Paris qu'à Venise . Il (Candide) n'y connaît qu'amertume et déception . un abbé retors et de fausses marquises et une fausse Cunégonde qui se révèlent être de vraies voleuses . Il (Candide) se fait même injustement arrêter et ne parvient à s'enfuir qu'en soudoyant un officier de police .
Il (Candide) embarque alors en compagnie de Martin pour l'Angleterre .
Il (Candide) assiste à l'exécution d'un amiral condamné pour n'avoir pas fait tuer assez de monde .
Finalement, il (Candide) refuse de débarquer en Angleterre et demande au capitaine du bateau de l'emmener directement à Venise .A Venise, il (Candide) ne retrouve ni Cacambo, ni Cunégonde mais tombe sur Paquette, l'ancienne suivante de la Baronne de Thunder-ten-Tronckh . Elle (l'ancienne suivante de la Baronne de Thunder-ten-Tronckh) vit en compagnie d'un moine, Giroflée . Ses confidences et celles du moine font apparaître à Candide des misères cachées . Candide décide alors de rendre visite au seigneur Pococurante qui a la réputation de n'avoir jamais eu de chagrin .Le jeune héros s'émerveille de l'univers et de la personnalité de son hôte . Pourtant celui-ci évoque a demi-mot le dégoût et la lassitude du blasé . Candide ressort pourtant de cet entretien avec l'impression que le seigneur Pococurante est le plus heureux de tous les hommes, car affranchi des biens matériels . Martin, lui, est plus pessimiste, il (Martin) estime que ce seigneur est écœuré de tout ce qu'il (le seigneur) possède .
Au milieu d'un souper de carnaval, alors que Candide dîne avec six malheureux anciens rois qui ont perdu leur royaume, il (Candide) retrouve Cacambo qui est devenu esclave .
Il (Cacambo) lui apprend que Cunégonde l'attend sur les bords de la Propontide, près de Constantinople .
Elle (Cunégonde) aussi est devenue esclave et est devenue très laide .
Candide se rend à Constantinople . Sur la galère, il (Candide) croit reconnaître parmi les galériens le docteur Pangloss et le jeune baron (tous deux mal tués) . Il (Aucun) les rachète au capitaine du navire .
Les deux anciens galériens racontent leurs aventures, mais le récit de leur malheurs ne perturbe pas Candide qui est toujours convaincu que tout est pour le mieux dans le meilleur des mondes .Candide retrouve Cunégonde, et il (Candide) est saisi d'horreur à la vue de cette femme hideuse et défigurée . Il (Aucun) la rachète ainsi que la vieille femme .
Il (Aucun) ne l'aime plus, mais l'épouse par bonté malgré le refus répété de son frère .
Candide se débarrasse du jeune baron en le renvoyant aux galères . Il achète avec ses derniers diamants une modeste métairie où viennent se réfugier Paquette , le frère Giroflée, Pangloss, Martin, Cunégonde et Candide . Un sage vieillard leur conseille le travail qui éloigne de nous trois grand maux, l'ennui , le vice et le besoin .Candide en arrive à cette conclusion qui recueille l'assentiment de tous ses compagnons . il (Candide) faut cultiver son jardin .
Quelques Citations de Candide .Pangloss enseignait la métaphysico-théologo-cosmolonigologie . Il (Pangloss) prouvait admirablement qu'il n'y a point d'effet sans cause, et que, dans ce meilleur des mondes possibles, le château de monseigneur le baron était le plus beau des châteaux et madame la meilleure des baronnes possibles .
s malheurs particuliers font le bien général; de sorte que plus il y a de malheurs particuliers et plus tout est bien .Tout est bien, tout va bien, tout va le mieux qu'il () soit possible .

Je n'ai que vingt arpents, répondit le Turc ; je les cultive avec mes enfants ; le travail éloigne de nous trois grands maux . l'ennui, le vice, et le besoin .

Travaillons sans raisonner, dit Martin ; c'est le seul moyen de rendre la vie supportable .Toute la petite société entra dans ce louable dessein ; chacun se mit à exercer ses talents . La petite terre rapporta beaucoup . Cunégonde était à la vérité bien laide ; mais elle (Cunégonde) devint une excellente pâtissière ; Paquette broda; la vieille eut soin du linge . Il n'y eut pas jusqu'à frère Giroflée qui ne rendît service ; il fut un très bon menuisier, et même devint honnête homme ; et Pangloss disait quelquefois à Candide . Tous les événements sont enchaînés dans le meilleur des mondes possibles ; car enfin, si vous n'aviez pas été chassé d'un beau château à grands coups de pied dans le derrière pour l'amour de Mlle Cunégonde, si vous n'aviez pas été mis à l'Inquisition, si vous n'aviez pas couru l'Amérique à pied, si vous n'aviez pas donné un bon coup d'épée au baron, si vous n'aviez pas perdu tous vos moutons du bon pays d'Eldorado, vous ne mangeriez pas ici des cédrats confits et des pistaches . -- Cela est bien dit, répondit Candide, mais il (frère Giroflée) faut cultiver notre jardin .



Bibliographie

Références Bibliographiques :

Abeillé, A. (1993). Les nouvelles syntaxes : grammaires d'unification et analyse du français. *Cahiers de praxématique,* Armand Collin, Paris, [En ligne]. http://praxematique.revues.org/2283

Achard-Bayle, G. (2001). L'anaphore pronominale en contexte évolutif. *Grammaire des métamorphoses : Référence, identité, changement, fiction*, Louvain-la-Neuve, Belgique : De Boeck Supérieur, 119-134.

Adam, C. (2007). *Traitement automatique de la coréférence. Pour une application de veille terminologique*. Mémoire de master, Université de Toulouse.

Adam, J.-M. (1992). *Les textes : types et prototypes. Récit, description, argumentation, explication et dialogue*. Paris, Éditions Nathan.

Adam, J.-M. (1994). *Le texte narratif*. Paris, Nathan.

Adam, J.-M. (2001). Types de textes ou genres de discours ? : comment classer les textes qui disent de et comment faire ?. *Langages, 141*, 10-27.

Adam, J.-M. (2008). Note de cadrage sur la linguistique textuelle. *Congrès Mondial de Linguistique Française*, 133. DOI: https://doi.org/10.1051/cmlf08329

Adam, J.-M. (2005). Les sciences de l'établissement du texte et la question de la variation. *Sciences du texte et analyse de discours*, Adam, J.-M., Heidmann, U. (éds.), Genève, Slatkine, 69-96.

Alkhatib, M. (2012). La cohérence et la cohésion textuelles : problème linguistique ou pédagogique. *Didáctica. Lengua y Literatura* ISSN : 1130-0531, vol. 24 45-64 http://dx.doi.org/10.5209/rev_DIDA.2012.v24.39916

Amsili, P., Beyssade, C., Garreta, A./Roussarie, L. (2002). Chaînes de références et résolveurs d'anaphores. Nancy, Juin 2002. Workshop associé à *TALN'02*

Apothéloz, D. (1995). *Rôle et fonctionnement de l'anaphore dans la dynamique textuelle*. Droz, Genève.

Ariel, M. (1996). Referring expressions and the +/- coreference distinction. In T. Fretheim & J.K. Gundel (eds.), 13-25.

Ariel, M. (2001). Accessibility theory : An overview. Sanders T, Schilperoord J., Spooren W., (dir.), *Text Representation*, Amsterdam, Benjamins, 29-87.

