Résolution des anaphores nominales pour la compréhension automatique des textes

Thi Nhung Pham

UNIVERSITÉ SORBONNE PARIS CITÉ
UNIVERSITÉ PARIS NORD – PARIS XIII

ÉCOLE DOCTORALE ERASME (ED 493)

Laboratoire Lexiques - Dictionnaires - Informatique (LDI)

THÈSE DE DOCTORAT

Discipline : Sciences du langage

Spécialité : Traitement Automatique de la Langue (TAL)

PHAM Thi Nhung

RÉSOLUTION DES ANAPHORES NOMINALES POUR LA COMPRÉHENSION AUTOMATIQUE DES TEXTES

Thèse dirigée par Dr Pierre-André BUVET

Soutenue le 27 janvier 2017

JURY

Pr.	Elizabete Aparecida	MARQUES	Université Fédérale du Mato Grosso do Sul (Brésil)	
Pr.	Xavier	BLANCO	Université Autonome de Barcelone	
Dr.	Pierre-André	BUVET	Université Paris 13	Directeur de thèse
Dr.	Iris	ESHKOL	Université d'Orléans	Rapporteur
Pr.	Salah	MEJRI	Université Paris 13	Président du jury
Dr.	Jean-Manuel	TORRES	Université d'Avignon	Rapporteur

Table des matières

REMERCIEMENTS

Je tiens à remercier tout d'abord M. Pierre-André Buvet, mon directeur de thèse pour son encadrement, ses remarques avisées, ses critiques constructives et enrichissantes ainsi que pour sa gentillesse et la disponibilité qu'il a eues à mon égard. Sans son aide, ce travail n'aurait pas été possible.

Je voudrais remercier chaleureusement à l'ensemble du jury de thèse pour avoir bien voulu examiner mes travaux.

Mes remerciements vont aussi à tous mes camarades du LDI qui ont contribué au succès de ce travail. Un grand merci à Belem, Sunock pour leurs conseils, et leur soutien dans mes périodes difficiles.

Un grand merci à Marguerite et Françoise pour leur relecture détaillée de ma thèse qui m'ont permis d'améliorer la rédaction de ma thèse.

En fin, j'adresse particulièrement mes remerciements à ma famille et mes amis pour leur soutien et leurs encouragements.

TABLE DES MATIERES (DETAILS)

TABLE DES ILLUSTRATIONS

LISTE DES ABREVIATIONS

ADJ	:	Adjectif
AEF	:	Automates à états finis
GN	:	Groupe nominal /Groupes nominaux
MUC	:	Message Understanding Conferences
SV	:	Syntagmes verbaux
TAL	:	Traitement automatique de la langue
LDI	:	Lexiques – dictionnaires – Informatique
DET	:	Déterminant
NLTK	:	Natural Language Toolkit
XML	:	Extensible Markup Language
UML	:	Unified Modeling Language
FTB	:	French Treebank
WOLF	:	Wordnet Libre du Français

LISTE DES TABLEAUX

RESUME

Toutes les informations présentes actuellement sur le web représentent une source d'informations colossale, qui s'enrichit de jour en jour. L'analyse automatique de ces informations, qui sont plus souvent non-structurées, constitue un véritable enjeu économique et scientifique. La résolution des anaphores nominales s'inscrit dans la structuration des informations grâce à l'identification du lien entre des groupes nominaux, elle permet de simplifier des tâches à différentes applications : la traduction automatique, le résumé ou l'extraction automatique d'information, le data mining etc.

Les objectifs de cette thèse concernent la conception d'un système de résolution des anaphores nominales pour le traitement des textes en langue française. La réalisation de ce système vise à tester la validité des concepts utilisés dans les analyses linguistiques théoriques; tester les méthodes et les algorithmes existants; et mettre en disponibilité le système pour les recherches scientifiques ou pour l'enseignement. Le travail que nous avons mené dans cette thèse évoque différentes méthodes de résolution des anaphores nominales de deux types : infidèles et associatives.

En nous fondant sur divers aspects autours de la notion d'anaphore nominale et des notions de voisinage comme la résolution d'anaphores pronominales, la résolution de coréférences ; en combinant des méthodes existantes avec des outils et des ressources disponibles pour la langue française, notre travail s'attache à trois modules : module de prétraitement du corpus, module de résolution des anaphores nominales et le module d'évaluation.

Au module de prétraitement, les ressources lexicales sont constituées et mobilisées grâces aux analyses au niveau linguistique des anaphores nominales. La plateforme Unitex est le principal outil utilisé à cette étape. Pour les anaphores du type infidèle, nous avons utilisé deux méthodes différentes : la première mobilise des ressources

lexicales simples avec les entrées de groupes nominaux uniquement ; la deuxième mobilise des ressources plus élaborées (les entrées de groupes nominaux et verbaux). Pour les anaphores associatives du type méronymique, nous nous fondons sur la théorie des classes d'objets afin de décrire le type de relation anaphorique établie entre l'expression anaphorique et son antécédent. Les ressources utilisées pour ce type d'anaphore sont ainsi divisées hiérarchiquement selon les classes et les domaines.

Le module de résolution est l'étape de décision, nous nous basons sur le calcul du poids de saillance de chacun des antécédents potentiels pour sélectionner le meilleur candidat. Chaque candidat peut avoir différents facteurs de saillance, qui correspond à sa probabilité d'être sélectionné. Le poids de saillance final est calculé par le moyen pondéré des poids de saillance élémentaires. Les facteurs de saillances sont proposés après les analyses syntaxiques et sémantiques du corpus.

L'évaluation de notre travail constitue un vrai enjeu à cause de la complexité de la tâche, mais elle nous permet d'avoir une vue globale sur nos méthodes de travail. La comparaison des résultats obtenus permet de visualiser l'apport de chaque paramètre utilisé. L'évaluation de notre travail nous permet également de voir les erreurs au niveau du prétraitement (l'extraction des syntagmes nominaux, des syntagmes verbaux…), cela nous a permis d'intégrer un module de correction dans notre système.

INTRODUCTION

Le domaine de traitement automatique du langage (TALN) ou est une discipline jeune dont les premiers travaux datent du milieu des années 60. Avec le développement exponentiel des techniques de télécommunications, associé aux avancées en informatique et à l'essor de la linguistique computationnelle, le TALN a suivi une courbe de progression impressionnante.

Les chercheurs en TALN partaient principalement du développement des systèmes de traduction automatique et focalisaient leurs travaux surtout sur les aspects syntaxiques de la langue sans véritablement intégrer l'aspect informationnel des textes. Par la suite, dès la fin des années 80, les chercheurs se sont intéressé à l'extraction de l'information c'est-à-dire la récupération des informations utiles dans les textes écrits en langue naturelle. Dans le cadre de ce processus, la résolution des anaphores a été étudiée.

L'importance de la résolution automatique des anaphores a conduit à l'émergence de travaux qui ont fait l'objet de multiples campagnes internationales d'évaluation. Plusieurs approches en TALN ont été proposées afin de résoudre les problèmes de résolution automatique des anaphores, parmi lesquelles deux principales approches sont : les techniques d'apprentissage automatique à partir de corpus, et les analyses fondées sur les principes linguistiques. Pourtant, ces recherches portaient principalement sur la résolution des anaphores pronominales. Les travaux sur les anaphores nominales sont encore rares, même de nos jours, compte tenu de la complexité qu'elles impliquent.

Nous nous intéressons dans cette thèse à la résolution automatique des anaphores nominales. Nous proposons un système de reconnaissance automatique des antécédents pour les expressions anaphoriques, autrement dit, le système appariera une anaphore et sa ou ses sources. Notre étude vise à traiter deux catégories

d'anaphores : les anaphores infidèles et les anaphores associatives méronymiques. Cette résolution est fondée principalement sur des analyses linguistiques et sur l'exploitation de corpus, en particulier les ressources lexicales développées au laboratoire LDI selon la théorie de trois fonctions primaires et des classes d'objet.

En premier lieu, nous dressons une vue globale sur l'objet étudié. Il a semblé opportun de s'interroger tout d'abord sur la notion de l'anaphore : sa définition, sa nature, son histoire… Viennent par la suite les problématiques propres aux anaphores nominales qui motivent les objectifs de notre recherche avant d'aborder un état de l'art pour faire le point sur les études et des pratiques actuelles liées à la résolution automatique des anaphores nominales.

Ensuite, nous présentons la méthodologie pour repérer, étiqueter les anaphores nominales, notamment les trois types privilégiés d'anaphores, en nous appuyant sur les automates à états finis (UNITEX) et les ressources existantes avant de décrire le fonctionnement de notre système de résolution des anaphores nominales.

Enfin, nous proposons une méthode pour évaluer la performance de notre système, des pistes d'amélioration et des perspectives en passant par des analyses des erreurs fréquentes du système. Les résultats obtenus apporteront la réponse aux trois questionnements :

- Comment exploiter l'étiquetage anaphorique pour l'interprétation des textes ?
- Quel est l'apport des ressources linguistiques pour le traitement automatique ?
- Quelle analyse rétroactive sur la théorie permet la mise en place de l'outil informatique ?

PARTIE 1. MOTIVATIONS ET OBJECTIFS

Chapitre 1 : Objet d'étude

Chapitre 2 : Etat de l'art

Chapitre 3 : Méthodologie

CHAPITRE 1. OBJET D'ETUDE

Dans ce chapitre, nous apportons des précisions sur la notion d'anaphore : ses définitions, sa nature, son histoire, ainsi que les manières de distinguer cette notion avec d'autres notions. Par la suite, nous mentionnons, d'une façon globale, les enjeux de notre sujet de recherche dans le contexte actuel du TALN qui suscite la nécessité du traitement des anaphores nominales, avant d'exposer les objectifs de notre thèse.

1. QU'EST-CE QU'UNE ANAPHORE NOMINALE

La langue naturelle offre une grande variété d'expressions qui permettent d'éviter de répéter toujours les mêmes termes pour désigner ce dont on parle mais qui posent problème au niveau du codage informatique. Une même entité, un même événement, une même dénotation peuvent être désignée par différents signifiants coréférents. L'anaphore désigne une reprise intraphrastique ou interphrastique de l'information véhiculée par un mot, un groupe de mots ou un segment de texte antérieurement mentionné. On appelle « anaphorique » le mot ou le groupe de mots qui concrétise cette reprise (Dispy, 2014).

Les deux types d'anaphores les plus fréquemment traités dans la littérature sont les anaphores pronominales, c'est-à-dire les anaphores déclenchées par un pronom personnel et les anaphores nominales (ou anaphores définies) - déclenchées par un groupe nominal défini.

Exemple :

> ***Une voiture*** *est garée dans mon jardin, mon mari vient de* ***l'****acheter.*
>
> ***Une Ferrari*** *est garée dans mon jardin. C****ette voiture*** *était mon rêve.*

L'expression « ***cette Ferrari*** » est appelée antécédent, car elle est reprise sous une autre forme : le pronom « **l'** » dans la première phrase qu'on l'appellera anaphore pronominale, et le groupe nominal « **une voiture** » dans la deuxième phrase, on appellera anaphore nominale.

Les phénomènes d'anaphore peuvent s'exprimer de différentes manières, soit en fonction du type d'antécédent ou de la position de l'antécédent (intra-phrastique ou inter-phrastique), soit en fonction du type ou de la position de la reprise. Le déroulement de ces phénomènes ne se fait pas de la même façon, nous allons l'aborder plus clairement dans cette partie.

1.1. DEFINITION

Le mot « anaphore » est présenté dans les dictionnaires comme un mot latin issu du grec, formé du préfix *ana* qui signifie « en haut », et du verbe *pherein* signifiant « porter ». Il a été réemprunté en grammaire vers le XXe siècle pour désigner la reprise du signifié d'un mot par le moyen d'un autre signe (pronom pour le cas des anaphores pronominales, nom pour le cas des anaphores nominales…) (Rey, 2010)

Les définitions de l'anaphore varient en fonction du point de vue adopté. Il convient tout d'abord que nous distinguions la notion d'anaphore, en rhétorique et en linguistique. En rhétorique, l'anaphore est la répétition d'un mot en tête de plusieurs éléments d'une phrase, pour obtenir un effet de renforcement ou de symétrie. En linguistique, l'anaphore est la reprise d'un segment de discours (antécédent) par un mot anaphorique. – *Le Nouveau Petit Robert de la langue française, 2007.*

1.1.1. ANAPHORE EN RHÉTORIQUE

En rhétorique, l'anaphore est utilisée comme un outil coordinatif de remplacement, un mouvement dans le texte consistant en la répétition d'un mot, d'un segment, d'une expression ou d'un groupe de mots, afin d'obtenir un effet d'ornement du discours, de renforcement ou de symétrie ; d' obtenir un effet d'écho ou d'insistance ; de souligner une juxtaposition, une obsession ; de créer des accumulations analogiques ou disparates ; de produire des parallélismes et de rythmer les phrases (Fontanier, 1977).

Par exemple, l'anaphore est souvent utilisée dans un discours politique (Magri-Mourgues, 2015).

On a capitulé devant l'idéologie de mai 68.

On a capitulé devant la logique de l'assistance.

On a capitulé devant l'immigration non maîtrisée.

On a capitulé devant le communautarisme.

On a capitulé devant une conception formelle et dogmatique de l'égalité. (Meeting de Charleville Mézières, 18 décembre 2006).

Les anaphores jouent un rôle dans le renforcement d'une idée ou dans l'acquisition de l'harmonie en poésie, elles apportent un effet mélodique et rythmique aux vers et elles mettent en valeur les idées principales. L'anaphore est utilisée dans la poésie comme une figure de style pour favoriser l'expressivité et la structure harmonieuse de l'énoncé. Elle renforce l'expression d'un sentiment : indignation, regret, enthousiasme et elle produit un effet d'insistance. La répétition sémantique se présente tantôt avec l'appui des synonymies, c'est-à-dire des termes ayant le même sens, tantôt avec l'appui de termes juxtaposés suivant une intensité sémantique croissante. On distingue aussi l'anaphore de la ***symploque*** consistant à commencer plusieurs phrases ou vers par un même mot et de l'***épiphore*** consistant à les terminer par le même mot (https://fr.wikipedia.org/wiki/Symploque)

***Qui** est l'auteur de cette loi ? **Rullus**.*

***Qui** a privé du suffrage la plus grande partie du peuple romain ? **Rullus**.*

***Qui** a présidé les comices ? **Rullus**. » (Le Grand Larousse du XX^{e} siècle)*

Selon Bonhomme (Bonhomme, 1998), les anaphores obtiennent non seulement une fonction esthétique visant à plaire au public, mais elles sont aussi utilisées comme des moyens argumentatifs. Dans leur rôle argumentatif, elles servent de « persuasion » tout en construisant l'amplification d'une même pensée, en fortifiant une pensée ou une idée particulière dans le message sans devoir nécessairement utiliser sans cesse le même terme.

1.1.2. ANAPHORE EN LINGUISTIQUE

En linguistique, l'anaphore désigne une reprise d'un segment de discours, posé antérieurement dans le texte, qu'on appelle antécédent, par un mot anaphorique. « *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 and Todorov, 1972)

L'anaphore est *un phénomène de dépendance interprétative de deux unités, dont la première, à laquelle se reporte la seconde, l'anaphorique, est appelée «interprétant»* (Ducrot and Todorov, 1972) *«antécédent», «contrôleur de l'anaphorique», ou encore «source sémantique»* (Perdicoyanni-Paléologou, 2001)

Si l'anaphorique et son antécédent sont situés dans la même phrase, on l'appelle une anaphore intraphrastique, au contraire, si l'antécédent se trouve dans une phrase différente de l'anaphore, on l'appelle une anaphore transphrastique (Dispy, 2014)

Par exemple :

***Une voiture** est garée dans mon jardin, c'est mon mari qui vient de **l'**acheter*

C'est une anaphore intraphrastique car l'antécédent et l'expression anaphorique se trouvent dans la même phrase.

***Une Ferrari** est garée dans mon jardin. C**ette voiture** était mon rêve*

C'est une anaphore interphrastique car l'antécédent et l'expression anaphorique se trouvent dans deux phrases différentes.

L'anaphore a été définie traditionnellement comme la « *reprise d'un élément antécédent dans un texte* » (Riegel et al., 1994). Les expressions anaphoriques ne sont pas autonomes car leur interprétation dépend d'une autre expression qui se trouve dans

le texte. Si la source sémantique se situe en aval de l'expression qui renvoie, on parle de ***cataphore***.