Ariel,M. (1990). *Accessing noun-phrase antecedents*. (Croom Helm Linguistics Series.) London & New York : Routledge.

Asher, N. (2011). *Lexical Meaning in Context : a Web of Words*. Cambridge University Press.

Asher, N./Lascarides, A. (2003). *Logics of conversation*. Cambridge University Press.

Auran, C. (2004). *Prosodie et anaphore dans le discours en anglais et en français : cohésion et attribution référentielle*. Thèse de doctorat : http://www.lpl-aix.fr/~fulltext/2307.pdf

Baldwin, B. (1997). CogNIAC: high precision coreference with limited knowledge and linguistic resources. *Proceedings of a Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts,* 38-45, Association for Computational Linguistics.

Bar-Hillel, Y. (1964). *Language and information*. Addison Wesley, Reading Mass.

Baylon, C./Mignot, X. (2000). *Initiation à la sémantique du langage*. Paris, Nathan.

Bègue, D. (1982). A propos d'un emprunt historique de la grammaire générative à la théorie des langages formels. *Histoire Épistémologie Langage*, *3*(1), 1-19.

Bègue, D. (1986). Les analyseurs syntaxiques. Linx , Volume 14, Numéro 1, 67-105. http://www.persee.fr/doc/linx_0246-8743_1986_num_14_1_1040

Berndtsson, J. (2014). Coreference resolution in BART. *Semantic Analysis in Language Technology Essay Assignment*, January 20, 1-7.

Berrendonner, A. (1983). Connecteurs pragmatiques et anaphore. *Cahiers de Linguistique Française*, 5, 215-246.

Biard, J. (1997). Signifier. *Guillaume d'Ockham : Logique et philosophie*, Paris: Presses Universitaires de France,15-54.

Bittar, A. (2006). *Un algorithme pour la résolution d'anaphores événementielles*. Mémoire de master, Université Paris, 7, Denis Diderot UFR de Linguistique.

Blache, P., Guizol, J., Lévy, F., Nazarenko, A., N'Guema, S., Pasero, R./ Sabatier, P. (1997). Evaluation des systèmes de compréhension de textes. *Rapport final, UREF, CNRS*.

Blanc, N./Brouillet, D. (2003). *Mémoire et compréhension : Lire pour comprendre*. In Press.

Blanche-Benveniste, C. (1996). De l'utilité du corpus linguistique. *Revue française de linguistique appliquée*, Dossier : *Corpus, de leur constitution à leur exploitation*, 25-42.

Bloomfield, L., Barnhart, C.-L., Pooley, R.-C./ Faust, G.-P. (1961). *Let's read : A linguistic approach*. Wayne State University Press.

Bobrow, D. (1968). *Natural language input for a computer problem solving system*. MIT Press, Cambridge.

Bodelot, C. (2004). Anaphore, cataphore et corrélation: approche générale de la problématique dans l'optique de la phrase complexe. *Anaphore, cataphore et corrélation en latin, Actes de la journée d'étude de linguistique latine.* Université Blaise Pascal-Clermont-Ferrand II, 7 janvier 2003, 13-26.

Bodineau, P. (1996). *Le résumé du texte narratif*. Poitiers : CRDP Poitou-Charentes.

Bonhomme, M. (2005). *Pragmatique des figures de style*. Paris, Champion.

Bosch, P. (1985). Constraints, Cohérence, Comprehension. *Text connexity, Text coherence, Aspects methods results, Hamburg, Buske*, 299-520.

Boudreau, S./Kittredge, R. (2006). Résolution d'anaphores et identification des chaînes de coréférence: une approche minimaliste. *Actes des 8èmes Journées internationales d'Analyse statistique des Données Textuelles (JADT)*, 201-210.

Bouquet, S. (2005). Après un siècle, les manuscrits de Saussure reviennent bouleverser la linguistique. *Texto, !* juin 2005 [en ligne]. http://www.revue-texto.net/Saussure/Sur_Saussure/Bouquet_Apres.html

Bourigault, D., Fabre, C. (2000). Approche linguistique pour l'analyse syntaxique de corpus. Cahiers de Grammaire 25, *Sémantique et Corpus*, 131-151.

Bresnan, J./Ronald, K. (1981). Lexical functional grammars : a formal system for grammatical representation. *The mental representation of grammatical relations*, MIT Press, Cambridge, Mass.

Broscheit, S., Poesio, M., Ponzetto, S.-P., Rodriguez, K.-J., Romano, L., Uryupina, O./ Zanoli, R. (2010). BART : A multilingual anaphora resolution system. *Proceedings of the 5th international workshop on semantic evaluation*, Association for Computational Linguistics, 104-107.

Brown, G./Yule, G. (1983). *Discourse analysis*. Cambridge, Cambridge University Press.

Byron, D.-K./Tetreault, J.-R. (1999). A flexible architecture for reference resolution. *Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics,* (EACL'99), Bergen, Norvège, 229-232.

Caillies, S., Denhière, G. (2001). The interaction between textual structures and prior knowledge : Hypotheses, data and simulations. *European Journal of Psychology of Education, 16*(1), 17-31.

Charaudeau, P. (2010). Dis-moi quel est ton corpus, je te dirai quelle est ta problématique. *Corpus* [En ligne]. 8 | 2009. http://corpus.revues.org/1674

Charaudeau, P./Maingueneau, D. (2002). *Dictionnaire d'analyse du discours*. Paris, Seuil.

Charniak, E. (1997). Statistical Techniques for Natural Language Parsing. *AI Magazine*, *18*(4), 33. https://doi.org/10.1609/aimag.v18i4.1320

Charniak, E./Elsner, M. (2009). EM works for pronoun anaphora resolution. *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, 148-156.

Charolles, M. (1978). Introduction aux problèmes de la cohérence des textes : (Approche théorique et étude des pratiques pédagogiques). *Langue française, (38)*, 7-41.

Charolles, M. (1988a). Les études sur la cohérence, la cohésion et la connexité textuelles depuis la fin des années 1960. *Modèles linguistiques, X (2),* 45-66.

Charolles, M. (1988b). Les plans d'organisation textuelle. Périodes, chaînes, portées et séquences. *Pratiques,* 57, 3-13.

Charolles, M. (1991). L'anaphore : définition et classification des formes anaphoriques. *Verbum*, 2-3-4, 203-21

Charolles, M. (2001). De la phrase au discours : quelles relations. A.Rousseau ed. *La sémantique des relations*, Université de Lille III, 237-260, 2001. ffhal-01404546f

Charolles, M. (2005). Cohérence, pertinence et intégration conceptuelle. *Des discours aux textes : modèles et analyses*. P. Lane. (éd). Rouen : Publications des Universités de Rouen et du Havre, 39-74.

Charolles, M. (2006). De la cohérence à la cohésion du discours. *Cohérence et discours*, Calas, F. (éd.), Paris : Presses de l'Université de Paris-Sorbonne, 25-38

Charolles, M. (2011). Cohérence et cohésion du discours. *Dimensionen der Analyse ,Texten und Diskursivent - Dimensioni dell'analisi di testi e discoursi*, Holker, K., Marello, C. ,Lit Verlag, 153-173, 2011, ffhal-00665838f

Charolles, M./Ehrlich, M.-F. (1991). Aspects of textual continuity : Linguistic approach. *Text and text processing*, Denhière, G., Rossi, J.-P. (Eds.), Amsterdam: North Holland.

Charolles, M/Schenedecker, C. (1993). Coréférence et identité, le problème des référents évolutifs. *Langages*, 112, 106-126.