Par exemple :

Elle était horrible, cette méchante sorcière

Elle est une cataphore dont la source sémantique est *cette méchante sorcière*

Le ***diaphore*** est une notion qui englobe l'anaphore et la cataphore (Maillard, 1974).

(Apothéloz, 1995a) a abordé les notions d'***endophore*** pour parler de l'expression qui complète son sens dans le texte et d'***exophore*** pour parler de l'expression qui complète son sens par la situation extralinguistique.

Par exemple :

Ça s'est bien passé

Nous devons nous fonder sur le contexte pour savoir à quoi le pronom *ça* se réfère. Pour cette propriété, l'exophore concerne notamment les anaphores pronominales.

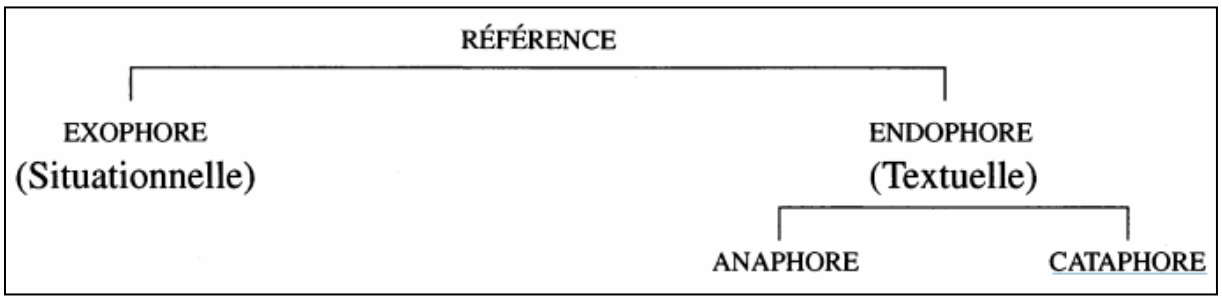


Figure 1 : Hiérarchie des types de références (Perdicoyanni-Paléologou 2001)

L'anaphore est définie de manière plus imagée par Milner (Milner, 1982): *il y a relation d'anaphore entre deux unités A et B, quand l'interprétation de B dépend de A, c'est-à-dire B n'est interprétable que si B reprend entièrement ou partiellement A, alors B est l'élément anaphorique, et A l'antécédent.*

Par exemple :

*J'ai oublié **mon livre de français** ! Peux-tu me prêter **le tien** ?*

L'expression anaphorique *le tien* est interprétable seulement s'il va avec l'antécédent *mon livre de français.*

Lorraine Pepin (2009) définit l'anaphore comme 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, par exemple :

*Jean a trouvé **une très belle théière** dans une boutique la semaine dernière, il voulait **l'**acheter mais il n'avait pas assez d'argent avec lui. Le lendemain, il est retourné à la boutique pour acheter **cet objet** tant convoité. Pourtant, une fois à la maison, Jean l'a mis sur une assiette et il a cassé **ce précieux objet**.*

Dans cet exemple, toutes les expressions anaphoriques *l', cet objet, ce précieux objet* se réfèrent à la même chose : *une très belle théière,* toutes les phrases du texte sont ainsi cohérentes même si on n'a pas repris sous la même forme.

1.1.3. DÉFINITION DE L'ANAPHORE NOMINALE

Pour (Mitkov, 1999), il y a anaphore nominale lorsque l'antécédent d'une anaphore est un groupe nominal défini qui représente le même concept (répétition) ou un concept sémantiquement proche (synonyme).

*Jean a acheté **un sapin** de Noel. **Ce sapin** est vraiment joli.* (Le mot *sapin* est répété)

*Jean a acheté **un sapin** de Noel. **Cet arbre** est vraiment joli.* (Le mot *sapin* est repris par un synonyme *arbre* qui lui est proche sémantiquement)

Mitkov s'est fondé uniquement sur la nature de l'antécédent pour définir l'anaphore nominale : *On parle d'anaphore nominale lorsque l'expression anaphorique, qui peut être un pronom, un groupe nominal défini ou un nom propre, dispose d'un groupe nominal comme antécédent* (Mitkov, 2003).

La résolution des anaphores nominales correspond à deux tâches : d'une part la mise en correspondance de groupes nominaux en prenant en compte leurs variations possibles et d'autre part la détermination des expressions anaphoriques nominales et de leurs antécédents.

1.2. TYPOLOGIE DES ANAPHORES

1.2.1. ANAPHORE & CORÉFÉRENCE

Avant de faire une synthèse sur la typologie des anaphores, nous proposons de faire une distinction entre une anaphore et une coréférence.

Le terme d'anaphore exprime un procédé linguistique consistant à employer une expression rappelant une autre entité précédemment introduite dans le discours (ou éventuellement dans un autre document) mais il n'y a pas nécessairement de relation d'identité entre ces entités. Dans la **coréférence,** ces entités pointent vers le même référent dans le monde réel. La **résolution de la coréférence** correspond à *la mise en correspondance d'entités - qu'on appelle une* ***chaîne de référence*** *- en prenant compte leurs variations possibles* (Jean-Louis, 2012).

La définition de la coréférence est plus étroite que la définition des anaphores : une coréférence est une anaphore, une expression anaphorique et son antécédent sont coréférentiels s'ils ont le même référent dans le monde réel.

Par exemple :

Pierre *n'avait pas de* ***voiture****, je* ***lui*** *ai prêté* ***la mienne****.*

Lui est une expression anaphorique dont l'antécédent est **Pierre. Lui** et **Pierre** sont coréférentiels car ces expressions désignent la même personne dans le monde réel.

La mienne est une expression anaphorique qui pointe vers **voiture** par une relation d'identité sémantique. Il s'agit bien d'une **voiture** mais il ne s'agit pas de la même entité réelle. En effet, **voiture** dans la première partie de la phrase n'a pas d'existence.

La mienne est aussi une expression anaphorique qui pointe vers l'antécédent **je** par une relation d'appartenance.

Autrement dit, lorsqu'une relation linguistique s'établit entre deux unités lexicales qui partagent le même référent, il y a une relation anaphorique entre elles. Lorsqu'elles désignent le même objet du monde réel, cette relation devient une relation coréférentielle, les deux unités lexicales partageant le même référent. Au contraire, lorsqu'elles ont les mêmes propriétés mais qu'elles ne désignent pas le même objet du monde réel, elles sont en relation de *coréférence virtuelle* (Milner, 1976).

Comme les anaphores, les coréférences peuvent également faire appel à des relations sémantiques plus complexes, telles que la relation méronyme-holonyme ou l'hyperonyme-hyponyme :

Par exemple :

*Je viens d'emprunter **un livre** sur la condition des femmes au XII^e^ siècle. Julie avait beaucoup ri de mon intérêt pour **cet ouvrage** et pour toutes **les autres vieilleries** que j'aimais lire. Mais pour moi, **ce livre** est un vrai chef-d'œuvre.*

Dans cet exemple, l'hyperonyme **cet ouvrage** et l'hyponymie **un livre** sont coréférentiels.

Dans un texte, une chaîne coréférencielle est une suite d'expressions qui sont coréférentes. Dans l'exemple, les trois éléments **« un livre », « cet ouvrage »** et **« ce livre »** forment une chaîne coréférencielle.

La résolution des anaphores englobe l'identification des chaînes de référence. La résolution des anaphores correspond à la recherche d'un antécédent à une ou plusieurs expressions anaphoriques (équivalent de « chaîne de référence »).

Les principales relations de dépendance qui influent sur les syntagmes nominaux sont les co-références et les anaphores. L'une des principales différences entre co-référence et anaphore (dans le sens large) est une question d'identité.

L'une de ces relations ne peut se passer d'une identité implicite, en l'occurrence la co-référence, l'autre, l'anaphore, ne sous-tend pas une identité et on peut alors justement appeler ces cas des anaphores non co-référentielles (cas de l'anaphore associative, par exemple). La coréférence sans anaphore est une forme particulière où la référence est double mais elle concerne le même objet dont l'identité est figée. C'est donc une relation sémantiquement informative sur l'objet mais qui n'affecte pas particulièrement l'une des deux facettes de son identité. (Beust and Nicolle, 1997)

Par exemple :

> *Le président de la république est l'ancien maire de Paris.*

Les syntagmes nominaux *président de la république* et *ancien maire de Paris* sont co-référentiels et la co-référence est marquée de façon supplémentaire par la copule avec la présence du verbe *être*.

1.2.2. TYPOLOGIE DES ANAPHORES

Il existe différents types d'anaphores que l'on peut classer selon le terme qui fait appel à l'entité pré-introduite, ce terme pouvant être soit un segment comme un pronom (personnel, relatif, démonstratif, possessif, indéfini), un GN (défini, indéfini, ou démonstratif), soit toute une proposition ou même une phrase.

(Maillard, 1974) distingue deux grands types d'anaphores: l'anaphore résomptive (ou anaphore événementielle, ou anaphore synthétisante) et l'anaphore segmentale. (Mitkov, 2003) a abordé un autre type d'anaphore qui existe dans certaines langues comme le chinois ou le japonais, c'est l'anaphore zéro. Une anaphore pronominale zéro se produit lorsque le pronom anaphorique est omis, mais l'élément omis est néanmoins compris.

L'anaphore zéro est également connue sous le nom d'anaphore «invisible» (Mitkov 2003), l'anaphore par l'ellipse ou l'anaphore elliptique. Il n'est pas explicitement mentionné mais déclenche un antécédent pour déduire le sens global d'un énoncé. Dans la langue française, il est aussi possible de créer une anaphore zéro tout simplement en ne mettant aucun élément anaphorisant, ou autrement dit, on omet la mention du référent, qui peut être :

- Un pronom, par exemple :

Mitterrand (François). Homme politique français. Mobilisé au début de la Deuxième Guerre mondiale, il fut fait prisonnier, Ø parvint à s'évader, Ø entra dans la Résistance et Ø fonda le Mouvement national des prisonniers. (Kleemann-Rochas et al., 2003)

Dans cet exemple, l'ellipse est marquée par Ø (= zéro), l'élément anaphorisant qui aurait été *« il »* est sous-entendu. Ce procédé ne s'emploie qu'avec des verbes coordonnés ou juxtaposés ayant le même sujet.

- Une proposition, par exemple :

*Est-ce qu'**il fera beau** ? Je pense. (a)*

*Est-ce qu'**il fera beau** ? Je **le** pense.(b)*

Dans l'exemple *(a)*, l'élément anaphorisant est inexistant. L'élément anaphorisant qui aurait été *« le »* (mis pour *qu'il fera beau*) est sous-entendu.

- Un groupe nominal, par exemple :
  *Je suis venu avec **mon parapluie**, je suis reparti sans.*

(Kleemann-Rochas et al., 2003) considèrent que la préposition *sans* est orpheline.

L'anaphore résomptive est aussi abordée par Mitkov (2003), elle concerne une unité de dimension supérieure ou égale à la phrase qu'elle condense, résume et qualifie. L'anaphore résomptive peut être évoquée grâce à une proposition ou un groupe nominal.

Exemple :

(a) La femme vend ses enfants. ***Cette pratique*** *est abominable*

(b) La femme vend ses enfants. ***Cela*** *est abominable*

Dans l'exemple (a), le contenu de la première phrase est résumé par le GN « *cette pratique* **».** (a) relève aussi d'un cas spécial de l'anaphore nominale.

Dans l'exemple (b), il est repris par le pronom « *cela* », **(**b) est aussi un cas d'anaphore pronominale totale (qui sera expliqué dans la partie anaphore pronominale), qui vise à remplacer la totalité d'une proposition par un pronom.

Si l'anaphore résomptive reprend une unité de dimension supérieure ou égale à la phrase, **l'anaphore segmentale** reprend une unité de dimension inférieure à la phrase (Maillard, 1974). En fonction de la catégorie grammaticale de ces unités, on peut diviser les anaphores segmentales en cinq catégories : les anaphores verbales, les anaphores adjectivales, les anaphores adverbiales, les anaphores pronominales et les anaphores nominales.

- Les anaphores verbales se produisent lorsque le sens d'un verbe est inféré par un verbe mentionné précédemment.

Exemple :

Isabelle bricole mieux que son mari ne le ***fait****.*

Le plus souvent, la locution verbale « *le faire* », qui représente un verbe dénotant un processus, sert de pro-forme verbale.

- Les anaphores **adjectivales** se réalisent généralement au moyen de l'adjectif « tel » pour représenter une proposition précédente.

Exemple :

*Le gouvernement se croit réussi. **Tel** n'est pas l'avis du parti d'opposition.*

- L'anaphore **adverbiale** est la reprise d'un terme à travers un adverbe de manière, du type "ainsi", "pareillement", de l'adverbe locatif "là", ou temporel « à l'époque »

Par exemple :

*Sa mère le priait d'aller chez le dentiste, mais c'était justement **là** qu'il ne voulait pas aller.*

- L'anaphore **pronominale** est l'un des types d'anaphore les plus étudié de l'anaphore dans la littérature (Mitkov, 2003). Comme son nom l'indique, l'anaphore pronominale reprend un élément par un pronom : personnel, possessif, démonstratif, réflexif etc.

Par exemple :

*La femme met au monde des monstres puis **elle** les vend.*

Par un pronom démonstratif :

*Elle a vendu ses enfants ; **ceux-ci** sont montrés à la foire.*

Par un pronom relatif :

*La femme met au monde des enfants **qu'**elle vend.*

*(*Les exemples sont pris dans *Maupassant, La Mère aux monstres, 1883)*

Par un pronom indéfini :

*Cette histoire vous paraît abominable mais **tout** est vrai.*

Lorsque le pronom anaphorique représente la totalité du groupe nominal anaphorisé, on parle d'une anaphore pronominale totale, ou d'une anaphore pronominale « par représentation totale ». L'anaphore est dite « par représentation partielle » lorsque la valeur référentielle du pronom anaphorique reprend une partie dénotée par le groupe nominal anaphorisé, et elle est dite « par représentation notionnelle » si elle ne reprend que le contenu lexical du groupe nominal anaphorisé (Sahiri, 2013). On remarque qu'une anaphore pronominale « par représentation totale » correspond au cas de l'anaphore résomptive.

- De son côté, **l'anaphore nominale** est le domaine le plus difficile et le mieux structuré linguistiquement. L'anaphore nominale reprend un élément par un nom ou un groupe nominal défini, en particulier un GN construit selon le schéma *le N* ou *ce N*. Si les anaphores pronominales sont étudiées surtout pour leur rôle cohésif en réalisant la continuité des idées dans le texte et n'évoquent pas d'informations nouvelles, l'anaphore nominale est remarquée pour son rôle dans l'introduction de nouvelles idées et informations qu'elle contient. Une anaphore nominale peut être aussi résomptive si l'expression anaphorique reprend la totalité de l'information évoquée, qui est supérieure ou égale à la phrase.

Par exemple :

Une femme *abominable, un vrai démon, met au jour chaque année, volontairement, des enfants difformes, hideux, effrayants, des monstres enfin ;* ***cette mère*** *les vend aux montreurs de phénomènes. Ces affreux marchands viennent s'informer de temps en temps si* ***la femme*** *a produit quelque avorton nouveau, et, quand le sujet leur plaît, ils l'enlèvent en payant une rente à* ***la mère****. Elle a onze rejetons de cette nature.* - (Maupassant, La Mère aux monstres, 1883).

1.2.3. TYPOLOGIE DES ANAPHORES NOMINALES

L'anaphore nominale est le type d'anaphore le plus fréquent dans les textes. Pourtant, actuellement, les nombreux travaux du TALN existants traitent principalement les anaphores pronominales car son traitement est moins compliqué que celui des anaphores nominales. Concernant la typologie des anaphores nominales, Georges Kleiber (1999) divise les anaphores nominales en deux catégories : les anaphores **fidèles** et les anaphores **infidèles**.

Tout d'abord, on parle des anaphores fidèles lorsqu'on reprend un nom ou un groupe nominal par le même nom ou le même groupe nominal en l'introduisant par un déterminant différent.