Chartrand, S. (2011). Produire des résumés de textes de genres universitaires http://www.enseignementdufrancais.fse.ulaval.ca/fichiers/site_ens_francais/modules/document_section_fichier/fichier__c192ac44c0b5__Produire_des_resumes_2011.pdf

Chaumartin, F.-R. (2013). Apprentissage d'une classification thématique générique et cross-langue à partir des catégories de la Wikipédia. *TALN - Traitement Automatique des Langues Naturelles - 2013*, ATALA, Jun 2013, Les Sables d'Olonne, France. 659-666. ⟨hal-00851794⟩

Cheng, H. (2001). *Modelling Aggregation Motivated Interactions* in *Descriptive Text Generation*. Ph.D. thesis, University of Edinburgh.

Cheng, H., Poesio, M., Henschel, R./Mellish, C. (2001). Corpus-based NP modifier generation. *Proc. of the Second NAACL*, Pittsburgh.

Chibout, K., Vilnat, A. (2000). SCALP : un système de compréhension automatique du lexique polysémique d'inspiration linguistique. *Traitement Automatique des Langues Naturelles (TALN 2000)*, Lausanne.

Chomsky, N. (1965). *Aspects of the Theory of Syntax.* MIT Press, Cambridge. http://faculty.georgetown.edu/irvinem/theory/Chomsky-Aspects-excerpt.pdf

Chomsky, N. (1969). *Structures syntaxiques*, Le seuil, Paris.

Chomsky, N. (1971). Deep structure, surface structure and semantic interpretation. *Semantics*, Cambridge University press, 183-216.

Chomsky, N. (1999). On the nature, use, and acquisition of language. W.C. Ritchie and T.K. Bhatia (eds.), *Handbook of Child Language Acquisition*. Cambridge: Cambridge University Press, 33-54.

Combettes, B. (2001). Grammaticalisation de la phrase complexe et évolution de la cataphore. *Langage et référence : mélanges offerts à Kerstin Jonasson*, Université d'Uppsala, 105-114, 2001. ⟨halshs-00004822⟩

Combettes, B. (2007). Évolution des structures thématiques en moyen français. Texte et discours en moyen français : *Actes du XIe Colloque international sur le moyen français*, 35-46.

Connolly, D., Burger, J.-D./ Day, D.-S. (1994). A machine learning approach to anaphoric reference. *Proceedings of International Conference on New Methods in Language Processing*, 255–261.

Corblin, F. (1985). Les chaînes de référence : analyse linguistique et traitement automatique. *Intellectia*, 123-143.

Corblin, F. (1987). *Indéfini, défini et démonstratif*. Genève, Droz.

Corblin, F. (1995). *Les formes de reprise dans le discours*. Rennes, Presses Universitaires de Rennes.

Cornish, F. (1986a). Anaphoric pronouns : Under linguistic control or signalling particular discourse representations?. A contribution to the debate between Peter Bosch, and Liliane Tasmowski and Paul Verluyten, *Journal of semantics, 5(3),* 233-260.

Cornish, F. (1986b). *Anaphoric Relations in English and French : A Discourse Perspective*. London, Croom Helm.

Cornish, F. (1990). Anaphore pragmatique, référence et modèles du discours. Kleiber G. & Tyvaert J. (eds.), *L'anaphore et ses domaines*, Paris, Klincksieck, 81-96.

Cornish, F. (1996). 'Antecedentless' anaphors : deixis, anaphora, or what ? Some evidence from English and French. *Journal of linguistic, 32,* 19-41

Cornish, F. (2009a). Inter-sentential anaphora and coherence relations in discourse : a perfect match. *Language Sciences,* 31 : 572-592.

Cornish, F. (2009b). Le rôle des anaphores dans la mise en place des relations de cohérence dans le discours : l'hypothèse de Hobbs, J.-R. *Journal of French Language Studies* : *Relations de cohérence et fonctionnement des anaphores, 19 (2).* Cornish, F. (ed.): 159-181.

Danlos, L. (2005). ILIMP : outil pour repérer les occurrences du pronom impersonnel il. *Actes de TALN'05*, Dourdan, France, 123-132.

De Beaugrande, R./Dressler, W. (1981). *Introduction to Text Linguistics*. London & New York : Longman.

de Chalendar, G./Grau, B (2000). Généralisation de graphes conceptuels. *Actes du congrès Reconnaissance des Formes et Intelligence Artificielle (RFIA)*, 359-368.

De Chalendar, G./Grau, B. (2000). Généralisation de graphes conceptuels. *Actes 12ème Congrès Reconnaissance des Formes et Intelligence Artificielle (RFIA),* volume II, 359-368, Paris.

De Week, G. (1991). *La cohésion dans les textes d'enfants*. Paris, Delachaux et Niestlé.

Demol, A. (2007). *Les anaphoriques celui-ci et il : étude des facteurs qui déterminent leur choix.* Thèse de Doctorat, Gante : Universidad de Gante.

Demol, A. (2011). Multiplicité des cadres théoriques et des terminologies. A. Demol, *Les pronoms anaphoriques il et celui-ci*, Louvain-la-Neuve, Belgique : De Boeck Supérieur. *Discourse*, London, Academic Press, 15-46.

Désoyer, A., Landragin, F., Tellier, I., Lefeuvre, A./ Antoine, J.-Y. (2014). Les coréférences à l'oral : une expérience d'apprentissage automatique sur le corpus ANCOR. *Traitement Automatique des Langues*, 55(2), 97-121.

Désoyer, A., Landragin, F./ Tellier, I. (2015). Apprentissage automatique d'un modèle de résolution de la coréférence à partir de données orales transcrites du français : le système CROC. *Vingt deuxième Conférence sur le Traitement Automatique des Langues Naturelles*, Jun 2015, Caen, France, 439-445. ffhalshs-01162174f.

Dessus, P., Trausan-Matu, S., Wild, F., Dupré, D., Loiseau, M., Rebedea, T./ Zampa, V. (2011). Un environnement personnel d'apprentissage évaluant des distances épistémiques et dialogiques. *Distances et savoirs*, 9(4), 473-492.

Dimitrov, M., Bontcheva, K., Cunningham, H./ Maynard, D. (2005). A Lightweight Approach to Coreference Resolution for Named Entities in Text. *Anaphora Processing: Linguistic, Cognitive and Computational Modelling*, Branco, A., Cenery, T./ Mitkov, R.-A. (eds.), John Benjamins.

Dubois, J. (1965). *Grammaire structurale du français. Nom et Pronom*. Paris, Larousse.

Ducrot, O./Schaeffer, J.-M. (1995). *Nouveau dictionnaire encyclopédique des sciences du langage*, Paris : Seuil.

Ducrot, O./Todorov T. (1972). *Dictionnaire encyclopédique des sciences du langage*, Paris, Seuil.

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

Dupuy, E. (2006). *La continuité référentielle en moyen français: règles syntactico-sémantiques*. Thèse de Doctorat. Le Mans : Université du Maine.

Dupuy, E. (2013). La cataphore. Approche diachronique et émergence dans la prose du moyen français. *Le Moyen Français, 73*, 49-87.

Duvallon, O. (2007). L'anaphore au sein des "configurations syntactico-discursives ". *Les cahiers de praxématique*, Montpellier : Presses universitaires de la Méditerranée, 163-192. ffhalshs00674834f

Ehlich, K. (1982). Anaphora and deixis : same, similar or different?. R. Jarvella & W. Klein (eds.), *Speech, Place and Action : Studies in Deixis and related Topics*, Chichester, John Wiley, 315-338.