« *L'anaphore est dite fidèle lorsque l'anaphore reprend toute l'information véhiculée par le mot ou par le groupe antérieurement utilisé, et n'en ajoute pas. Le cas le plus flagrant est celui de la répétition.* » - (Dispy, 2014)

(Mitkov, 2003) a abordé de ce type d'anaphore sous le nom d'anaphore nominale lexicale.

(Le Pesant, 2002a) appelle *anaphores non attributives* les anaphores *qui ne sont pas enrichies sémantiquement : elles ne s'accompagnent d'aucun apport d'information. Certaines anaphores non attributives sont de simples répétitions de l'antécédent ou de la tête de l'antécédent.*

Exemple :

*Il rencontra **un ami** qu'il n'avait plus vu depuis des années. **Cet ami** lui avait autrefois été très proche.*

Dans le cas de l'anaphore fidèle, l'emploi du démonstratif est plus fréquent et on ne confondra pas la fonction anaphorique du démonstratif avec son emploi comme déictique.

Anne Theissen (Theissen, 2001) s'intéresse au degré de fidélité des anaphores en insistant sur la présence du modifieur dans le groupe nominal anaphorique.

Par exemple :

*Nous avons vu **un petit chien** dans le jardin du voisin. **Le (petit) chien** est vraiment adorable.*

Le GN « ***Le petit chien*** », qui est repris complètement avec le maintient du modifieur, est appelé le GN défini totalement fidèle, et l'anaphore devient l'**anaphore hyperfidèle** dans ce cas.

Parmi les anaphores fidèles, nous pouvons compter l'anaphore nominale de type « ce N », que Condamines (Condamines, 2005) appelle l'anaphore **démonstrative.**

Au contraire de l'anaphore nominale fidèle, une anaphore nominale est considérée comme **infidèle** lorsque le nom de la forme de rappel est différent de celui de la forme introductrice. Il s'agit le plus souvent d'un synonyme ou d'un hyperonyme (Georges Kleiber, 1999).

Les anaphores infidèles peuvent s'illustrer par l'alternance de diverses structures de mots, des modes, du temps et des voix, il peut s'agir:

- Des sigles, exemple :

*Je viens de signer un **CDI**, mon premier **contrat à durée indéterminée** de ma vie.*

- Des synonymes, exemple :

*Elle est allergique à tous les **fards à joues**. Lorsqu'elle met ce **produit de beauté**, elle doit subir les démangeaisons insupportables.*

- Des nominalisations, exemple :

*Nous avons passé près de cinq ans pour **construire** cette maison. **La construction** a été réalisée par une équipe inexpérimentée.*

- Des périphrases synonymiques – une figure de rhétorique qui substitue une suite de mots au terme propre et unique qui le définit de manière imagée, par exemple :

Le dollar *a profité de l'écart de croissance, la valeur de* ***ce billet vert*** *a été revue en hausse.*

Les anaphores infidèles peuvent être divisées en types directs et indirects.

Il s'agit du type **anaphore directe** lorsque la relation entre l'expression anaphorique et l'antécédent relève d'un *synonyme,* d'où le nom *anaphore infidèle synonymique, d'un hyperonyme,* d'où le nom *anaphore infidèle hyperonymique* (Le Pesant, 2002a), d'une *généralisation* ou d'une *spécialisation*. Les anaphores infidèles directes sont aussi appelées les *anaphores attributives - « celles qui sont enrichies sémantiquement : elles sont l'occasion d'un apport d'information. Il y en a plusieurs sortes ; certaines exigent un déterminant démonstratif, alors que d'autres acceptent l'alternative ce/le. L'enrichissement sémantique peut concerner soit la tête de l'anaphore, soit le modifieur* » (Le Pesant, 2002a). Par exemple :

Un artisan *s'est présenté au commissariat, (ce, le)* ***malheureux*** *a été cambriolé hier.*

Martin *s'est présenté au commissariat, (ce, le)* ***jeune artisan*** *a été cambriolé hier.*

Les anaphores infidèles directes peuvent également être réalisées par l'utilisation des *concepts*, d'où le nom anaphore **conceptuelle** (Riegel et al., 1994). La définition de l'anaphore conceptuelle rencontre des divergences selon les points de vue de chaque linguiste. Pour (Dispy, 2014), l'anaphore est dite conceptuelle lorqu'elle résume, d'un mot ou d'un groupe de mots, le sens d'un fragment antérieur. De notre point de vue, l'anaphore **conceptuelle** est la reprise d'un groupe nominal ou d'un segment qui n'apparait pas explicitement dans la partie précédente du texte. Elle résume le contenu d'une phrase, d'un paragraphe ou d'un fragment de la partie du texte qui précède.

Par exemple :

Dans certains pays, signer un papier avec l'encre rouge porte malheur. ***Cette superstition*** *est de moins en moins connue chez les jeunes.*

Si l'anaphore résomptive, qui reprend *entièrement les informations données* dans la phrase par un mot qui condense ou résume, est une anaphore conceptuelle, par exemple :

Une mère vend ses enfants, cette pratique est intolérable

Cette pratique reprend entièrement la proposition *Une mère vend ses enfants.*

Il réfléchit souvent à cela… cette activité lui fait du bien

Cette activité reprend entièrement la proposition *Il réfléchit souvent à cela*

L'anaphore conceptuelle est utilisée pour résumer *une unité lexicale donnée*. Très souvent, la reprise prend la forme d'une nominalisation et, dans ce cas, le groupe anaphorique contient un nom ou un groupe nominal formé sur une base verbale ou adjectivale, qui apparaissait dans le contexte précédent.

Par exemple (Le Pesant, 2002b):

*Ici, l'****inondation*** *durera plusieurs mois ; (l', cet)* ***événement*** *est exceptionnel*

Ici, ***ce sera inondé*** *pendant plusieurs mois ; (l', cet)* ***événement*** *est exceptionnel*

Ici, ***ce sera inondé*** *pendant plusieurs mois ; (l', cette)* ***inondation*** *est exceptionnelle*

Lorsqu'une anaphore infidèle s'accompagne d'une relation métonymique, on parle de **l'anaphore indirecte**. L'anaphore indirecte dénote des configurations où un anaphorique ne reprend pas le référent évoqué par son antécédent textuel … mais pointe vers un autre qui pourra lui être associé d'une façon ou d'une autre (relation partie-tout, relation de métonymie, relation d'instance-classe ou plus généralement relation associative) (Cornish, 2001). La résolution de l'anaphore indirecte nécessite ainsi des connaissances générales.

(Cornish, 2001) répertorie trois types d'anaphore indirecte : l'anaphore inclusive, l'anaphore exclusive et anaphore créée.

L'anaphore inclusive implique une relation méronymique (partie-tout), d'ensemble-attribut ou d'ingrédient entre le référent du marqueur anaphorique et celui de son antécédent textuel. Ce type d'anaphore est également appelé l'anaphore **associative.** Elle est introduite par les groupes nominaux du type partie-de ou membre-de (Kleiber, 1999a). La valeur référentielle de l'anaphore associative se fait par la relation d'une idée à une autre en vertu d'une connaissance encyclopédique, elle relève d'un savoir partagé, d'une association d'idées, souvent dans une relation du tout aux parties.

Par exemple :

*Je vais à l'école avec **le vélo** qu'on m'a offert. Ce n'est pas possible quand **le pneu** est crevé.*

Dans cet exemple, ***le pneu*** est une partie du ***vélo***, on s'est rapporté à l'objet par un de ses éléments (métonymie) et cela nécessite de savoir qu'une bicyclette a ***le pneu***.

L'anaphore exclusive introduit une partition au sein d'un ensemble plus englobant entre un sous-ensemble dont fait partie le référent de l'antécédent (déclencheur extra-linguistique) et un autre sous-ensemble qui comprend le référent de l'anaphorique, par exemple :

*Jean coupe **la pomme** en quatre, il donne **un morceau** à sa voisine et mange **le reste.***

Un morceau et *le reste* sont les anaphores du type exclusives.

L'anaphore créée est différente des deux autres types, du fait que son référent est inféré à partir d'une ou de plusieurs propositions ou de l'évocation d'un événement. Par exemple :

Jean a **pris le car** pour Marseille. **Le voyage** était pénible.

L'association du GN *voyage* au GV *prendre le car* est faite par le truchement d'une inférence, selon laquelle, prendre le car pour se rendre à un lieu constitue un voyage.

(Kleiber, 2000) distingue différent types d'anaphores associatives :

- Les anaphores associatives méronymiques (Kleiber, 1996) qui repose sur une relation partie-tout :

  Il s'abrita sous ***un vieux tilleul***. ***Le tronc*** *était tout craquelé*

- Les anaphores associatives locatives (Kleiber, 1997)

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

- Les anaphores associatives actancielles (Kleiber, 1997) :

  Une vieille dame ***a été assassinée***. ***Le meurtrier*** *n'a pas été retrouvé*

  Le meurtrier est l'agent de l'***assassinat***

- Les anaphores associatives fonctionnelles :

  Nous entrâmes dans ***le village*** *et demandâmes à voir* ***le maire***

- Les anaphores associatives du type membre-collection :

  Le régiment a été défait. ***Les soldats*** *n'ont pas eu le temps de combattre*

- Les anaphores associatives du type argumental :

  La poule a pondu, l'œuf est encore chaud

- Les anaphores associatives du type prédicatif :

  Des œufs sont encore dans le nid, la ponte date d'hier

Dans le type membre-collection, quand l'expression anaphorique est déclenchée par un possessif, (Mathilde Salles, 2010) utilise le terme **anaphore possessive** :

Le régiment a été défait. ***Ses soldats*** *n'ont pas eu le temps de combattre*

L'anaphore possessive est considérée comme un sous-type de l'anaphore associative, l'usage du possessif est possible dans la plupart des cas où existe la relation collectifs-membre, mais c'est impossible dans certains cas (Mathilde Salles, 2010):

*Dans **les familles** d'origine immigrée notamment, **la mère / *leur mère** [= des familles] est en porte à faux entre sa culture d'origine et sa volonté d'intégration, elle est complètement larguée au niveau scolaire et **les enfants / *leurs enfants** [= des familles] en profitent.* (Extrait du journal *Dernières nouvelles d'Alsace* du 18 janvier 1998; Kleiber, 2000 : 57 et 59, 2008 : 320)

***Un couple** m'a rendu visite hier. **Le mari / *Son mari** [= du couple] était insupportable.* (Milner, 1982 : 28)

Pour les parties du corps, les anaphores associatives ne semblent pas acceptables (Mathilde Salles, 2010) :

*(*)**Jacques** est tombé du premier étage. **Les pieds** sont cassés.*

*J'ai fait tomber **la petite table**. **Les pieds** sont cassés.*

***Jacques** est tombé du premier étage. **Ses pieds** sont cassés.*

Par conséquent, la présence d'un marqueur supplémentaire tel le connecteur « En effet », « Du coup », « De ce fait », « Ainsi », « Donc » est indispensable dans ce cas :

*Jean a été étranglé. **Le cou** est en effet tout couvert de bleus.* (Kleiber et al., 1994)

L'anaphore associative est utilisée pour obtenir la valeur explicative, avec des descriptions, par exemple :

***Sa mère** a demandé qu'il ne sorte pas après 20h. **Cette directrice d'un lycée réputé** souhaite de bons résultats pour son fils.*

ou des énumérations, par exemple :

Les glaces claires laissaient voir ***la salle. Le fond humide de couleur vert tendre*** *laissa apparaître des guirlandes de feuillages, de pampres et de grappes.*

La typologie des anaphores nominales est représentée par le schéma suivant :

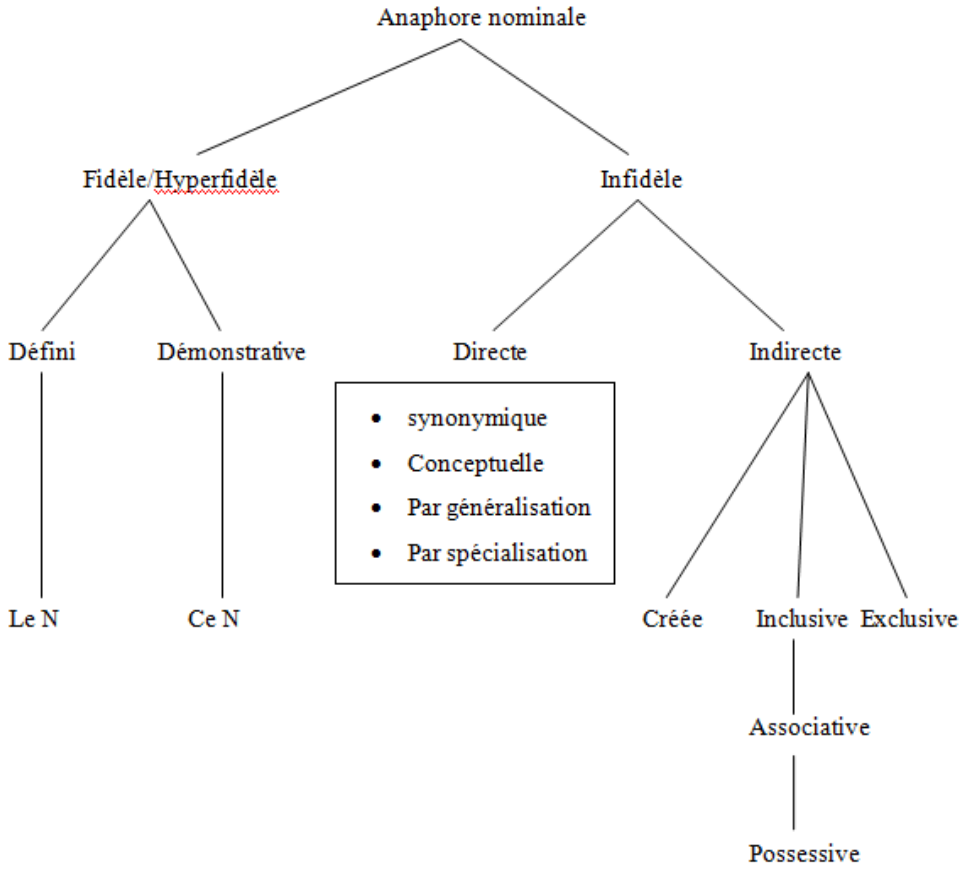


Figure 2 : Typologie des anaphores nominales

2. PROBLEMATIQUE

Dans cette partie, nous reviendrons sur les raisons qui motivent notre recherche avant de détailler les enjeux théoriques et les applications possible de notre sujet d'étude.

2.1. MOTIVATIONS

Les anaphores ont suscité des débats contradictoires quant à leur statut, leur rôle et leur utilisation. La résolution des anaphores demeure une problématique centrale du TALN car malgré les nombreuses études linguistiques en rapport avec ce sujet, les systèmes dédiés au traitement de l'information textuelle ne prennent pas en compte ces faits de langue d'une manière satisfaisante. Or, la résolution des anaphores participe tout d'abord à l'identification des *chaînes de référence* dans les discours, et, de ce fait, est fondamentale pour la compréhension automatique des textes puisque les chaînes de référence contribuent à l'organisation d'un texte, notamment en ce qui concerne sa cohésion (Corblin, 1995). Autrement dit, les relations anaphoriques constituent une source essentielle d'information sur la cohésion textuelle qui existe entre les différents segments discursifs.

La reconnaissance des anaphores en général et des anaphores nominales en particulier permet une meilleure qualité des systèmes d'extraction d'information (détecter et extraire automatiquement toutes les informations et enregistrer sous un même terme un ensemble d'informations partageant un point commun). Il s'agit de logiciels capables de « lire » des textes, de les « interpréter » dans une certaine mesure, non pas pour fournir après une recherche des mots clés mais pour répondre à des questions qui lui sont posées ; logiciels capables d'extraire les bonnes

informations des bons textes pour répondre à des questions. Pour déterminer la pertinence d'une information par rapport à une requête (une demande d'extraction), nous pouvons nous appuyer sur la fréquence des termes communs à la requête ou encore déterminer si le document contient tous les termes de la requête dans un contexte étroit. En résolvant préalablement les relations anaphoriques, nous pouvons non seulement ajouter à la fréquence des termes la fréquence de leurs anaphores mais aussi reconnaître des situations de présence de tous les termes d'une requête dans un contexte étroit même si l'un des termes est réalisé par une reprise anaphorique. Dans le cas où le système de traitement des textes doit fusionner des informations provenant de différentes sources, il est utile de détecter des relations anaphoriques entre deux ou plusieurs termes.