Ehrlich, M.-F. (1994). *Mémoire et compréhension du langage*. Lille : Presses universitaires de Lille.

Ericsson, K.-A./Kintsch, W. (1995). Long-term working memory. *Psychological review*, *102*(2), 211.

Estival, D. (1994). Anne Abeillé (1993), Les nouvelles syntaxes : grammaires d'unification et analyse du français. *Cahiers de praxématique*, (22), 181-185.

Fayol, M. (1985). Analyser et résumer des textes : une revue des études développementales. *Études de Linguistique Appliquée, n° 59*, 54-64.

Fayol, M. (1992). Le résumé : un bilan des recherches de psychologie cognitive. *L'activité résumante* Charolles, M., Petitjean, A. (Eds.), Metz : Université de Metz, 105-124.

Fayol, M. (1997). *Des idées au texte : Psychologie cognitive de la production verbale, orale et écrite*, Paris : Presses Universitaire de France.

Fillmore, C.-J. (1968). The case for case. Bach and Harms (Ed.) *Universals in Linguistics Theory*, NY: Holt, Rinehart and Winston, 1-88.

Flavell, J. H. (1976). Metacognitive aspects of problem solving. Dans Resnick, L. B. (Ed.). *The nature of intelligence*, Hillsdale, NJ : L. Erlbaum, 231-235.

Flavell, J.-H. (1985). Developpement métacognitif. *Psychologie développementale, problèmes et réalités*, 29-41.

Fournier, N. (2008). La gestion des anaphoriques en discours au XVIIe siècle : l'exemple du cardinal de Retz. Bertrand. O., Prévost, S., Charolles, M., François, J./Schnedecker, C., (éds). *Discours, diachronie, stylistique du français* ; *Etudes en hommage à Bernard Combettes*, Peler Lang SA, Editions scientifiques internationales, 325-341, 2008. ffhalshs-00387868f

Frege, G. (1892). Über Sinn und Bedeutung. *Zeitschrift für Philosophie und philosophische Kritik*, 22-50.

Frege, G. (1971). Sens et dénotation. *Ecrits logiques et philosophiques*, Paris, Seuil, 102-126 (traduction de Frege 1892a).

Fries, P.-H. (2004). What makes a text coherent?. *Text and texture*, Banks, D. (éd.), Paris, L'Harmattan.

Fuchs, C. (2009). L'ambiguïté : du fait de langue aux stratégies interlocutives. *L'ambiguïté*. Nanterre, France. 3-16. ffhal-00551367f

Gardelle, L., Rossi, C./Vincent-Durroux, L. (2019). La gestion de l'anaphore en discours : complexités et enjeux. *Cahiers de praxématique* [En ligne], *72 | 2019*, URL : http://journals.openedition.org/praxematique/5368

Gardent, C./Manuélian, H. (2005). Création d'un corpus annoté pour le traitement des descriptions définies. *TAL, 46(1)*, 115–139.

Garside, R. (1997). *Corpus Annotation, Linguistic Information from Computer Text Corpora*. Leech, G. & A. McEnery (eds.), Longman : London.

Gasperin, C., Karamanis, N./ Seal, R. (2007). Annotation of anaphoric relations in biomedical full-text articles using a domain-relevant scheme. *Proceedings of DAARC,* 19–24, Lagos, Portugal.

Gaudinat, A., Goldman, J.-P./ Wehrli, E. (1998). Le système de synthèse FIPSVox: syntaxe, phonétisation et prosodie. *Actes des XXIIe journées d'études sur la parole, Martigny, Suisse,* 139-142.

Ge, N., Hale, J./ Charniak, E. (1998). A statistical approach to anaphora resolution. In *Proceedings of the 6th Workshop on Very Large Corpora*, Montréal, Canada, 161-170.

Gilles, P. (2005). Cohésion et cohérence. Études de linguistique textuelle, sous la direction d'Anna Jaubert, Lyon, ENS-Éditions, *L'Information Grammaticale, N. 112*, 2007. 51-52.

Gillon, B. S. (1990). Ambiguity, generality and indeterminacy : tests and definitions. *Synthese 85*, 391-416.

Gillon, B. S. (2004). Ambiguity, indeterminacy, deixis and vagueness. S. Davis & B . S. Gillon (eds.), *Semantics : a reader*, 157-187. Oxford University Press.

Givón, T. (1992). The grammar of referential coherence as mental processing instructions. *Linguistics*, 30: 5-55.

Givón, T. (1983). *Topic continuity in discourse : a quantitative cross-language study.* Amsterdam : John Benjamins Publishing.

Green Jr, B. F., Wolf, A. K., Chomsky, C./Laughery, K. (1961). Baseball : an automatic question-answerer. *Papers presented at the May 9-11, 1961, western joint IRE-AIEE-ACM computer conference,* 219-224.

Grevisse, M./Goosse, A. (1995). *Nouvelle grammaire française*. 3éme Ed. Louvain-la-Neuve : De Boeck-Duculot.

Gross, G., 1996. *Les expressions figées en français : noms composés et autres locutions*. Edition Ophrys.

Grosz, B.-J., Joshi, A.-K./ Weinstein, S. (1995). Centering : a framework for modeling the local coherence of discourse. *Computational Linguistics, 21, 2,* 203-225. DOI : 10.21236/ADA324949

Guillot, C. (2004). Entre anaphore et deixis : l'anaphore démonstrative à fonction résomptive. David Trotter. Actes du *XXIVe Congrès international de linguistique et de philologie romanes*, Aberystwyth. Vol. 3, M. Niemeyer, 307-315, 2007. ⟨halshs-00324174⟩

Gundel, J.-K., Hedberg, N./Zacharski, R.(1993). Cognitive status and the form of referring expressions in discourse. *Language,* 274-307.

Habert, B., Nazarenko, A./ Salem, A. (1997). Les linguistiques de corpus. *U Linguistique*. Armand Colin/Masson, Paris.

Halliday, M. (1973). *Explorations in the functions of language*. Arnold, Londres.

Halliday, M.A.K./Hasan, R. (1976). *Cohesion in English*. London, Longmans.

Heindrickx, I., Bouma, G., Coppens, F., Daelemans, W., Hoste, V., Kloosterman, G., Mineur ,A.-M., Van Der Vloet, J./Verschelde J.-L. (2008). A coreference corpus and resolution system for Dutch. *Proc, LREC'2008*.

Hidi, S./ Anderson, V. (1986). Producing written summaries: task demands, cognitive operations, and implications for instruction. *Review of Educational Research*, 56(4), 473-493.

Hinrichs, E., Kübler, S., Naumann, K./ H. Zinsmeister. (2005). Recent developments in linguistic annotations of the TüBa-D/Z Treebank. *27th Meeting of the German Linguistic Association*, Köln.

Hobbs, J.-R. (1979). Coherence and coreference. *Cognitive science, 3(1),* 67-90. DOI : 10.1207/s15516709cog0301_4

Hobbs, J.-R. (1990). *Literature and Cognition*. Stanford University : CLSI : 83-114.

Huang Y., (2006). Anaphora, cataphora, exophora, logophoricity. K. Allan (ed.), *Concise Encyclopedia of Semantics,* Amsterdam, Elsevier, 18-25. DOI : 10.1016/B0-08-044854-2/01084-1

Huang, Y. (2000). *Anaphora : a cross-linguistic study*. *Oxford Studies in Typology and Linguistic Theory*, Oxford, Oxford University Press.