La résolution des anaphores est également nécessaire pour le *résumé automatique* qui consiste encore simplement en la suppression de certaines phrases. Le problème majeur réside dans la compréhension et l'interprétation des textes (Buvet, 2013). L'interprétation de texte nécessite en fait un minimum de compréhension et fait appel à de multiples mécanismes (par exemple, l'identification des relations anaphoriques pour donner le bon antécédent à une expression anaphorique, ceci grâce aux informations données par le texte ou grâce à des connaissances encyclopédiques, recours au contexte, identification de certaines structures, du niveau de langue, etc.), exercice forcément difficile pour une machine. Le système de compréhension de textes est tout système qui produit une représentation du sens du texte et qui utilise cette représentation sémantique comme point de départ d'un processus inférentiel (Nazarenko, 2004). Ce système comprend aussi bien l'extract - l'extraction des phrases saillantes d'un texte, et l'abstract - un texte généré après la phase de compréhension du texte. Dans le cas de l'extract, la résolution des anaphores contribue à pallier le manque de cohésion du texte obtenu par la simple superposition des phrases extraites. Pour trouver la réponse à une question dans un texte : si les liens anaphoriques ont été résolus, les systèmes d'extraction d'information pourront les utiliser comme points d'ancrage. Dans l'abstract, les

chaînes les plus longues sont effectivement présumées renvoyer aux entités les plus importantes du texte, et les différents syntagmes nominaux qui les composent donnent accès à la manière dont sont envisagées ces entités. Si la compréhension générale de texte est assez difficile à traiter, cela est dû à l'insuffisance des ressources grammaticales et lexicales, à l'omniprésence de l'ambigüité et au manque de modélisation des connaissances extra-linguistiques. La compréhension partielle de texte (qui vise à extraire uniquement les informations pertinentes pour une application donnée) est exploitable :

- Au niveau de la phrase (ex. EN, relations sémantiques, etc.)
- Au niveau du texte (ex. chaînes de référence)
- Au niveau du corpus (ex. regroupement de documents)

La résolution des anaphores est indispensable pour la *traduction automatique* et pour *l'aide à la traduction*, qui nécessite d'abord une interprétation du sens du texte pour pouvoir ensuite le reformuler dans la langue cible. (Mitkov, 1996) montre le rôle crucial d'une phase de résolution des liens anaphoriques pour un système de traduction automatique, notamment pour la traduction des pronoms (par exemple « le » = « he »/ « it » ?).

La résolution des anaphores joue aussi un rôle important dans la *reconnaissance de la parole* pour lequel un modèle linguistique vient supporter le modèle phonologique. L'intégration d'un tel module entre l'analyse sémantique et les applications se révèle un facteur indéniable pour la qualité de celles-ci.

La résolution des anaphores favorise la *gestion des langues contrôlées*. La notion de langue contrôlée provient du souci d'éradiquer dans certains textes – manuels de fabrication - tout risque d'ambiguïté : interdiction de certains mots trop vagues, de certaines structures trop complexes pouvant porter à confusion, comme la pronominalisation : logiciels capables de contrôler si la langue contrôlée est correctement utilisée.

Du point de vue de la linguistique didactique dont l'un des objectifs est l'amélioration de l'enseignement et l'apprentissage d'une langue, le phénomène de l'anaphore influence fortement la réalisation d'une lecture et la compréhension du texte. De nombreuses études ont été ainsi consacrées au rôle de l'anaphore dans la lecture du texte en français langue étrangère. Elles ont confirmé que ce phénomène est important non seulement pour la compréhension superficielle du texte, mais aussi pour sa compréhension en profondeur.

2.2. LES ENJEUX

Le phénomène de l'anaphore occupe une place importante dans les recherches récentes et la résolution des anaphores est désormais devenue une des tâches majeures en TALN. La résolution des anaphores nominales s'inscrit particulièrement dans la perspective de l'automatisation de ce type de tâches.

Si l'humain détecte aisément et instantanément les anaphores dans une ou plusieurs phrases, même à l'échelle textuelle, la machine est confrontée à plusieurs défis pour y parvenir. Elle doit reconnaître les groupes nominaux et les apparier en se fondant sur la relation entre eux. Pour un tel travail, il faudra délimiter au plus juste les unités et les catégoriser convenablement. Le système automatique doit ainsi prendre en considération tous les paramètres nécessaires : le contexte, les critères morphosyntaxiques et sémantiques des unités lexicales, etc. Les phénomènes d'homonymie, de synonymie et de métonymie viennent encore compliquer la tâche.

Les deux grands axes de ce travail de thèse sont :

- L'annotation des groupes nominaux selon le thème traité

- L'identification des antécédents pour un groupe nominal considéré comme anaphore (le procédé de l'identification doit se fonder sur le poids de saillance de l'antécédent).

Pour la résolution de l'anaphore associative du type partie-tout par exemple, il s'agit de savoir si l'encodage des relations partie-tout dans le lexique est fait d'une manière systématique ou, au contraire, si cet encodage procède par association.

Nous pouvons comparer ces deux exemples de (Kleiber, 1999a):

L'église était ouverte. Le clocher était penché (a)

L'église était ouverte. Le toit était penché (b)

Dans (a), le nom *clocher* est nécessairement constitutif du nom *église,* ou autrement dit, *le clocher* est un méronyme du nom *église*.

Dans (b), le nom *toit* n'est pas uniquement constitutif du nom *église* mais de tous les noms de bâtiment.

Le premier enjeu qui se pose, en particulier pour ce type d'anaphore, est celui-ci : peut-on imaginer un encodage plus dynamique de telle sorte que le fonctionnement méronymique d'un nom ne soit pas encodé en fonction du seul nom holonymique (Buvet, 2013). Le deuxième enjeu vise à mesurer la part de l'extralinguistique dans le traitement des anaphores nominales, qui est l'un des aspects les plus délicats à traiter en TALN.

Pour notre part, nous montrerons comment la notion lexicographique de domaine peut apporter des réponses formelles à des questions qui ne se prêtent guère au formalisme.

A côté de ces enjeux d'ordre théorique, le développement d'application dans le traitement des anaphores nominales permet de tester la validité des concepts utilisés dans les analyses linguistiques théoriques et cet aller-retour entre l'applicatif et le théorique est fondamental pour faire avancer les connaissances.

Puisque la résolution des anaphores nominales fait appel à des connaissances de nature lexicale, syntaxique, sémantique et pragmatique, ainsi qu'à une compréhension du contexte, pour lever les éventuelles ambiguïtés, le premier enjeu au niveau applicatif concerne la constitution et la structuration des ressources électroniques, telles que les dictionnaires électroniques, exploitées par l'application. Plus précisément, la question est : quelles sortes d'informations métalinguistiques constituant la microstructure du dictionnaire sont nécessaires pour détecter des relations anaphoriques (Buvet, 2013) ?

Pour ce qui est du deuxième enjeu applicatif, nous remarquons que les travaux effectués jusqu'ici sont nombreux, variés et obtiennent de très bons résultats. Il n'est pas envisageable de rivaliser avec ces travaux menés durant plusieurs années par les spécialistes. Notre étude vise à élaborer nos propres graphes, nos propres dictionnaires électroniques en nous appuyant sur des corpus enrichis d'informations métalinguistiques et sur des expérimentations personnelles.

Le troisième enjeu applicatif concerne les stratégies d'appariement (ou de rattachement) des groupes nominaux au niveau textuel, en utilisant le calcul des poids de saillance. Cette problématique n'est pas totalement nouvelle et des solutions proposées peuvent être catégorisées selon deux paradigmes: les approches relevant de l'ingénierie des connaissances et les approches à base d'apprentissage statistique. Les approches relevant de l'ingénierie des connaissances se caractérisent par l'utilisation d'un ensemble de règles créées par des experts du domaine de la linguistique. Les systèmes à base de règles servent généralement à constituer un ensemble de règles ou patrons d'extraction totalement attaché à un domaine, soit de façon manuelle (Hobbs, 1993) soit de façon automatique (Aone and Ramos-Santacruz, 2000), les approches à base d'apprentissage statistique cherchent à diminuer l'intervention des linguistes en utilisant des méthodes statistiques. Les systèmes à base d'apprentissage statistique, quant à eux, se servent des traits caractéristiques contenus dans des exemples, dits d'entraînement, pour apprendre à reconnaître les informations à extraire. Dans le cadre de notre recherche, nous

reposons principalement sur des ensembles de règles d'agrégation fortement liés à un domaine. Certes, cette méthode est faiblement portable par rapport aux approches plus globales à base d'apprentissage automatique, mais elle rend les règles plus performantes.

Le dernier enjeu applicatif concerne l'analyse des corpus, en particulier les corpus provenant du web. Toutes les informations présentes actuellement sur le web représentent une source d'informations colossale, qui s'enrichit de jour en jour. Or la plus grande partie des informations disponibles librement sur le Web se présentent sous la forme textuelle, c'est-à-dire non-structurée. L'analyse des documents non structurés constitue un véritable enjeu économique mais également un enjeu scientifique car ils sont révélateurs d'un nouveau mode de communication écrite. Les avantages des corpus web se trouvent dans la richesse des textes. On peut y trouver librement une variété de thèmes, de style, de niveau de langue etc. Ces corpus peuvent être écrits ou parlés, publics ou privés, leur contenu peut être général ou spécialisé, en sens unique (articles de journal) ou en interaction (forum, blog), leurs destinataires sont de différents âges, de différentes professions ; et les niveaux de langage sont ainsi variés d'un destinataire à un autre.

Pourtant, les désavantages des corpus aspirés du web sont aussi importants. Tout d'abord, les façons de structurer les pages web sont différentes d'un site à un autre, mais elles changent aussi d'un jour à un autre. De plus, les pages web du même site rédigées à différents moments peuvent être structurées de différentes manières.

Comme la réalisation des pages web passe par différents protocoles dont les deux principaux sont http et https, ce n'est pas toujours facile d'aspirer tout ce que nous souhaitons, surtout pour les contenus cryptés (pour les raisons de sécurité) ou les contenus réservés au abonnés (pour les raisons financières).

La difficulté au niveau de l'aspiration des pages web provient aussi à la manière d'écriture de code html des webmasters (des personnes qui se chargent d'élaborer des sites web). Par exemple, techniquement, il est très difficile d'aspirer le texte

imbriqué dans le nœud <h3> suivant, car la balise ouvrante contient un espace et la balise fermante ne le contient pas :

```
<h3 ><a href="http://www.60millions-mag.com/forum/rue-du-commerce-f5
6a789bfdc08885#p93837">Re: Publicité et promo mensongères</a></h3>
```

Le code html tolère les espaces (permet l'existence des espaces, qui seront ignorés automatiquement lorsqu'on affiche par un navigateur) alors que la plupart des parseurs html, qui servent à parcourir les balises et extraire le contenu texte, ne les acceptent pas. Les parseurs vont trouver un manque d'équivalence entre la balise ouvrante et la balise fermante, et le nœud sera ignoré. Ainsi, le texte ne peut être extrait.

3. OBJECTIF DU TRAVAIL

3.1. HYPOTHESES

Face à cette évolution du monde de l'information, l'objectif des linguistes informaticiens consiste à exploiter efficacement de l'information : trouver les bonnes informations, les analyser, les structurer, les classer dans un ensemble de données pour les utiliser plus tard dans diverses applications. Les méthodes créées pour cette tâche sont nombreuses, en fonction des ressources existantes et des outils de traitement choisis, et les résultats obtenus sont donc très variés.

La modélisation linguistique fondée sur la théorie des trois fonctions primaires sert de cadre méthodologique aux études des faits de langue menées au laboratoire LDI. Elle a intégré dans son dispositif descriptif une grande partie de la typologie des anaphores nominales de Kleiber car la typologie présente l'avantage d'être

compatible avec la modélisation linguistique pour ce qui est du traitement des anaphores nominales (Buvet, 2013). Nous présentons dans cette partie des hypothèses sur : 1) la nécessité des ressources lexicales fondées sur la théorie des trois fonctions primaires dans la résolution des anaphores nominales ; 2) l'importance de la méthode fondée sur le calcul de la saillance des unités lexicales.

Les hypothèses concernant l'importance de certains éléments que nous supposons nécessaires au traitement automatique des anaphores connu actuellement peuvent être démontrées si nous arrivons à répondre au questionnement suivant :

- Comment exploiter l'étiquetage anaphorique pour l'interprétation des textes ?
- Quel est l'apport des ressources linguistiques pour le traitement automatique ?
- Quelle analyse rétroactive sur la théorie permet la mise en place de l'outil informatique ?

3.1.1. LA NOTION DE SAILLANCE

La notion pragmatique de saillance est souvent utilisée par les linguistes dans la résolution de l'anaphore. La question de l'anaphore apparaît tout à fait centrale en sémantique et en logique, car elle est indispensable à l'individuation des termes et des prédicats.

Pour Apothéloz (Apothéloz, 1995a), la saillance, en tant que propriété des objets dans la représentation discursive, est une dimension clé dans le traitement de l'anaphore. Il distingue deux types de saillance : la saillance locale et la saillance cognitive. *La saillance locale est liée à la dimension nécessairement séquentielle du texte: l'objet le plus saillant localement, en un point donné du texte, est celui qui a été activé le plus récemment.* La saillance cognitive relève du fait que dans un univers d'objets élaboré discursivement, certains objets sont généralement plus centraux, ou plus pertinents, et d'autres plus périphériques. L'objet le plus saillant cognitivement est alors celui

qui occupe la position la plus centrale dans l'univers d'objets considéré : c'est aussi celui dont l'effet organisateur est le plus fort dans cet univers. *La saillance locale a donc à voir avec la situation immédiate, avec les évidences perceptives. Tandis que la saillance cognitive relève des connaissances et des représentations.* Apothéloz (Apothéloz, 1995a) rappelle que la notion de saillance permet de prédire le choix d'une expression référentielle.

Kleiber a fait apparaître diverses dimensions permettant le choix d'une expression référentielle, comme une réaction émotionnelle, une contrainte liée à la structure argumentative ou à l'organisation textuelle.

Nous supposons que la saillance grammaticale joue un rôle moins important pour résoudre les anaphores nominales que pour les anaphores pronominales.

Nous remarquons que le traitement sémantique des textes se conçoit généralement de deux façons différentes selon qu'il porte sur des **unités lexicales ou sur les unités énonciatives**. Lorsqu'il néglige les paramètres énonciatifs et s'intéresse aux seules unités lexicales, le traitement sémantique, quelle que soit sa qualité, perd une grande partie de sa pertinence. Inversement, tenir compte des paramètres énonciatifs en omettant les contenus propositionnels qu'impliquent les relations syntaxiques entre les unités lexicales restreint la portée du traitement informationnel de textes (Buvet, 2011)

La mise en œuvre du tableau de poids de saillance sémantique permet de dépasser ce clivage, car il s'agit de tenir compte de tous les facteurs sémantiques, aussi bien d'ordre lexico-syntaxique que d'ordre énonciatif, et de déterminer en quoi leurs interactions contribuent à la compréhension automatique des textes.

De plus, relevant à la fois du niveau intra-phrastique et du niveau inter-phrastique, les anaphores n'ont pas le même statut informatif selon qu'elles mettent en jeu des reprises totales (anaphore infidèle par synonymie) ou partielles (anaphore associative fondée sur une relation partie-tout). Les reprises totales et les reprises partielles ne seraient donc pas dotées du même poids sémantique puisqu'elles ne permettent pas

d'effectuer les mêmes inférences sur le contenu textuel. Ainsi, le calcul du poids de saillance permet d'inférer le contenu informatif du texte, notamment les principales thématiques traitées, et sa nature, par exemple les discours structurés ou non structurés, les argumentaires, les descriptions, les narrations, les communications unidirectionnelles ou interactives etc.

Notre dernier objectif est l'élaboration d'un protocole d'évaluation pour établir la portée des résultats obtenus du point de vue de la compréhension automatique des textes. D'autres conditions formelles seront prises en compte pour améliorer les résultats.