Hull, D. A., Grefenstette, G., Schulze, B. M., Gaussier, E., Schütze, H./Pedersen, J. O. (1997). Xerox trec-5 site report : Routing, filtering, nlp, and spanish tracks. *Proceedings of the Fifth Text REtrieval Conference (TREC-5)*, 167–180.

Hutchins, J. (1987). Summarization : Some problems and methods. *Meaning : The frontier of informatics, 9*, K. P. Jones (Ed.), Aslib, Londres, 151-173.

Iida, R., Mamoru, K., Kentaro, I./Yuji, M. (2007). Annotating a Japanese text corpus with predicate-argument and coreference relations. *Proc. Linguistic Annotation Workshop*, Stroudsburg,132-139.

Jackendoff, R. (1990). *Semantic Structures*. The MIT Press, Cambridge Mass.

Jackendoff, R. (2002). *Foundations of Language : Brain, Meaning, Grammar, Evolution*. Oxford : University Press.

Joseph, J. (1989). *Pensées, jugements et notations, anthologie critique établie par Rémy Tessonneau*. José Corti, Paris.

Joubert, J. (1850). Pensées, essais, maximes et correspondance. *Gallica* www.bnf.fr

Kara, M./Wiederspiel, B. (2011). Anaphore résomptive conceptuelle et mémoire discursive : entre identité et altérité. *Itinéraires* [En ligne], *2011-2* . URL : http://journals.openedition.org/itineraires/134 ; DOI : 10.4000/itineraires.134

Karamanis, N. (2003). *Entity Coherence for Descriptive Text Structuring*. Ph.D. thesis, University of Edinburgh, Edinburgh, Scotland.

Kay, M. (1979). Functional grammars. *Annual meeting of the Berkeley linguistic society*, 142-158.

Kayser, D. (1987). Une sémantique qui n'a pas de sens. *Langages, (87)*, 33-45.

Kennedy, C./Boguraev, B. (1996). Anaphora for everyone : Pronominal anaphora resolution without a parser. *Proeedings of the 16th International Conference on Computational Linguistics-Volume (1)*, Association for Computational Linguistics, 113-118.

Kesik, M. (1989). *La cataphore*. Paris : PUF.

Kęsik, M. (2014). Cataphore / anaphore : complémentarité référentielle, sémantique et syntaxique. René Daval; Pierre Frath; Emilia Hilgert; Silvia Palma. *Les théories du sens et de la référence. Hommage à Georges Kleiber*, Éditions et presses universitaires de Reims, Res per nomen, 9782915271805, 567-578. ⟨hal-01864278⟩

Kintsch, W. (1988). The role of knowledge in discourse comprehension : A construction-integration model. *Psychological review*, *95*(2), 163-182.

Kintsch, W. (1992). A cognitive architecture for comprehension. H. L. Pick, Jr., P. W. van den Broek, & D. C. Knill (Eds.), *Cognition : Conceptual and methodological issues*, Washington, DC, US: American Psychological Association, 143-163. http://dx.doi.org/10.1037/10564-006

Kintsch, W. (1993). Information accretion and reduction in text processing: Inferences. *Discourse processes*, *16*(1-2), 193-202.

Kintsch, W. (2000). Metaphor comprehension : A computational theory. *Psychonomic bulletin &review*, *7*(2), 257-266.

Kintsch, W., Patel, V.-L./ Ericsson, K.-A. (1999). The role of long-term working memory in text comprehension. *Psychologia*, *42*(4), 186-198.

Kintsch, W./Van Dijk, T.-A. (1978). Toward a model of text comprehension and production. *Psychological review*, *85*(5), 363.

Kintsch, W./Van Dijk, T.-A. (1983). *Strategies of discourse comprehension*. Orlando, Academic Press.

Kleiber, G. (1988). Peut-on définir une catégorie générale de l'anaphore?. *Vox Romanica*, *47*,1.

Kleiber, G. (1991). Anaphore-déixis. Où en sommes-nous?. *Information grammaticale*, 51 : 3-18.

Kleiber, G. (1992a). L'anaphore : d'un problème à l'autre. *Le français moderne*, 60, I, 1-22.

Kleiber, G. (1992b). Anaphore-deixis : deux approches concurrentes. : M.-A. Morel & L. Danon-Boileau (éds) *La deixis. Colloque en Sorbonne 8-9 juin 1990*. Paris : PUF, 613-626.

Kleiber, G. (1994a). *Anaphores et pronoms*. Louvain-la-Neuve : Duculot.

Kleiber, G. (1994b). Contexte, interprétation et mémoire : approche standard *vs* approche cognitive, *Langue Française,103,* 9-23.

Kleiber, G. (1997). Sens, référence et existence : que faire de l'extra-linguistique?. *Langages*, *31*(127), 9-37.

Kleiber, G. (2001). *L'anaphore associative*, Paris, PUF.

Kleiber, G. (2002). Marqueurs référentiels et théorie du centrage. *Linx*, 47, 107-119. URL : http://journals.openedition.org/linx/588

Kleiber, G. (2006). Démonstratifs : emploi à la mode et mode(s) d'emploi, *Langue Française,* 152, 9-23.

Kleiber, M./Riegel, G. (1989). Une sémantique qui n'a pas de sens n'a vraiment pas de sens. *Lingvisticae Investigationes, 13*(2), 405-417.

Kleiber, G./Vassiliadou, H. (2007). Sur les approches intuitives de la relation d'Elaboration. *Scolia*, *22*, 147-161.

Lamiroy, B. (2003). Maurice Gross (1934-2001). *Travaux de linguistique (1),* 145-158. www.cairn.info/revue-travaux-de-linguistique-2003-1-page-145.htm.

Landragin, F. (2011). Une procédure d'analyse et d'annotation des chaînes de coréférence dans des textes écrits. *Corpus 10*, 61-80.

Landragin, F. (2012). La saillance : questions méthodologiques autour d'une notion multifactorielle. *Faits de langues*, Peter Lang, 2012, 15-31. ffhalshs-00690831f

Landragin, F. (2015). Coreference in the light of pronouns with indefinite reference. *Conference on R-impersonals,* Paris.

Lappin, S./Leass, H.-J. (1994). An algorithm for pronominal anaphora resolution. *Computational linguistics, 20(4),* 535-561.

Le Pesant, D. (2002). La détermination dans les anaphores fidèles et infidèles. *Langages*, (145), 39-59.

Le Priol, F. (2000). *Extraction et capitalisation automatiques de connaissances à partir de documents textuels : Seek-Java : identification et interprétation de relations entre concepts.* Doctoral dissertation, Paris 4.

Lefebvre, A. (1991). La cataphore. *La Linguistique, 27(1),* 161-163. http://www.jstor.org/stable/30248641

Lefeuvre, A., Antoine, J.-Y./Schang, E. (2014). Le corpus ANCOR_Centre et son outil de requêtage : application à l'étude de l'accord en genre et nombre dans les coréférences et anaphores en français parlé. *Congrès Mondial de Linguistique Française – CMLF 2014* SHS Web of Conferences 8

Lefeuvre, F. (2014). *Etude grammaticale du français classique dans les textes*. PU de la Sorbonne Nouvelle, Paris. (halshs-01138852).

Lemaire, B., Bianco, M., Sylvestre, E./ Noveck, I. (2001). Un modèle de compréhension de textes fondé sur l'analyse de la sémantique latente. *La cognition entre individu et société,* 309-320.

Lemaire, B., Dessus, P. (2003). Modèles cognitifs issus de l'Analyse de la sémantique latente. *Cognito-Cahiers Romans de Sciences Cognitives, 1(1),* 55-74.

Léon, J. (2002). Le CNRS et les débuts de la traduction automatique en France. *La Revue pour l'histoire du CNRS, (6),* [En ligne]. URL : http://histoire-cnrs.revues.org/3461

Léon, J. (2015). *Histoire de l'automatisation des sciences du langage*. ENS Éditions.

Loaiciga Sanchez, S. (2013). Résolution d'anaphores et traitement des pronoms en traduction automatique à base de règles. *Traitement Automatique du Langage Naturel 2013* (TALN 2013), 683-690.

Longo, L. (2013) Un corpus pour optimiser l'identification automatique des chaînes de référence. *Cahiers de praxématique* [En ligne], 54-55 | 2010, consulté le 29 juin 2016. URL : http://praxematique.revues.org/1172

Longo, L. (2013). *Vers des moteurs de recherche "intelligents" : un outil de détection automatique de thèmes. Méthode basée sur l'identification automatique des chaînes de référence.* Doctoral dissertation, Université de Strasbourg.

Lundquist, L. (1980). *La cohérence textuelle : syntaxe, sémantique, pragmatique*. Kobenhavn : Nyt Nordisk Forlag, Arnold Busck.

Maâloul, H. (2012). *Approche hybride pour le résumé automatique de textes. Application à la langue arabe*. Thèse pour obtenir le titre de Docteur en Informatique. https://tel.archives-ouvertes.fr/tel-00756111v1/document

Maes, A./Noordman, L.-G.-M. (1995). Demonstrative nominal anaphors : A case of non-identification al markedness. *Linguistics*, *33*, 255–282.

Magri-Mourgues, V. (2015). L'anaphore rhétorique dans le discours politique. L'exemple de N. Sarkozy. *Sémio-Linguist*. Textes Discours.

Maillard, M. (1974). Essai de typologie des substituts diaphoriques (Supports d'une anaphore et/ou d'une cataphore). *Langue française, 21*, 55-71.

Mandin, S., Dessus, P./ Lemaire, B. (2006). Comprendre pour résumer, résumer pour comprendre. P. Dessus & E. Gentaz. *Apprentissages et enseignement : sciences cognitives et éducation*, Dunod, 107-122.

Mann, W.-C./Thompson, S.-A. (1988). Rhetorical Structure Theory : Toward a functional theory of text organization. *Text,* 8 (3) : 243-281. DOI: 10.1515/text.1.1988.8.3.243

Manning, C.-D./Schutze, H. (1999). *Foundations of statistical natural language Processing,* The Massachusettes Institute of Technology Press, Cambridge, Massachusettes.

Masseron, C./Schnedecker,C. (1988). Le mode de désignation des personnages. *Pratiques*, 60(1), 98-123.

Masson, N. (1998). *Méthodes pour une génération variable de résumé automatique : vers un système de réduction de texte*. PhD thesis, Université Paris 11-Orsay.

Mayaffre, D. (2005). Rôle et place du corpus en linguistique. Réflexions introductives. Pascale Vergely. Actes du colloque *JETOU'2005*, Université de Toulouse-Le Mirail, 5-17. 〈hal-00553742〉

Mayaffre, D. (2006). Les corpus politiques : objet, méthode et contenu Introduction. *Corpus* [En ligne]. 4 | 2005a. URL : http:// corpus.revues.org/292

McNamara, D.-S./Kintsch, W. (1996). Learning from texts: Effects of prior knowledge and text coherence. *Discourse processes*, *22*(3), 247-288.

Mehler, J./Dupoux, E. (1987). De la psychologie à la science cognitive. *Le Débat*, 47, 65-87.

Mellet, S. (2003). Corpus et recherches linguistiques. *Corpus*, [En ligne]. 1 | 2002. URL : http://corpus.revues.org/7

Milner , J. C (1982). *Ordres et raisons de langue*. Paris, Gallimard.

Milner , J.C . (1978). *De la syntaxe à l'interprétation*. Paris, Gallimard.

Milner, J.-C. (1976). Réflexions sur la référence. *Langue française 30*, 61-71.

Milner, J.-C. (1989). *Introduction à une science du langage*. Paris, Seuil.

Minel, J.-L. (2002). *Filtrage sémantique de textes. Problèmes, conception et réalisation d'une plate-forme informatique*. Habilitation à diriger des recherches, Université Paris Sorbonne.

Minsky, M. (1974). *A framework for representing knowledge*. MIT, Cambridge Mass.

Mitkov, R. (1994). An integrated model for anaphora resolution. *COLING 1994 Volume 2: The 15th International Conference on Computational Linguistics*.

Mitkov, R. (1997). Factors in anaphora resolution : They are not the only things that matter: a case study based on two different approaches. *Proceedings of a Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts*, Association for Computational Linguistics, 14-21.

Mitkov, R. (1998). Robust pronoun resolution with limited knowledge. *Proceedings of the 17th international conference on Computational linguistics-Volume (2)*, Association for Computational Linguistics, 869-875.

Mitkov, R. (1999). Multilingual anaphora resolution. *Machine Translation, 14*(3-4), 281-299.

Mitkov, R. (2002). *Anaphora Resolution*. Longman

Mitkov, R., Boguraev, B./ Lappin, S. (2001). Introduction to the special issue on computational anaphora resolution. *Computational Linguistics, 27(4),* 473– 477.

Mitkov, R., Evans, R./ Orasan, C. (2002). A new, fully automatic version of mitkov's knowledge-poor pronoun resolution method. *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics*, Springer, Berlin, Heidelberg, 168-186.

Mitkov, R., Evans, R., Orasan, C./ Pekar, V. (2007). Anaphora resolution : To what extent does it help NLP applications?. *Discourse Anaphora and Anaphor Resolution Colloquium*, Springer, Berlin, Heidelberg, 179-190.

Moeschler, J./Reboul, A. (1994). *Dictionnaire encyclopédique de pragmatique*. Paris, Seuil.

Montague, R. (1970). English as a formal language, *Formal philosophy*. Yale University Press.

Mullet, V./Denhière, G. (1997). Accès au lexique et ambiguïtés lexicales nominales: effet de la polarité des homographes et de la nature du contexte. *Sémantique linguistique et psychologie cognitive. Aspects théoriques et empiriques*, 51-74.

Muninn, P. (2001). *Une stratégie d'extraction d'informations dans des corpus spécialisés par application de méthodes d'analyse linguistique de surface et de représentation conceptuelle des structures sémantiques*. Thèse de doctorat en informatique, université de Bourgogne.

Muzerelle, J., Schang, E., Antoine, J., Eshkol I./Maurel, D. (2012). Annotation en relations anaphoriques d'un corpus de discours oral spontané en français. *Congrès Mondial de Linguistique Française, CMLF'2012*, Jul 2013, Lyon, France. 15.

Nazarenko, A. (2004). *Donner accès au contenu des documents textuels. Acquisition de connaissances et analyse de corpus spécialisés*. Habilitation à Diriger des Recherches, Université Paris-Nord.

Nedoluzzhko, A., Mírovský, J., Ocelák, R./J. Pergler. (2009). Extended coreference relations and bridging anaphora in the Prague Dependancy Treebank. *Proc.DAARC'2009*, 1-16. Chennai Goa, Indica.

Neveu, F. (2004). *Dictionnaire des sciences du langage*. Paris : Colin.