3.1.2. LE CHOIX DU CORPUS ET DU GENRE TEXTUEL

Un corpus est un regroupement de textes utilisés pour une étude particulière (Habert, 1997). Le corpus assure un rôle important dans les analyses de textes. Quels sont les critères d'un bon corpus ? La réponse se trouve dans l'objectif de chaque analyse qu'on souhaite mener. Généralement, le choix d'un corpus doit répondre à des critères : taille du corpus, type du corpus, choix du genre textuel du corpus, degré de spécialisation du corpus, représentativité du corpus, etc.

Il s'agit tout d'abord de constituer des corpus conséquents et de les étudier pour analyser les différents types d'anaphores nominales qu'ils contiennent. En analysant les propriétés des anaphores qui correspondent à des conditions formelles pour la résolution automatique d'anaphores, nous nous rendons compte de l'importance du choix des corpus et du genre textuel dans notre travail.

LA TAILLE DU CORPUS

La représentativité d'un corpus dépend tout d'abord de sa taille. *D'un point de vue statistique, on peut considérer un corpus comme un échantillon d'une population* (Habert, 1997) Quand un échantillon est trop petit pour représenter avec précision la population réelle, l'incertitude survient. Ainsi, la taille d'un corpus doit être assez grande pour que cet échantillon ait un caractère représentatif. La représentativité d'un corpus dépend non seulement de la taille du corpus mais aussi de son degré d'homogénéité, car *un corpus doit apporter une représentation fidèle sans être parasité par des contraintes externes*.

LES TYPES DE CORPUS

Une particularité de ce travail tient aux types de corpus étudiés. Il existe différents types de corpus dont les deux plus connus sont les corpus annotés et les corpus non annotés.

Les corpus annotés sont les corpus structurés avec les informations métalinguistiques, grammaticales, syntaxiques, sémantiques etc. Dans le TAL, les corpus annotés sont, pour la plupart, représentés au format XML, un format qui permet d'étiqueter les textes avec les balisages logiques et qui vise à expliciter la structure du corpus. Le seul désavantage des corpus annotés se trouve dans le fait qu'ils sont assez coûteux, en temps et en argent.

Un des corpus annotés les plus utilisés dans le traitement automatique de la langue française est le corpus French Treebank. Ce corpus arboré, souvent appelé le FTB pour French Treebank (Abeillé, 2001) contient un million de mots et regroupe tous les articles du journal Le Monde de 1989 à 1993. Ce corpus est disponible depuis 2003 sur demande. La taille du FTB est de 385 458 mots et 12531 phrases ce qui fait de lui un corpus assez petit, mais il est le premier corpus arboré de la langue française.

Le FTB est disponible et mis à disposition au format XML par Anne Abeillé ; néanmoins pour les besoins de la recherche, de nombreuses versions sont mises à disposition par l'équipe Alpage qui fait varier les jeux d'annotations selon les versions.

```
<?xml version="1.0" encoding="iso-8859-1"?>
<!-- Dernier correcteur, Foucault Nicolas -->
<text>
   <SENT nb="1000">
      <NP>
         <w cat="PRO" ee="PRO-card-mp" ei="PROmp" lemma="six" mph="mp" subcat="card">Six</w>
         <PP>
            <w compound="yes" cat="P" ee="P" ei="P" lemma="d'entre">
               <w catint="P">d'</w>
               <w catint="P">entre</w>
            </w>
            <NP>
               <w cat="PRO" ee="PRO-3mp" ei="PRO3mp" lemma="eux" mph="3mp" subcat="pers">eux</w>
            </NP>
         </PP>
      </NP>
      <VPpart>
         <w cat="PONCT" ee="PONCT-W" ei="PONCTW" lemma="," subcat="W">,</w>
         <w cat="ADV" ee="ADV" ei="ADV" lemma="seulement">seulement</w>
         <w cat="V" ee="V--Kmp" ei="VKmp" lemma="blesser" mph="Kmp" subcat="">blessés</w>
         <w cat="PONCT" ee="PONCT-W" ei="PONCTW" lemma="," subcat="W">,</w>
      </VPpart>
```

Figure 3 : L'étiquetage dans le FTB : l'annotation morphosyntaxique du corpus

L'avantage des corpus prétraités se trouve à la normalisation des textes, l'enrichissement des informations grammaticales, mais le désavantage se trouve au temps de traitement, parce qu'ils sont souvent des corpus assez volumineux.

D'un autre côté, les corpus non annotés sont, le plus souvent, des corpus de texte brut. Le corpus de texte brut peut être spécialisé ou non selon l'objectif des recherches. Les techniques de collecte des corpus de texte brut sont assez variées, mais, face à une quantité géante d'informations qui grandit de jour en jour sur les sites web, le plus souvent, les linguistes informaticiens utilisent les techniques d'aspiration de texte sur les sites web.

Notre étude ambitionne de faire une comparaison entre les résultats obtenus sur différent corpus, dans l'hypothèse que certains corpus facilitent le traitement des anaphores nominales plus que d'autres.

LE CHOIX DU GENRE TEXTUEL

Le choix du genre textuel est aussi important que le choix du corpus dans la qualité et la fiabilité du système de traitement. Le choix du genre textuel permet, fondamentalement, de prendre en compte des affinités entre les caractéristiques extra-linguistiques et les fonctionnements langagiers.

Les études d'Anne Condamines (Condamines, 2005) montrent l'importance du choix du genre textuel dans la résolution des anaphores infidèles du type démonstratif. La mise en place de son étude se fonde sur trois paramètres : le corpus, le choix du genre textuel et la méthodologie du traitement automatique.

En comparant trois genres textuels différents : un manuel spécialisé, un corpus de textes littéraires et un corpus de discours, les études d'Anne Condamines ont montré des résultats différents pour la même méthodologie du traitement automatique.

Ses recherches montrent que l'anaphore infidèle avec démonstratif ne fonctionne pas de manière très différente dans les manuels par rapport à d'autres genres textuels, comme le roman ou le mensuel. Pourtant, du point de vue de la relation hyperonymique, les supplétifs dans les manuels techniques pourraient facilement jouer le rôle de têtes de taxinomies.

Notre recherche suppose que le paramètre « genre textuel » joue un rôle important dans l'augmentation de la précision des algorithmes.

3.1.3. LES RESSOURCES LEXICALES

La résolution d'anaphores nominales demande une quantité importante et diversifiée des connaissances lexicales ou encyclopédiques nécessaires pour comprendre un texte. Plusieurs travaux exploitent des indices linguistiques de surface ou proposent des modèles linguistiques pour la résolution des anaphores (Salmon-Alt, 2001). Par conséquent, l'état de l'art des systèmes de résolution qui gèrent les anaphores dépend fortement des ressources « artisanales », tels que la hiérarchie lexicale WordNet.

Ce n'est qu'en 1985 que le projet WordNet voit le jour. C'est la première base de ressource sémantique à vocation universelle tant en terme de langue d'usage que de domaines traités. Wordnet contient douze grandes classes nominales : organisme, entité, entité abstraite, caractères psychologiques, phénomènes naturels, activité, événement, groupe, location, possession, précision, état. La synonymie est la relation de base dans WordNet puisque les unités lexicales dans WordNet sont regroupées dans des *synsets* (les différents sens de l'unité lexicale). A chaque entrée lexicale sont associées des informations morphologiques (lemme, partie du discours) et des informations syntaxiques (cadre de sous-catégorisation : fonctions syntaxiques profondes + réalisations possibles).

L'équivalent de Wordnet pour le français est WOLF - une ressource lexicale sémantique libre pour le français, construit à partir du Princeton WordNet (PWN) et de diverses ressources multilingues (Sagot and Fišer, 2008). WOLF contient 32 000 synsets pour 38 000 lexèmes. Les lexèmes sont organisés en partie du discours (verbes, noms, . . .) et la hiérarchie de synsets contient l'ensemble de lexèmes synonymes.

Une autre ressource pour le lexique français est Le Lefff (Lexique des Formes Fléchies du Français). Avec 110 000 lemmes pour 520 000 entrées, Le Lefff est un lexique morphologique et syntaxique à large couverture (Sagot and Danlos, 2008) .

3.2. OBJECTIFS

L'un des principaux objectifs de la thèse concerne l'intérêt de valider le plan théorique et descriptif des anaphores nominales, qui consiste à établir : **quelles sont les propriétés des différents types d'anaphore nominale qui peuvent correspondre à des conditions formelles** permettant d'implémenter dans un module de résolution d'anaphores afin de traiter deux types de relations anaphoriques, anaphore infidèle et anaphore associative.

Un autre objectif majeur de la thèse s'inscrit dans le cadre de la mise en œuvre d'un **module de résolution automatique** d'anaphores après avoir mis en évidence la nature et la structure des outils et des ressources linguistiques nécessaires à la spécification des conditions. La mise en œuvre de ce module sera l'occasion de tester puis de valider les conditions retenues à partir de l'analyse des propriétés des anaphores à l'étape précédente. La réalisation de cet objectif dépend ainsi de l'analyse des propriétés des anaphores nominales dans la mesure où les outils et les ressources linguistiques doivent permettre un traitement exhaustif des propriétés correspondant aux conditions. Il s'agira d'une part d'exploiter des outils et des ressources existants et d'autre part, d'en élaborer de nouveaux.

Notre troisième objectif est **d'élaborer des tableaux de poids sémantiques** de façon raisonnable. Ces tableaux relient deux indices : des propriétés sémantiques des anaphores d'une part et leurs poids sémantiques équivalents de l'autre. Le poids sémantique est une valeur comprise entre 0 et 1 - que l'on alloue à des éléments textuels ou des ensembles d'éléments textuels en fonction de leur rôle dans la compréhension globale du document traité, du double point de vue lexical et énonciatif. Le calcul du poids sémantique global d'un texte s'appuie sur les différents poids sémantiques qui caractérisent les paramètres pris en compte pour analyser et interpréter un texte. L'élaboration du tableau de poids de saillance est un travail arbitraire et modifiable, mais d'une grande importance pour l'analyse automatique.

3.3. RESULTATS SOUHAITES

Le rôle du développement d'applications dans le traitement des anaphores nominales est indéniable car le résultat obtenu permet de tester la validité des concepts utilisés dans les analyses linguistiques théoriques. Nous souhaitons obtenir les résultats de façon visible et compréhensible par les linguistes et par les informaticiens.

Pour illustrer ce propos, prenons l'extrait d'un commentaire de forum, à partir duquel nous cherchons à trouver les anaphores nominales et leur source. Il s'agit ici du traitement des anaphores associatives à partir d'un texte authentique.

Je viens de recevoir cette montre mais celle-ci ne fonctionne pas, j'ai beau essayer de la régler mais rien ne se passe, la trotteuse ne bouge pas, et l'heure reste figée.

La sortie attendue peut être affichée au format XML :

```
<TEXT>

...

<ANTE ID="11">cette montre</ANTE> mais celle-ci ne fonctionne pas, j'ai beau essayer de la régler mais rien ne se passe, <ANA ID="12" REF="11">la trotteuse</PRO> ne bouge pas, et l'heure reste figée.

</TEXT>
```

Pour visualiser le texte avec les anaphores résolues, nous allons choisir le format HTML en sortie. Les anaphores à résoudre sont identifiées par un numéro d'identifiant, et mises en valeur par une couleur (bleu par exemple). Les antécédents attribués, également numérotés, ont une autre couleur (rouge par exemple). Après appariement, les anaphores retiendront le numéro d'identifiant de l'antécédent correspondant.

Le code source HTML attendu est :

```
<html>
<head>
  <meta content="text/html; charset=ISO-8859-1"
 http-equiv="content-type">
  <title></title>
</head>
<body>

<span style="font-size: 11pt; line-height: 150%;">Je viens
de recevoir <span style="color: red;">cette montre <sub>11</sub></span>
mais celle-ci ne fonctionne pas, j'ai beau essayer de la régler mais
rien ne se passe, <span style="color: blue;">la
trotteuse <sub>12</sub></span><sub> </sub><sup><span
 style="color: red;">11</span></sup> ne bouge
pas, et l'heure reste figée.</span>

</body>
</html>
```

Figure 4 : Le résultat attendu - au format html

Et la visualisation sur le web devient :

Je viens de recevoir cette montre $_{11}$ mais celle-ci ne fonctionne pas, j'ai beau essayer de la régler mais rien ne se passe, la trotteuse $_{12}{}^{11}$ ne bouge pas, et l'heure reste figée.

Explication :

→ cette montre $_{11}$: le numéro 11 est l'identifiant du GN « *cette montre* » qui est l'antécédent choisi

→ la trotteuse $_{12}{}^{11}$: le numéro 12 est l'identifiant du GN anaphorique, le numéro 11 rappelle l'identifiant de son antécédent choisi, ici le GN anaphorique « *la trotteuse* » se réfère au GN « *cette montre* »

Ce format nous permettra d'explorer de façon rapide et efficace les divers phénomènes mal traités par les modules de résolution.

CHAPITRE 2. ETAT DE L'ART

Dans ce chapitre, nous reviendrons sur les concepts qui se servent de bases pour notre recherche. En nous appuyant sur les travaux existants effectués dans le domaine de la linguistique et du TAL, nous confronterons les définitions des anaphores puis nous rappellerons les étapes sur lesquelles se base notre recherche. Nous évoquerons également les méthodes de repérage et d'étiquetage existantes.

1. LES ANAPHORES NOMINALES DU POINT DE VUE DES THEORIES LINGUISTIQUES

1.1. LA THEORIE DES TROIS FONCTIONS PRIMAIRES

Le modèle de données que nous exploitons pour nos analyses linguistiques est la théorie des trois fonctions primaires qui a pour finalité d'expliquer les mécanismes langagiers en privilégiant le lexique comme objet d'étude.

En effet, l'analyse de la phrase fondée sur la structure prédicat-argument consiste à distinguer les éléments de la phrase selon leurs fonctions, prédicative, argumentale et actualisatrice (Mejri, 2009). Il s'agit d'analyser conjointement les propriétés morphologiques, syntaxiques et sémantiques des unités linguistiques selon leur fonction.

Les trois fonctions primaires permettent de catégoriser les éléments linguistiques sur le plan syntactico-sémantique dans le cadre de la phrase et d'expliquer le rôle qu'ils jouent dans la construction d'une phrase (Buvet, 2011)

LE PRÉDICAT

Du point de vue de la théorie des trois fonctions primaires, la phrase repose toujours sur un élément essentiel : le prédicat. Il n'y a pas de phrase sans prédicat.

Dans son livre sur la sémantique lexicale *Introduction à la lexicologie explicative et combinatoire*, Mel'cuk (Mel'cuk et al., 1995) utilise les concepts de prédicat sémantique et d'argument de prédicat sémantique. Un prédicat sémantique (souvent un verbe) a nécessairement un certain nombre d'arguments spécifique, même s'ils ne sont pas tous mentionnés : il faut donc leur faire une place à tous, même si on ne peut pas les

spécifier dans la modélisation. Le nombre d'arguments pour un prédicat varie selon Mel'cuk (Mel'cuk et al., 1995) de 1 à 6 en fonction du prédicat présent et des langues (constat qui n'a qu'une valeur théorique). Par exemple :

dormir : 1 argument (x dort)

poids, peser : 2 arguments (x pèse un poids de y grammes)

donner : 3 arguments (x donne y à z)

vendre : 4 arguments (x vend y à z pour une somme w)

louer : 5 arguments (x loue y à z pour une somme de w pendant une période t)

exiler : 6 arguments (x exile y de lieu1 vers lieu2 pour cause z pour une période t). Les arguments x, y, l, z, t, w sont les actants sémantiques du prédicat : ils sont nécessaires pour spécifier le sens du prédicat (c'est pourquoi tous les arguments sont toujours implicitement présents dans le sens même des prédicats).

Mais tout ce qui accompagne un prédicat n'est pas forcément un argument (les compléments). Par exemple :

Pierre travaille dans le jardin ;

Pierre a vendu sa voiture à Paris

Ces groupes sont des adjoints, ce ne sont donc pas des arguments ; une action se passe certes toujours en un lieu et en un temps donné, mais ces groups adjoints ne sont pas des informations nécessaires qui spécifient le sens, c'est-à-dire le prédicat, ce ne sont donc pas des arguments. D'autre part, tous les adjoints ont tendance à aller avec n'importe quel prédicat, alors qu'un prédicat particulier appelle certains arguments en particulier.

Un prédicat peut aussi occuper différentes formes :

- Nominal : *Luc fait un **câlin** à Léa*

- Adjectival : *Luc est **fatigant***
- Prépositionnel : *Le livre est **sous** la table*

Un argument est aussi polymorphe :

- un groupe nominal : *Luc dresse un **caniche***
- une complétive : *Luc dit que **Léa viendra***
- ou une infinitive : *Luc pense **partir**.*

Selon Polguère (Alain Polguère 2003), les unités prédicatives sont distinguées des unités non prédicatives par les caractéristiques suivantes :

- Les prédicats sémantiques dénotent des situations, ou des faits, impliquant des participants. Par exemples une action (manger, marcher, addition etc.), un état (sourire, être malade, aimer…)
- Les semi-prédicats dénotent des entités impliquant des participants. Par exemple un individu ayant une certaine relation avec un autre (mari, fils, cousine…), une partie de quelque chose (fond, dessus…)
- Les noms sémantiques dénotent des entités. Par exemple un objet physique (caillou, arbre…), une substance (eau, terre…)

L'ACTUALISATEUR

Autre constituant de la phrase, les **actualisateurs** permettant l'insertion d'une structure prédicat-argument dans une situation de communication ou au niveau énonciatif de la phrase. La fonction actualisatrice concerne l'actualisation de tous les éléments qui s'inscrivent dans la phrase, c'est-à-dire le prédicat et les arguments, dans une situation d'énonciation (Buvet, 2011). Elle prend en charge les notions de temps, d'espace, de modalité, de personne, de genre et d'aspect, etc.

Les verbes sont actualisés par la conjugaison, les substantifs prédicatifs sont actualisés par les verbes supports, les substantifs argumentaux sont actualisés par les déterminants, et les adjectifs sont actualisés par le verbe « être ».

1.2. LE GROUPE NOMINAL SELON LA THEORIE DES TROIS FONCTIONS PRIMAIRES

Le traitement des anaphores nominales se fonde sur le classement et l'analyse des arguments substantifs et leur rôle du point de vue syntaxique et sémantique. En effet, du point de vue sémantique, les arguments substantifs sont les arguments qui donnent le sens au prédicat lorsqu'il est verbal. La nature sémantique des arguments aide à la reconnaissance de chaque emploi prédicatif en réduisant la polysémie et l'ambigüité des prédicats. Du point de vue syntaxique, les arguments substantifs fonctionnent autour du prédicat qui fait la relation dans la phrase.

Pourtant, les substantifs du français ne sont pas syntaxiquement homogènes (Buvet, 1998). Certains peuvent fonctionner, tantôt comme des prédicats et tantôt comme des arguments, alors que d'autres sont toujours des arguments. Par exemple, les noms comme <partie du corps> fonctionnent toujours comme des arguments puisqu'ils sont des noms élémentaires, alors que les noms désignant les <humain> peuvent être un prédicat ou un argument.

De plus, les propriétés syntactico-sémantiques des arguments ne permettent pas de désambiguïser l'ensemble des prédicats. Par exemple, dans la phrase *Luc voit loin,* la séquence verbale peut être interprétée comme *avoir une vue qui porte loin* ou *être prévoyant* même si on ne change pas le domaine d'arguments.

Certains mots sont ambigus par nature, c'est à dire qu'ils ont des sens différents selon le contexte, ces différents sens (prédicats) étant souvent distingués les uns des autres par leurs nombres respectifs d'arguments.

Par exemple :

Marie est une fille.

C'est l'attribut qui devient prédicat, le verbe *être* (verbe d'état) n'est jamais prédicat. On a un mot *fille* à 1 argument.

Jacques connaît la fille de Pierre.

On a un mot *fille* à 2 arguments : x est la fille de y

L'ambiguïté lexicale se traduit souvent dans le nombre d'arguments, critère qui coïncide souvent avec le nombre d'entrées par sens différents dans un dictionnaire.

Le critère du nombre d'arguments est relativement efficace dans la composition de nos ressources lexicales (les dictionnaires).

2. LES ETUDES LINGUISTIQUES DE L'ANAPHORE NOMINALE

Les références anaphoriques peuvent être classées à partir des motifs morpho-sémantiques, ou c'est-à-dire en fonction de leur forme, ou de motifs syntaxiques fondés sur les relations entre l'anaphore et l'antécédent. Les connaissances morphologiques et lexicales comme le genre, le nombre et les personnes sont essentielles pour la résolution d'anaphores et ce pour un certain nombre de raisons. Ces informations permettent d'identifier un type fréquent d'anaphores, et elles sont souvent suffisantes pour lever l'ambiguïté sur le choix du bon antécédent.

Les connaissances syntaxiques fournissent des informations cruciales sur les constituants (par exemple GN, propositions, phrases, etc.). Elles sont la base importante non seulement pour l'identification des anaphores et des antécédents éventuels, mais également pour délimiter la portée de la recherche des antécédents.

Les connaissances sémantiques nous donnent des informations sur la restriction des sélections. Elles sont très utiles pour élaborer les contraintes.

En contribuant directement à la mesure de la cohésion et de la cohérence du discours, la connaissance de la structure du discours est nécessaire à la résolution des anaphores. En effet, certains types d'anaphores exigent seulement une notion locale de la structure du discours (par exemple, les pronoms), tandis que d'autres demandent une notion globale (par exemple, les anaphores nominales définies).

Nous présentons brièvement dans cette section les théories linguistiques qui servent de base pour notre recherche sur les anaphores nominales.

2.1. POINT DE VUE LEXICAL

D'un point de vue lexical, les langues ont comme caractéristiques essentielles la polysémie, la polymorphie, le figement, la paraphrase et la vraisemblance d'occurrence, et l'étude du lexique doit s'effectuer conjointement sur les plans morphologique, syntaxique et sémantique de telle sorte que le sens des mots est défini en fonction de leurs caractéristiques morphosyntaxiques (Buvet, 2009). Toutes ces caractéristiques influencent fortement sur le phénomène d'anaphore nominale, notamment sur la relation entre les unités lexicales.

Nos études lexicales des anaphores nominales se basent sur le modèle des classes d'objets et la théorie des trois fonctions primaires, développée au laboratoire LDI. Dédiée au TALN, le modèle des classes d'objet sert de base à la conception et à la constitution des dictionnaires électroniques qui seront implémentés dans les systèmes de traitement de l'information textuelle.

2.1.1. LE LEXIQUE SELON LE MODELE DES CLASSES D'OBJET

LES CLASSES D'OBJETS

Les travaux classiques ont accordé beaucoup d'importance à la propriété de référer des groupes pronominaux. Pourtant, les travaux de Kleiber (Kleiber, 1981) (1981) ont montré que les noms ont aussi une force référentielle. En effet, la notion de classe référentielle de Kleiber affirme la nature référentielle des noms par rapport aux autres catégories linguistiques. La classe référentielle est définie comme l'ensemble conceptuel de tous les particuliers réunis sous l'étiquette de l'item lexical.

D'après une définition de G. Gross, la classe référentielle s'exprime à travers la notion de classe d'objet. Ainsi, les mots qui partagent les mêmes restrictions de sélection tendent à former des **ensembles sémantiquement cohérents**, on les appelle les classes d'objets. Les classes d'objets permettent de rendre compte de la polysémie, de la synonymie et du figement (Gross, 1998) et vice-versa, le traitement de la polysémie permet de catégoriser sémantiquement les unités lexicales sous forme de classes d'objets.

Il existe souvent une **relation d'appropriation** entre un verbe et un nom. Une classe d'objets est conçue à partir de cette relation. Par exemple, les verbes de la classe <paroles> ont des adverbes appropriés comme *distinctement* et *clairement*; les noms de la classe <crime-délit> ont des verbes supports appropriés comme *commettre* ; les noms de la classe <style> des adjectifs appropriés comme *percutant*. Le nom d'une classe provient de la forme nominale éponymique, synonymique ou hyperonymique de l'unité la plus représentative de cette classe. Par exemple : la classe <colère> caractérise les prédicats colèr-, courrou-, irrit-, rag-, rogne ; La classe <mode de nourriture> est constituée des prédicats hyponymes anthropophag-, autophagi, cannibale, carnassier, carnivore, carpophag-, etc. Les éléments dans une classe d'objets sont dotés des mêmes propriétés. Par exemple, des substantifs comme

camembert, frite, steak sont caractérisés par la classe d'objets <aliment> parce qu'ils sont tous en position complément des verbes *avaler, engloutir, ingurgiter et manger* qui définissent cette classe (Buvet, 2009)

Le modèle des classes d'objets met à part la spécificité morphologique des unités lexicales, il définit également les unités monolexicales (formes simples) et polylexicales (formes complexes).

LE NOM ÉLÉMENTAIRE & LE NOM PRÉDICATIF

Les mots ont des emplois différents d'une phrase à une autre et ne peuvent pas être interprétés de la même façon. L'étude et la description des mots doit se faire dans leur contexte phrastique.

Selon (Polguère, 2003), les prédicats sémantiques dénotent des situations, ou des faits, impliquant des participants. Par exemple une action (manger, marcher, addition etc.), un état (sourire, être malade, aimer…). Les arguments sont les éléments avec lesquels les prédicats se combinent. A chaque fois qu'on change les arguments, le prédicat change d'emploi.

Le nom prédicatif et le verbe prédicatif, malgré leur unité prédicative, ont des propriétés linguistiques spécifiques qui sont difficilement compatibles avec une représentation hiérarchisée. Nous remarquons un double rattachement de certaines de classes de noms prédicatifs aux traits généraux action et événement : *Luc a commis l'assassinat de Tom ; Il y a eu un assassinat ;* en revanche, le verbe *assassiner* s'interprète uniquement comme un prédicat d'action – *Luc a assassiné Tom*.

Un emploi prédicatif est une occurrence d'un prédicat sous forme d'adjectif, de nom ou de verbe. Un prédicatif peut correspondre à un seul emploi prédicat. Lorsque plusieurs emplois se rapportent à un même prédicat, ils peuvent avoir des formes différentes (le verbe *croire*, le nom *croyance* et l'adjectif *croyant*). Un prédicat est conçu

comme une **racine prédicative** : celle-ci est un **lemme** lorsque le prédicat a un seul emploi, **une base** lorsque le prédicat correspond à plusieurs emplois de forme différente *(cour-* est la racine prédicative des emplois *courir* et *course*). En revanche, une racine est commune à plusieurs prédicats (*boucl-*), chaque prédicat a un domaine spécifique d'arguments.

Les travaux de P.A Buvet (Buvet, 2009) se fondant sur le modèle des classes d'objets insistent sur une opposition entre les noms élémentaires et les noms prédicatifs. En effet, les noms élémentaires, par définition, sont les substantifs incompatibles avec la fonction prédicative. Ils sont catégorisés par des traits généraux :

- inanimé concret, par exemple : *galet* dans *Luc lance un galet dans la mer* ;
- inanimé abstrait, *mètre* dans *La planche mesure deux mètres* ;
- locatif, *maison* dans *Luc habite dans cette maison* ;
- végétal, *herbe* dans *Luc ramasse de l'herbe pour son lapin* ;
- animal, *lapin* dans *Luc donne de l'herbe à son lapin*

Les noms de la classe <humain> appartiennent à la fois aux noms élémentaires et aux noms prédicatifs.

La subdivision sémantique plus fine des noms élémentaires constitue des **classes d'arguments**. Les classes d'arguments contribuent à l'analyse de la polysémie des prédicats. Par exemple, la classe correspond aux noms élémentaires qui ont comme prédicats appropriés définitionnels les verbes enfiler, essayer, mettre et porter : Luc (enfile + essaye +met + porte) une (casquette + chemise + culotte + gabardine + tunique…).

Les arguments sont spécifiés par des traits généraux et, lorsque les traits ne sont pas assez discriminants, ils sont spécifiés par des classes d'objets, voire par une ou plusieurs unités lexicales. Par ailleurs, chaque acception est caractérisée par une classe sémantique.

La description des arguments est hiérarchisée : les nœuds inférieurs se rapportent aux noms élémentaires ; les nœuds supérieurs correspondent aux traits sémantiques généraux ; les nœuds intermédiaires sont constitués de classes et d'hyperclasses. En revanche, il est difficile d'élaborer une taxonomie hiérarchisée pour les prédicats car elle rend difficile l'interprétation de leur relation. Par exemple, le rattachement des adjectifs *en colère* et *colérique* aux hyperclasses <affect> et <comportement> rendrait difficile l'interprétation de la phrase *Il se met souvent en colère mais n'est pas vraiment colérique* (Buvet, 2009)

Des dictionnaires électroniques du modèle des classes d'objets sont étudiés et développés au laboratoire LDI par P.A. Buvet, visant à associer toutes les entrées lexicales à des informations métalinguistiques standardisées qui sont suffisamment explicites pour donner lieu à des procédures informatisées. Le dictionnaire des arguments est constitué de noms élémentaires et de descripteurs métalinguistiques comme la catégorie grammaticale et une spécification de genre, le trait syntactico-sémantique, la classe d'objets et le domaine d'emploi. Ils décrivent aussi bien les noms simples que les noms composés. Le dictionnaire des noms prédicatifs est constitué de l'entrée (verbes, adjectifs, noms) et d'indications sur sa structure argumentale (arg1, arg2..), une spécification de trait syntactico-sémantique et une spécification de verbe support.

ASPECT INHÉRENT ET LE SCHÉMA D'ARGUMENTS

L'aspect se conçoit comme une marque de temps référentiel et ne nécessite aucun repère temporel. L'aspect inhérent est un critère définitionnel d'un emploi prédicatif. Il y a différents types d'aspect inhérent selon que les emplois prédicatifs sont du type action et événement, d'une part, ou du type état, d'autre part. Par exemple, pour les emplois prédicatifs d'état, l'aspect inhérent peut être le provisoire, comme dans *La porte est ouverte*, ou le permanent, comme dans *Il est intelligent.*

En revanche, l'aspect cotextuel est pris en charge par des actualisateurs, il peut être le ponctuel, l'itératif, l'inchoatif, le terminatif, le continuatif, etc. Si l'aspect inhérent joue plutôt sur le niveau lexical, l'aspect cotextuel joue sur le niveau grammatical.

Le verbe *lire* équivaut à deux emplois :

Luc lit un livre ;

Luc lit surtout des romans.

Les deux emplois partagent le trait « *action* » mais chacun a un aspect inhérent spécifique : respectivement, le **duratif perfectif et le duratif imperfectif.** Le verbe *lire* dans le premier cas requiert un GN au singulier alors que le deuxième cas requiert un GN au pluriel.

Il y a deux emplois du verbe *aimer :*

Luc aime Léa

Luc aime rêver

Les deux emplois sont caractérisés par la construction X0 V X1. Le X0 dans les deux cas est le nom d'<humain> mais le X1 est différents. Le premier exemple se rapporte à un prédicat d'<amour>, la place de X1 est prise par un groupe nominal, le second se rapporte à un prédicat de <goût>, la place de X1 est occupée par une infinitive.

De même, une entrée dans le dictionnaire des prédicats est représentée ainsi sous cette forme :

	Distribution 1	Distribution 2
Dévorer : X0 V X1	X0=GN X1=GN	X0=<humain> X1=<aliment>
Dévorer : X0 V X1	X0=GN X1=GN	X0=<humain> X1=<texte>

2.1.2. PLURIEL ET EXPRESSION DE LA QUANTITÉ

Dans un travail mené par Tyne Liang et al. sur l'utilisation du Tagger pour la résolution des anaphores nominales dans la langue anglaise, (Liang and Lin, 2005) traite le nombre probable d'antécédents, grâce au phénomène du singulier et du pluriel de l'expression anaphorique, comme les syntagmes nominaux commencés par un adjectif démonstratif ou quantitatif (par exemple : 'either', 'this', 'both', 'these', 'the', et 'each').