Newell, A./Simon H.-A. (1956). The Logic Theory Machine : A complex information processing system. *Transactions on Information Theory*, 61-79.

Nouioua, F. (2007). Heuristique pour la résolution d'anaphores dans les textes d'accidents de la route. *Actes de la Journée d'étude de l'Association pour le Traitement Automatique des Langues (ATALA) sur La résolution des anaphores en Traitement Automatique des Langues.*

Ogrodniczuk, M., Kopeć, M., Głowinska, K., Savary, A./ M. Zawisławska. (2013). Polish coreference corpus, *LTC'2013*.

Paquay, L./Lauwaers, A. (1992). Résumer un texte : les procédures prescrites dans les ouvrages méthodologiques sont-elles appliquées ? sont-elles applicables ? Charolles, M./Petitjean, A. (Eds.). *L'activité résumante*, Metz : Université de Metz, 159-181.

Pepin, L. (2009). La coréférence dans la narration, première partie. *Semiotica*, 4, 335-363.

Perdicoyanni-Paléologou, H. (2001). Le concept d'anaphore, de cataphore et de déixis en linguistique française. *Revue québécoise de linguistique*, *29(2)*, 55–77. doi:10.7202/039441ar

Pereira, F.-C./Warren, D.-H. (1980). Definite clause grammars for language analysis—a survey of the formalism and a comparison with augmented transition networks. *Artificial intelligence*, *13*(3), 231-278.

Perret, M. (2000). Quelques remarques sur l'anaphore nominale aux XIVe et XVe siècles. *L'Information grammaticale*, 87(1), 17-23.

Petöfi, J.-S. (1973). Towards an empirically motivated grammatical theory of verbal texts. *Studies in Text Grammar*, Petöfi, J.-S., Rieser, H (eds), Dordrecht, Reidel, 205-275.

Petrick, S.-R. (1973). Transformational analysis. Natural language processing, Rustin, R. (ed.) Academic press, New York, 27-41.

Pironneau, M., Brunelle, E./Charest, S. (2014). Pronoun anaphora resolution for automatic correction of grammatical errors (correction automatique par résolution d'anaphores pronominales) [in french]. *Proceedings of TALN 2014*, 113–124, Marseille, France.

Poesio, M. (2003). Associative descriptions and salience. *Proc. of the EACL Workshop on Computational Treatments of Anaphora*, Budapest.

Poesio, M. (2004). Discourse Annotation and Semantic Annotation in the GNOME Corpus. http://www.aclweb.org/anthology/W04-0210

Poesio, M./Vieira, R. (1998). A corpus-based investigation of definite description use. *Computational Linguistics*, 24(2) June:183–216.

Poibeau, T. (2014). Le traitement automatique des langues pour les sciences sociales. *Réseaux*, (6), 25-51.

Popescu-Belis, A., Robba, I./Sabah, G. (1998). Reference resolution beyond coreference : a conceptual frame and its application. *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 2*, Association for Computational Linguistics, 1046-1052.

Popescu-Belis, A. (1999). *Modélisation multi-agent des échanges langagiers : application au problème de la référence et son évaluation*. Thèse d'université, Université de Paris XI (Paris-Sud).

Porge, E. (2003). Nommer quoi ? À propos de la nomination dans la passe. *Essaim, 11, (1),* 39-56. DOI:10.3917/ess.011.0039.

Pradhan, S., Ramshaw, L., Weischedel, R., MacBride, J./Micciula, L. (2007). Unrestricted coreference : identifying entities and events in On to Notes. *Proc. Int. Conf. on Semantic Computing (ICSC'07)*. Washington, DC. USA. IEEE., 446-453.

Prandi, M. (2007). Les fondements méthodologiques d'une grammaire descriptive de l'italien. *Langages*, 167, (3), 70-84. DOI:10.3917/lang.167.0070.

Rastier, F. (1998). Le problème épistémologique du contexte et le statut de l'interprétation dans les sciences du langage. *Langages 129*, 97-111.

Rastier, F. (2001). *Arts et sciences du texte,* Paris, Presses universitaires de France.

Rastier, F. (2005). Enjeux épistémologiques de la linguistique de corpus. *La linguistique de corpus*. Williams, G. (éd.), Rennes : Presses Universitaires de Rennes, 31-46. http://www.revue-texto.net/ Inedits /Rastier/ Rastier_Enjeux.html

Rastier, F., Cavazza, M./ Abeill, A. (1994). Sémantique pour l'analyse : de la linguistique l'informatique. *Sciences cognitives*.

Reboul, A. (1989). Résolution de l'anaphore pronominale : sémantique et/ou pragmatique. *Cahiers de Linguistique Française 10*, 77-100.

Reboul, A. (1990). Rhétorique de l'anaphore. G. Kleiber & J.-E. Tyvaert (éds) *L'anaphore et ses domaines*. Paris : Klincksieck, 279-300.

Reboul, A. (1997). Cohérences et anaphores : Mythes et réalités. *Relations anaphoriques et cohérences*, Rodopi, 297-314.

Recanati, F. (2007). *Sens littéral : langage, contexte, contenu*. Edition de l'éclat.

Recasens, M. (2010). *Coreference : Theory, Annotation, Resolution and Evaluation*. Mémoire de doctorat de l'Université de Barcelone, Espagne.

Recasens, M. Martí, A./Taule, M. (2009). First mention definite : More than exceptional cases. *The Fruits of Empirical Linguistics*, 2:217.

Reichler-Béguelin, M.-J. (1988). Anaphore, cataphore et mémoire discursive. *Pratiques*, 57(1), 15-43.

Reichler-Béguelin, M.-J. (1989). Anaphores, connecteurs, et processus inférentiels. Modèles du discours. *Recherches actuelles en Suisse Romande*, Rubattel, C. (éd.), Berne : P. Lang, 302-336.

Reinhart, T. (1980). Conditions for text coherence. *Poetics today, 1(4),* 161-180.

Rialle, V. (1996). L'intelligence artificielle et sa place dans les sciences de la cognition. *Bulletin de l'Association Française de l'Intelligence Artificielle. n° 26,* 8-12.

Ricoeur, P. (2006). Mémoire, histoire, oubli. *Esprit, (3),* 20-29.

Rondelli, F. (2008). La cohérence textuelle : rapport à la langue, à soi, à l'autre. *Actes des Xèmes Rencontres Jeunes Chercheurs* de l'école doctorale 268 Langages et Langues, Paris : ILPGA.

Rosenbaum, P.-S. (1965). *A grammar English Predicate Complement Constructions*. Doctoral dissertation, Massachusetts Institute of Technology.

Roux, L. (1987). Introduction aux Actes de l'Atelier sur l'Anaphore. *Atelier de linguistique*, SAES Bordeaux, : 7-8.

Sabah, G. (1988). *L'intelligence artificielle et le langage : représentation des connaissances*. Hermès, Paris.

Sabah, G. (1996). Le sens dans les traitements automatiques des langues-le point après 50 ans de recherches. *Actes de journée ATALA (un demi-siècle de traitement automatique des langues : état de l'art),* 46. URL : http://users.dcc.uchile.cl/~abassi/WWW/Lengua/Sabah97.html

Sabah, G. (2004). Intelligence artificielle, linguistique et cognition. *La linguistique cognitive* [en ligne]. Paris : Éditions de la Maison des sciences de l'homme.

Sabah, G./Grau, B. (2000). Compréhension automatique de textes. *Ingéniérie des langues, Informatique et systèmes d'information*, 293-310.