Pour le corpus du domaine biomédical, Tyne Liang et al. filtrent les structures :

« *this* » ou « *the* » + *noms propres avec des majuscules ou des chiffres*

Par exemple, les anaphores commencées par « *this* » ou « *the* » + GN au singulier sont les anaphores singulières, elles ont un seul antécédent. Pourtant, d'autres sont considérées comme des anaphores nominales plurielles et le nombre de leurs antécédents est présenté dans le tableau suivant :

Anaphore	Nombre d'antécédents probable
Either + GN	2
Both + GN	2
Each + GN	Plusieurs
They, Their, Them, Themselves	Plusieurs
The + Nombre + GN	Nombre
Those + Nombre + GN	Nombre
These + Nombre + GN	Nombre

Tableau 1 : Nombre d'antécédents - proposé par Tyne Liang

Les syntagmes nominaux qui précèdent une anaphore reconnue dans la gamme de deux phrases seront traités comme des candidats et seront affectés d'un niveau de saillance initial égal à zéro. Les antécédents singuliers peuvent être sélectionnés suivant deux stratégies :

- Le meilleur en premier : le système sélectionne l'antécédent ayant le plus haut score de saillance et qui est supérieur à un seuil donné.
- Le plus proche en premier : le système sélectionne l'antécédent le plus proche et dont la valeur de saillance est supérieure à un seuil donné.

Pour les anaphores plurielles, leurs antécédents sont choisis comme suit:

- Si le nombre d'antécédents est connu, alors le système sélectionne le même nombre d'antécédents.
- Si le nombre d'antécédents est inconnu, alors le système sélectionne les candidats dont les scores sont supérieurs à un seuil et dont les motifs grammaticaux sont les mêmes que le candidat ayant le meilleur score.

Tyne Liang et al. a réalisé des résolution des anaphores nominales dans des corpus de text du domaine de biomédicale. Les résultats varient entre 64% - 71% selon le type de corpus choisi, entre 64% - 50.5% selon la stratégie choisie (le meilleur en premier et le plus proche en premier), de 78% (avec analyse sémantique) et 59% (sans analyse sémantique).

Le pluriel des groupes nominaux est également abordé par A.-M. Dessaux (Dessaux, 1976). Elle propose les types des noms quantitatifs dans la liste suivante :

- Noms collectifs ou noms d'ensembles, d'humains (armée, régiment etc.), d'animaux (troupeau, harde, meute etc.), de choses (arsenal, répertoire etc.)
- Noms de contenants : wagon, camion, assiette, bol, panier etc.
- Noms désignant une disposition : ribambelle, chapelet, cordon etc.
- Noms désignant des référents petits : aumône, zeste, ombre, atome, goutte etc. Certains noms prenant le sens de quantités petites comme déterminants

nominaux ne renvoient pas spécialement à des référents petits : un (nuage + soupçon) de lait.

- Noms désignant des masses : mer, flux, torrent, avalanche, monument, mur.

Elle propose également les noms dénotant intrinsèquement la notion de quantité :

- Noms désignant des quantités importantes mais indéterminées d'un autre Nom, ex. : nombre, quantité, masse, flopée, tapée, myriade, kyrielle, pléthore, multitude, abondance, foison, profusion etc.
- Noms désignant l'ensemble des autres Noms : totalité, ensemble, total
- Noms désignant des mesures : mètre, kilo etc.
- Noms de quantité numérique : centaine, millier etc.
- Noms de parties de l'ensemble des autres Noms : fragment, portion, morceau, quartier, mesure, fraction, minorité, majorité, parcelle, etc.
- Noms collectifs : collection, série, ligne, rangée, etc.
- Noms en -ée à base nominale dénotant une quantité définie par le nom de base : cuillerée, marmitée, bolée, etc.
- Des noms reliés à des verbes désignant une disposition (amas, grouillement, fourmillement, entassement etc.) dont on ne peut dire ni qu'ils dénotent intrinsèquement la quantité, ni qu'ils soient métaphorisables, puisqu'ils acceptent normalement tous les types d'un autre Nom, comme le verbe correspondant accepte tous les types de sujets.

2.1.3. LA PARTICULARITÉ DES CLASSES SÉMANTIQUES

LA CONCEPTION DES ARTÉFACTS

Le choix d'étudier des anaphores associatives de type partie-tout impliquant des noms d'artefact est une conséquence des propriétés remarquables de ces substantifs.

Par définition dans wikipédia, un artéfact est un objet façonné par l'homme. Le mot d'origine latine est composé de *ars, artis* (art), et du participe passé de *facere* (faire). Il apparaît d'abord en anglais (*artefact*), et n'est utilisé en français qu'à partir de 1921.

Selon (Kassel, 2009), « les primitives conceptuelles introduites pour les noms d'artefacts sont celles d'unité artificielle, de production intentionnelle d'objets, de capacité à exercer un rôle dans des actions d'un type donné, de fonction et d'entité fonctionnelle ». Kassel insiste sur la propriété artificielle et fonctionnellle des artéfacts, et sur le fait qu'ils sont nés d'une production intentionnelle.

Puisque la résolution des anaphores nominales consiste à donner aux machines des facultés d'interprétation, une tâche importante consiste à résoudre les dépendances sémantiques qu'une expression anaphorique entretient avec les autres mots. Ainsi, le classement sémantique des artefacts (ou des objets du monde d'une façon générale) est nécessaire.

Pour décrire les objets du monde, (Le Moigne, 1984) a proposé trois dimensions dans la conception des artefacts : la dimension ontologique (qu'est-ce que c'est ?), la dimension fonctionnelle (qu'est-ce que ça fait ?) et la dimension génétique (d'où ça vient et qu'est-ce que ça devient ?). Ces traits donnent un cadre à la catégorisation de l'objet en cause via son identité sortale et son identité individuelle (Beust and Nicolle, 1997).

Les noms d'artéfact sont incompatibles avec la fonction prédicative. Pour autant, la nature des objets qu'ils dénotent impliquant une finalité fonctionnelle, une partie

d'entre eux spécifient clairement une prédication en rapport avec cette finalité fonctionnelle car, lorsqu'il s'agit de mots construits, ils ont trait à des prédicats nominaux ou verbaux (*arrosage/arroser → arrosoir*) et, lorsque ce sont des mots simples, ils peuvent donner lieu à des prédicats nominaux ou verbaux (*marteau → martellement/marteler*). Il s'ensuit que les noms d'artefact constituent un observatoire privilégié pour étudier la fonction argumentale.

Les noms d'artéfact du français sont parfois des mots simples (par ex *couteau)* et souvent des mots construits. Les mots construits sont principalement des noms dérivés (*hachoir*) ou des noms composés (*minuterie automatique*). Les autres mots construits sont des abréviations comme *APN*, dont la source est *appareil photo numérique* ou des apocopes comme *ampli* dont la source est *amplificateur ;* ils sont relativement peu nombreux.

Par ailleurs, les noms d'artefact se conçoivent comme des holonymes (*voiture*) ou des méronymes (*roue*). Nous avons choisi de ne pas encoder spécifiquement le fait qu'un nom d'artefact est un méronyme ou un holonyme. Pour traiter les relations partie-tout, nous exploitons des descripteurs généraux : appareil, moyen de transport ou bien dispositif, organe pour les méronymes. Nous avons considéré que la relation partie-tout est construite avant tout dans le discours (notamment par le biais de l'anaphore associative (*Il a acheté une maison … le toit*) mais aussi par le biais de la possessivation (*Il a acheté une maison… son toit*). Il s'agit donc de l'anaphore possessive.

LA CLASSE <HUMAIN>

Pour le traitement des anaphores infidèles, nous avons choisi le thème des mots désignant des humains. Ainsi, nous avons créé la classe <humain> avec les groupes nominaux désignant les personnes avec la classification en genre, nombre.

Selon (Gross, 1998), le trait humain caractérise à la fois des noms élémentaires (homme) et des noms prédicatifs (fils). Ils constituent une catégorie syntactico-sémantique à part.

L'objet à traiter a été choisi en fonction du type d'anaphore nominale étudié. Par exemple, pour les anaphores du type infidèle, nous avons choisi un thème autour des humains. Pour ce genre de traitement, un corpus pris dans la rubrique faits-divers du journal *lemonde.fr* s'avère le plus approprié.

En effet, le choix du sujet de nom d'humain a été choisi car les noms d'humain peuvent fonctionner soit comme des prédicats :

Paul est journaliste

soit comme des arguments :

Le journaliste a interviewé Paul.

Puisque les noms d'<humain> jouent la double fonction, il nous faut préciser les particularités de cette classe.

En effet, en étudiant les anaphores associatives (Kleiber, 1999b) a indiqué que la particularité sémantique des *noms de parenté* comme *fils, père, mari, épouse*, *parent* etc. est de donner lieu à une relation sémantique converse ou réciproque :

X est père de Y, alors, Y est fils de X...

Il a remarqué que cette réciprocité ne se retrouve dans aucun cas d'anaphore associative, ni méronymique, ni locative, ni actancielle, ni collective, ne peut donner lieu à un enchaînement associatif. Lorsqu'on compare les deux phrases suivantes qui viennent de la même structure *x est mère de y*, il existe un blocage :

(a) *?? J'ai rencontré une jeune fille très malheureuse. La mère lui rend la vie impossible*
(b) *On m'a présenté une jeune fille à marier, mais la mère était impossible*

Un blocage est trouvé dans cet exemple (a) mais dans l'exemple (b), le blocage n'existe plus car « *la mère* » (b) est non pas directement comme *la mère* de *la jeune fille* mais comme la mère en général, *la mère de la famille* dont la jeune fille à marier est la fille.

Même explication pour un autre exemple :

Un beau mariage a eu lieu hier à Plaffenheim. ***Les mariés*** *étaient en blanc, le curé a fait un grand sermon et le maire a prononcé un discours républicain.* ***Les parents*** *étaient ravis.*

Il n'y a pas de blocage dans cet exemple car le GN *les parents* n'est pas défini directement vis-à-vis de *leurs enfants* qui *se marient* mais il se trouve saisi via l'ensemble famille (celle des *mariés*) au sein duquel ils sont *les parents*, autrement dit, *les parents* sont uniquement considérés comme membres de la famille des mariés et la structure *x est les parents de y* n'est pas valide. Les noms de parenté ici sont compris dans la dimension membre-collection.

Certains noms d'humains de la langue générale sont recensés dans le dictionnaire des humains, tels que père, mère, adulte, fils… cela revient de la fréquence d'utilisation de ces noms dans les textes. D'autres noms humains fonctionnent toujours comme des arguments, par exemple les noms propres.

Cette variation dans le fonctionnement des noms d'humains nécessite de distinguer s'ils sont plutôt des noms communs ou plutôt des noms propres. Ensuite, il faut distinguer les emplois des noms communs qui peuvent être variables, d'une phrase à l'autre.

2.2. POINT DE VUE MORPHOLOQUE ET SYNTAXIQUE

Nous analysons dans cette partie les anaphores nominales du point de vue de la morphologie et de la syntaxe, en faisant état de leurs caractéristiques morphosyntaxiques.

2.2.1. LA DETERMINATION DU NOM ANAPHORIQUE ET DE L'ANTÉCEDENT

LE DÉTERMINANT DÉFINI DANS LES ANAPHORES ASSOCIATIVES

Les déterminants sont divisés en définis *(le, la, les + N)* et indéfinis. Les indéfinis sont subdivisés en sous-classes en fonction des critères définis par trois propriétés syntaxiques suivantes :

- Les déterminants qui peuvent se combiner directement avec des Noms : *un livre, des livres…*
- Les déterminants qui peuvent se combiner avec un GN au moyen de la préposition « de » : *beaucoup de mes livres…*
- Les déterminants qui peuvent fonctionner comme adverbes *: je lis beaucoup, je lis très peu…*

Les deux premiers cas concernent les structures de base des groupes nominaux indéfinis. Dans ces structures, les déterminants dénotent le concept de quantité.

Les déterminants définis sont utilisés pour introduire les entités nouvelles dans les discours. En linguistique, les entités nouvelles introduites dans les discours sont divisées en deux types : les non-anaphoriques qui annoncent une nouvelle entité du discours et les anaphores associatives dont l'interprétation nécessite une connexion

sémantique avec un antécédent. D'après Kleiber, seules les *descriptions définies* permettent d'installer une anaphore associative :

Je suis arrivé dans le village. L'église était en travaux.

**Je suis arrivé dans le village. Elle était en travaux.*

**Je suis arrivé dans le village. Cette église était en travaux.*

Nous pouvons remarquer une importance de la stéréotypie lexicale, qui est l'emploi de *l'article défini*. De ce fait, le repérage de cet article peut être de nature à faciliter la compréhension des discours de la langue cible.

Selon (Posturzynska-Bosko, 2009), l'anaphore associative est distinguée des autres constructions comportant le possessif ou le pronom par le fait qu'elle exige que le référent soit interprété comme déjà connu et que la reprise soit en relation indirecte avec son antécédent, alors que les tournures comportant le possessif ou le pronom sont en relation directe avec leur antécédent :

*Les policiers examinèrent la voiture. **Ses roues** étaient pleines de boue.*

*Les policiers examinèrent la voiture. Les roues **en** étaient pleines de boue.*

Pour le cas des anaphores associatives, un référent nouveau est introduit dans un texte sous forme d'un GN précédé d'un article défini :

*Les policiers examinèrent la voiture. **Les roues** étaient pleines de boue.*

Pourtant, il ne s'agit pas d'un GN défini complet de type « Le président de la République Française» dont la valeur définie s'explique par la description elle-même, mais elle comporte l'identité de son référent. On remarque aussi que la reprise peut être introduite par un article indéfini comme dans l'exemple :

*Les policiers inspectèrent la voiture. **Une roue** était pleine de boue.*

Dans cette phrase, le référent du GN indéfini constitue une partie intégrale de la voiture mentionnée et le GN *une roue de la voiture* permet l'interprétation: une des roues de cette voiture concrète.

« *…une telle situation ne met pas en question l'exclusivité de l'emploi des définis dans les anaphores associatives, mais au contraire, elle prouve que dans le cas des indéfinis de ce type, le critère de la définitude reste sauvé* » (Posturzynska-Bosko, 2009). En effet, dans certains emplois, l'interprétation des GN amène à la construction d'une relation référentielle de caractère associatif (du type partie/tout) « *mais cette relation n'est que conjoncturellement associative, elle n'a aucun caractère de nécessité* », comme c'est le cas des GN définis.

Nous pouvons aussi constater que l'indéfini associatif *un N* est possible dans les cas où le pluriel défini *les N* lui correspond et tous les deux se situent dans le même site associatif.

***La table** est tombée, **un pied** est cassé*

***La table** est tombée, **les pieds** sont cassés*

Dans le cas contraire, les expressions définies ne sont pas toujours anaphoriques. En effet, (Grevisse, 1961) montre que les articles définis s'emploient devant les noms désignant quelque chose de bien connu ou quelque chose qui est l'objet d'un fait habituel.

Cette présupposition d'existence du référent du GN défini se traduit, dans le cas de référent spécifique. Dans l'exemple :

*Passe-moi **le livre***

Le GN défini *le livre* désigne un ou des artéfacts particuliers, alors que dans l'exemple suivant, le GN défini *le livre* désigne l'ensemble d'une classe ou sous-classe d'individu :

*Il faut que **le livre** cesse d'être une marchandise comme une autre*

« *le livre* » dans cet exemple n'est pas anaphorique car il ne se réfère à aucun élément.

LE DETERMINANT DE L'ANAPHORE DEMONSTRATIVE

Dans la résolution des anaphores nominales, la structure « Ce N » est assez complexe à traiter car elle peut aussi relever des anaphores fidèles de type démonstratif, des anaphores infidèles de type hyperonymie ou relever de la deixis (expression renvoyant à la situation d'énonciation).

La fonction anaphorique distinguait les pronoms qui renvoient à des **objets déictiques** et les pronoms qui renvoient à des **segments de discours** (les pronoms anaphoriques). Cela signifie qu'un élément anaphorique renvoie au contexte linguistique, et non pas à une réalité extralinguistique qui, elle, est renvoyée par les éléments déictiques (Householder, 1981)

Ainsi dans la phrase :

Ce livre est beau.

Le démonstratif *ce* a un emploi déictique, car nous devons nous fonder sur le contexte pour savoir de quel livre on parle.

En revanche, dans la phrase suivante :

Il y a un livre sur la table ; je veux ce livre.

Le démonstratif *ce* a un emploi anaphorique, car le groupe nominal *ce livre* renvoie à l'expression *un livre* dans la même phrase.

Du point de vue de leur forme, les déictiques ne sont pas nécessairement différents des anaphoriques mais les déictiques permettent à l'auteur de la phrase ou du texte d'entretenir une relation avec son lecteur, soit en se référant à son écrit, soit en se référant au temps et à l'espace commun (Condamines, 2005).