Sabah, G./Rady, M. (1983). A Deterministic Syntactic-Semantic Parser. *IJCAI* , 707-709.

Sager, N., Friedman, C./ Lyman, M. (1987). *Medical Language Processing: Computer Management of Narrative Data*, Addison-Wesley, Reading, MA.

Saldanha, G. (2009). Principles of corpus linguistics and their application to translation studies research. https://www.academia.edu/11969887/Principles_of_corpus_linguistics_and_their_application_to_translation_studies_research

Salkoff, M. (1973). *Une Grammaire de chaîne du Français, analyse distributionnelle.* Paris, 1973.

Salles, M. (2010). Anaphore associative et relations de cohérence : une expression particulière de la relation *Assertion-Indice. Discours. Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics*, (6). [En ligne]. URL : http://discours.revues.org/7739 ; DOI : 10.4000/discours.7739

Salmon-Alt, S. (2001). *Référence et Dialogue finalisé : de la linguistique à un modèle opérationnel*. Thèse de Doctorat, Université H. Poincaré, Nancy.

Salmon-Alt, S. (2002). Le projet ANANAS : Annotation Anaphorique pour l'Analyse Sémantique de Corpus. *Actes de TALN*, Nancy, 24-27 juin, 163-172.

Salmon-Alt, S. (2004). Résolution automatique d'anaphores infidèles en français : Quelles ressources pour quels apports ?. *Actes de TALN*, session poster, Fès, 19-21 avril.

Salmon-Alt, S., Bick, E., Romary, L./Pierrel, J.-M. (2004). La FREEBANK : vers une base libre de corpus annotés. *Actes de TALN*, Fès, 19-21 avril.

Saussure, F. (1916). *Cours de linguistique générale*.

Schank, R.-C. (1975). The structure of episodes in memory. *Representation and understanding*, Morgan Kaufmann, 237-272.

Schnedecker, C. (2015a). Contraintes pesant sur les anaphores à nom général dans les chaînes de référence renvoyant à des entités humaines. *Travaux de linguistique, 70* (1), 39–72.

Schnedecker, C. (2015b). Un problème à la croisée des disciplines linguistiques : Les noms d'humains comme interface entre morphologie, syntaxe et sémantique. *La sémantique et ses interfaces*, Actes du colloque 2013 de l'ASL, Association des sciences du langage, Rabatel, A. et al. (éds.), Limoges : Lambert-Lucas, 111–141.

Schnedecker, C. (1997). Nom propre et chaînes de référence. *Recherches Linguistiques, 21*, Paris : Klincksieck.

Schnedecker, C./Charolles, M. (1993). Coréférence et identité. Le problème des référents évolutifs. *Langages*, 106-126.

Shannon, C.-E. (1940). *An Algebra for Theoretical Genetics*. MIT Ph.D. thesis, Department of Mathematics,. MIT Institute Archives.

Smith, C.-S. (2003). Modes of Discourse. *The local structure of texts*, Cambridge: Cambridge University Press.

Sparck, K. (1998). Automatic summarizing : Factors and directions. *Automatic text summarisation*, Mani, I., Maybury, M. (Eds.), MIT Press, Cambridge, 1-12.

Tesnière, L. (1965). *Eléments de syntaxe structurale*.

TLFi, 9ème édition : http://atilf.atilf.fr/tlf.htm

Tognini Bonelli, E. (2001). *Corpus Linguistics at Work*. Amsterdam : John Benjamins.

Tutin A., (2002). A corpus-based study of pronominal anaphoric expressions in French. *Proceedings of DAARC 2002 (Discourse Anaphora and Anaphora Resolution)*, Lisbon.

Van Dijk, T.-A. (1972). *Some Aspects of Text Grammars. A Study in Theoretical Linguistics and Poetics*. The Hague, Mouton.

Van Dijk, T.-A. (1977). *Text and Context. Explorations in the Semantics and Pragmatics of Discourse*. London, Longman.

Van Dijk, T.-A./Kintsch, W., (1983). *Strategies of discourse comprehension*. New York: Academic Press.

Vargas, C. (1992). *Grammaire pour enseigner. Une nouvelle approche théorique et didactique*. Paris : Armand Colin.

Vendler, Z. (1957). Verbs and times. *The philosophical review*, *66*(2), 143-160.

Verschelde, J.-L. (2008). A coreference corpus and resolution system for Dutch. *Proc. LREC'2008*.

Versley et *al.* (2008) BART: A Modular Toolkit for Coreference Resolution. Proceedings *of the ACL-08*, Columbus, June 2008, Association for Computational Linguistics, 9-12.

Voutilainen, A., Heikkila, J./ Antilla, A. (1992). A constraint grammar of English: A performance oriented approach. *University of Helsinki, Department of General Linguistics, Publication*, (21), 11-13.

Walker, J.-P./Walker, M.-I. (1998). *Centering theory in discourse*. Oxford University Press.

Walker, M.-A. (2000). Vers un modèle de l'intégration du centrage avec la structure globale du discours. *Verbum*, t. XXII, n° 1, 31-58.

Ward, W. (1990). The CMU air travel information service : understanding spontaneous speech. Dans les actes de *Human Language Technology, workshop on Speech and Natural Language*, Morristown, NJ, USA, 127–129. Association for Computational Linguistics.

Weaver, W. (1955). Translation. *Machine translation of languages*, Technology press of MIT, New York, 15-23.

Wehrli, E. (1997). *L'analyse syntaxique des langues naturelles : problèmes et méthodes*. Paris, Masson.

Wehrli, E./Nerima, L. (2009). L'analyseur syntaxique Fips. *Proceedings of the IWPT 2009 ATALA Workshop : What French parsing systems*.

Weissenbacher, D. (2008). Influence des annotations imparfaites sur les systèmes de Traitement Automatique des Langues, un cadre applicatif : la résolution de l'anaphore pronominale. *Informatique et langage*, Université Paris-Nord - Paris XIII.

Weizenbaum, J. (1966). ELIZA---a computer program for the study of natural language communication between man and machine. *CACM, 9*, 26-45.

Wiederspiel, B. (1989). Sur l'anaphore : du modèle "standard" au modèle "mémoriel". *Travaux de linguistique et de philologie 27*, 95-113.

Winograd, T. (1972). *Understanding natural language*, Academic press, Edinburgh.

Woods, W.-A. (1970). Transition network grammars for natural language analysis. *Communications of the Association for Computing Machines*, 13(10), 591–606.

Woods, W.-A. (1975). What's in a Link : Foundations for Semantic Networks. *Representation and understanding*, Morgan Kaufmann, 35-82.

Woods, W.-A., Bates, M., Brown, G., Cook, C.-C./ Bruce, B.-C. (1976). Speech understanding systems, final technical progress report, volumes iv. Rapport *technique*, Cambridge, MA.

Yule, G. (1982). Interpreting anaphora without identifying reference, *Journal of Semantics,1*, 4, 315-322.

Yvon, F. (2007). Une petite introduction au Traitement Automatique de la Langue. Notes introductives d'un cours sur le traitement des langues naturelles.

Zampa, V. (2005). Utilisation de l'analyse sémantique latente pour tenter d'optimiser l'acquisition par exposition à une langue étrangère de spécialité. *Alsic. Apprentissage des Langues et Systèmes d'Information et de Communication, 8*(2).

Zribi-Hertz, A. (1996). *L'anaphore et les pronoms*. Paris : Presses Universitaires du Septentrion.