Dans les textes qui comportent des dialogues, des déictiques beaucoup moins prévisibles apparaissent, qui font référence à des éléments supposés communs aux deux interlocuteurs, par exemple, les noms qui renvoient au lieu et au temps communs (ce moment, le lendemain, la dernière fois, en ce temps…), les noms qui renvoient au support (ce livre, ce chapitre, cet article…) , les noms qui interviennent dans la progression du discours (cette fois, cette logique, ce contexte…).

Par exemple :

*Il pensait à ce qu'il pouvait faire **en ce moment** avec son épouse.*

Pour (Condamines, 2005), certaines structures « ce N » ne sont ni anaphoriques ni déictiques, ce sont des structures d'**autoréférentiels**. Elles construisent une référence soit aux structures comme :

« *un/une de ces* + *GN* »

« *avec/en l'absence de ce* + *GN* »

« *à l'instar de/ tel/comme ce* + *GN* »…

soit à une structure accompagnant d'un modifieur faisant appel à une **connaissance supposée commune avec le lecteur**.

Par exemple :

*On sentait là-dedans le renfermé, **le cuir des meubles, le vieux tabac et l'imprimerie** ; on sentait **cette odeur particulière** des salles de rédaction que connaissent tous les journalistes* (Bel Ami).

Si la structure « le N » peut correspondre à une métonymie, ce n'est pas le cas pour la structure « ce N ».

(Georges Kleiber 1990) a remarqué que l'anaphore démonstrative ne peut pas être une anaphore associative. En effet :

Nous entrâmes dans le village. L'église était située sur une grande place

** Nous entrâmes dans le village. Cette église était située sur une grande place.*

Selon (Condamines, 2005), la structure « ce N » peut établir une relation avec la nature langagière de l'antécédent. Une de ses études visait à traiter les anaphores infidèles du type hyperonymique et à limiter le nombre de cas à étudier. La recherche automatique est réalisée ainsi :

- Lorsqu'un « ce/cet/cette/ces N » apparaît dans le corpus, on va chercher si cette même forme N apparaît dans le paragraphe précédent. Si c'est le cas, on

peut considérer que l'anaphore n'est pas « infidèle » et donc qu'il n'y a pas de relation hyperonymique.

- Concernant la structure « ce N », il est nécessaire de laisser des données métalinguistiques (ou déictiques, autoréférentielles…) dans une stop-list : d'abord les noms qui renvoient au lieu et au temps communs à l'auteur et à l'interlocuteur ou aux protagonistes du récit. Par exemple dans la littérature, il s'agit des éléments comme : *heure, instant, jour, année, date, matin, soir, mois, printemps, été, automne, hiver, siècle, nuit, soirée, matinée*… ; ensuite, les noms qui renvoient au support comme : *article, manuel, ouvrage, chapitre, livre, paragraphe, édition, version, colonnes*… Enfin, les noms qui interviennent dans la progression du discours comme : *cas, circonstance, condition, date, égard, endroit, fait, façon, fin, fois, genre, logique, contexte, niveau, occasion, perspective, point, propos, rythme, stade, sujet, temps, titre, époque, temps* etc.
- Une fois éliminés les « ce N » non anaphoriques (cataphoriques, métalinguistiques, déictiques, autoréférentiels, figures), les « ce N » restants se répartissent dans différents types de relation : hyperonymes, supplétifs, synonymes, déverbaux, dérivés d'un adjectif, dérivés d'un nom…

Les études de P.A. Buvet (2012) montrent que l'opposition entre le déterminant défini et le déterminant démonstratif constitue un mode d'identification de l'antécédent d'une reprise.

Exemples :

*Je viens de commander **une table** basse chez […] mais **un pied** était déjà cassé. Vaut mieux que je répare **ce pied** ou simplement échanger la table?*

*Je viens de commander **une table** basse chez […] mais **un pied** était déjà cassé. Vaut mieux que je répare **la table** ou simplement l'échanger ?*

LE DÉTERMINANT INDÉFINI

Les groupes nominaux indéfinis sont les GN introduit par un déterminant du type *un N*, ou plus généralement, préfixé par : *deux, trois, plusieurs, aucun, quelques, chaque…*

Dans la résolution des anaphores, le déterminant indéfini est plutôt utilisé pour l'antécédent car la fonction du déterminant indéfini est d'introduire une nouvelle notion, un nouvel élément, ou un nouveau référent dans la phrase, il est ainsi rarement utilisé pour l'expression anaphorique qui marque la reprise d'un élément dans la phrase.

En général, les GN indéfinis sont employés de manière générique. L'existence du GN indéfini ne marque pas automatiquement la présence d'une expression de reprise. Par exemple :

Le mari a tué sa femme avec ***un fusil***

Le GN indéfini *un fusil* est une expression générique, qui ne garantit pas la présence d'une reprise (*le fusil)* quelque part.

Du point de vue du traitement automatique des anaphores, la résolution des GN indéfinis demandera un recours au contexte plus important que celle des GN définis, et nous ne chercherons pas à traiter automatiquement tous les GN indéfinis dans les corpus.

Pourtant, lorsque nous limitons notre sujet à traiter au GN de la classe <Personne>, des structures telles que :

GN indéfini […] GN défini […]

dont le GN indéfini est l'antécédent du GN défini, sont assez répandues dans nos corpus.

Par exemple :

Un couple *s'est donné la mort dans la nuit de jeudi à vendredi dans un grand hôtel parisien.* ***Les deux octogénaires*** *ont été retrouvés main dans la main […]*

Parfois, la reprise n'est que partielle :

*La police était à la recherche d'**un couple** […]. **La femme** a été arrêtée ce matin […]*

LE DÉTERMINANT SPÉCIAL « LEDIT »

Un autre cas d'anaphore nominale susceptible d'être étudié est le cas des anaphores nominales introduites par « ***ledit*** ». En présentant les résultats d'une étude sur les anaphores nominales de ce type, (Whittake and Handelshøyskole, 2002) ont révélé une autre fonction spécialisée de « *ledit* » : marquer la prise en charge du nom de l'antécédent par un autre locuteur. Dans le corpus littéraire, « *ledit* » peut aussi avoir des effets discursifs qui sont directement ou indirectement attribuables à son rôle de contrôleur d'ambiguïtés.

Ce déterminant spécial « *ledit* » est présenté comme propre à la langue juridique et administrative, donc comme relevant de la langue spécialisée. Pourtant, cette structure est assez fréquente dans des textes non spécialisés. Afin d'étudier ce phénomène plus en détail, Whittake et Handelshøyskole ont utilisé trois corpus de textes différents qui se distinguent aussi bien du point de vue stylistique que du point de vue de leur visée générale, et le résultat prouve que ces deux aspects ne sont pas indépendants l'un de l'autre. Son emploi dans le corpus juridique semble être conditionné exclusivement par la faible saillance soit de l'antécédent, soit de l'anaphore, soit des deux.

Le corpus journalistique a révélé une autre fonction spécialisée de « *ledit* » : marquer la prise en charge du nom de l'antécédent par un autre locuteur.

Dans le corpus littéraire, « *ledit* » peut aussi avoir des effets discursifs qui sont directement ou indirectement attribuables à son rôle de contrôleur d'ambiguïtés.

2.2.2. LE NOM ANAPHORIQUE

Les relations de dépendance nominales influent sur la constitution des noms anaphoriques et de leurs antécédents. Les noms anaphoriques sont construits à partir de syntagmes nominaux et ils construisent ou maintiennent l'identité des antécédents. Selon (Kleiber, 1993) le syntagme nominal qui introduit un référent a deux fonctions :

- il décrit le référent et son contenu, fournit des prédicats qui permettent de l'identifier
- il dénomme le référent.

Pour le cas de l'anaphore nominale associative, un phénomène langagier qui mobilise au moins trois domaines d'investigation de la linguistique : la sémantique lexicale, le lexique mental et l'analyse textuelle, Kleiber a proposé les caractéristiques suivantes :

- L'anaphore associative consiste en l'introduction d'un référent nouveau
- Le référent nouveau est introduit au moyen d'un GN défini
- Il exige une autre entité mentionnée auparavant dans le texte
- La relation entre l'entité antécédente et l'entité nouvelle est une association uniquement discursive ou contextuelle, elle relève d'un savoir a priori ou conventionnel associé aux lexèmes en question.

GN SANS NOM

Il existe une forme d'anaphore nominale particulière dont la tête lexicale est absente, nous les appelons les GN sans nom. Tout d'abord il faut citer les groupes nominaux « *autre* » et « *même* » qui s'accordent en genre avec leur antécédent, mais pas

obligatoirement en nombre :

*Marc a choisi un **pull** bleu, j'ai pris **le même** (pull).*

*Marc a choisi un **pull** bleu, j'ai pris **le** (pull) **rouge.***

*Marc a choisi plusieurs **pulls** bleus, j'en ai pris **un autre** (pull).* (Van Peteghem, 2001):

Il y a débat autour de leur classification : anaphore nominale ou pronominale.

En effet, « autre » et « même » ne fournissent aucune information de type prédicatif sur le référent, mais donnent uniquement des instructions référentielles. "Autre" peut se combiner avec tous les déterminants possibles, peut être postposé au substantif ou figurer comme attribut, ou il peut s'utiliser sans nom comme pronom, par exemple :

Les montagnes y avaient des lignes tout autres.

"Même" est moins adjectival et plus déterminatif :

Pierre et Paul sont les mêmes

**Pierre et Paul sont même*

Une autre étude des groupes nominaux sans nom concerne les expressions « le premier », « le second », « l'un », « l'autre » etc. réalisée par Corblin et Laborde (Corblin and Laborde, 2004).

« Le premier » et « le second » sont considérés comme un type un peu différent de groupes nominaux sans nom. Ce sont des GN sans nom admettant généralement le double fonctionnement : anaphore nominale et référence mentionnelle. Mais la dislocation à droite est acceptable pour les anaphores nominales, alors qu'elle n'est pas acceptable pour la référence mentionnelle. Par exemple :

Passe-moi le premier, de pull

** Le roman est paru avant la pièce, mais le premier, de roman, est bien meilleur.*

Dans certains exemples, il est nécessaire de se référer au discours pour saturer l'interprétation :

Le film** est paru avant **le roman** ; je préfère **le premier** au **second

Il est clair que la situation d'énonciation est prise en compte pour interpréter cet exemple.

"*L'un*" et « l'autre » ont également un fonctionnement mentionnel comme le cas de « le premier » et « le second ». Pourtant, les mentions antérieures (ou les antécédents) n'introduisent pas explicitement. Par exemple :

Pierre discute avec Jean. L'un est enthousiaste, l'autre non.

Pour la référence mentionnelle, la dislocation à droite est impossible :

(*) Pierre discute avec Jean. L'un, de Pierre, est enthousiaste, l'autre non.

« *L'un* » et « l'autre » ont également un fonctionnement nominal, la mention antérieure introduit explicitement deux individus du même type, par exemple :

Il a deux voitures, l'une (de voiture) est en panne.

Il a deux enfants. L'un est marié et l'autre célibataire

LA RÉCURSIVITÉ DES GROUPES NOMINAUX

La récursivité est un phénomène par lequel un groupe de mots d'une certaine catégorie grammaticale peut se trouver à l'intérieur d'un groupe plus large de même catégorie, éventuellement plusieurs fois de suite. Un tel phénomène semble permettre à la grammaire d'engendrer une infinité de phrases, mais on se heurte vite aux limites de la mémoire et de la compréhension humaines, qui viennent contraindre les engendrements grammaticaux. On assiste donc à une conjonction et à une interaction de critères quant à l'acceptabilité des énoncés ainsi engendrés. La structure syntaxique d'une phrase fait apparaître la façon dont les constituants sont emboîtés les uns dans les autres.

L'ami de la fille de ma voisine mangeait une pomme.

Le GN sujet a 3 constituants immédiats :

le Déteminant[l'] + le Nom [ami] + le GP [de la fille de ma voisine]

Ici aussi le GN est récursif. Nous parlons de récursivité quand nous avons affaire aux emboîtements successifs de syntagmes de même catégorie. Ainsi, dans cet exemple, un autre type de syntagme est récursif : le groupe pronominal (GP), puisque le GP [de ma voisine] est emboîté dans le GP [de la fille de ma voisine], lui-même emboîté dans le GP [de l'ami de la fille de ma voisine].

Le mécanisme de récursivité existe dans de très nombreuses langues du monde, pour ne pas dire dans toutes, ce qui fait dire aux génératives que c'est un aspect de la Grammaire Universelle.

Du fait de la récursivité, la grammaire est un système fini qui engendre un nombre infini d'énoncés (ce qui ne signifie pas que tous seront acceptables, même s'ils sont conformes aux règles syntaxiques).

Dans la résolution automatique des anaphores nominales, la récursivité des groupes nominaux est un phénomène assez complexe à traiter, il implique le traitement des groupes nominaux par ordre privilégié, du GN le plus long au GN simple.

LES NOMS COMPOSÉS

Les noms composés sont formés par « deux ou plusieurs termes originairement distincts, mais qui se rencontrant fréquemment en syntaxe, au sein d'une phrase, se soudent en une unité absolue et difficilement analysable » (Gross, 1996). Par définition d'Apothéloz (Apothéloz, 2002) les noms composés sont formés par construction d'une unité lexicale complexe au moyen d'un morphème grammatical non affixal et d'un morphème lexical, ou d'au moins deux morphèmes lexicaux libres ou liés.

Selon Grévisse qui utilise des critères syntaxiques, le nom composé est différent d'un syntagme car un nom composé est une unité lexicale permanente alors que le syntagme est une forme libre occasionnelle (Grevisse, 1986).

En informatique, un mot est souvent défini comme une chaîne de caractères séparés par deux espaces. Comment permettre à la machine de reconnaître automatiquement les mots composés, c'est une de nos problématiques actuelles. Notre recherche limite ce sujet à l'identification des noms composés de métiers.

Un nom composé de métier est souvent formé à partir d'un nom de métier complet de base, en ajoutant un modifieur, par exemple : *Directeur commercial, directeur de service..*

Un nom de métier composé peut aussi être formé à partir d'un morphème lexical (technico, électro, télé…) en y ajoutant un autre morphème lexical, un nom de base, ou un mot grammatical (sous, vice…), par exemple : *Electro-mécanicien ; vice-président…*

Le modifieur ajouté sous-spécifie le contenu du travail que le nouveau nom de métier implique. Le modifieur ajouté peut être une séquence introduite par une préposition, il peut aussi être un adjectif ou plusieurs adjectifs. Parfois, le modifieur est une construction à partir d'un verbe au participe passé.

Par exemple : *Ingénieur électronique spécialisé en…*

Il existe également des noms de métiers composés qui sont formés par la combinaison de deux noms de métiers. La combinaison peut être réalisée en juxtaposant deux noms de métier par un espace, par une préposition ou même par une conjonction.

Par exemple : *Boulanger-pâtissier, charcutier-traiteur…*

LA SIGLAISON

Définie comme « La réduction d'un syntagme à la lettre initiale de chacun de ses composants» (Le Petit Robert 1990), la siglaison est une forme de troncation d'une seule forme graphique du mot, ou une abréviation, au sens large. Lorsque l'unité

monolexicale est représentée par la lettre initiale, on la dénomme sigle simple. Lorsqu'elle est représentée par la lettre initiale suivie d'une ou de plusieurs lettres de ce même mot, on la dénomme sigle composé. Si elle est représentée par la lettre initiale puis par des lettres choisies arbitrairement dans le mot, on la dénomme sigle acronymique. Ces procédés de troncation englobent également l'acronyme. Exemple :

*Je viens de signer un **CDI**, le premier **contrat à durée indéterminée** de ma vie.*

LE MODIFIEUR

Dans la grammaire transformationnelle dont la classification se fonde entièrement sur des critères formels, le groupe nominal est défini par la forme :

Det + N + modifieur

Le modifieur peut être un adjectif :

…un livre intéressant

un adjectif modifié par l'adverbe :

…un très beau livre

ou une série d'adjectifs (un *beau vieux livre*)

…un très beau livre intéressant

Le modifieur peut également être une proposition relative sur laquelle on peut ajouter une autre proposition relative et sur laquelle on peut poursuivre des ajouts sans limite :

…le vieux livre qu'il m'a prêté..

Anne Theissen (Theissen, 2001) s'intéresse au degré de fidélité des anaphores en insistant sur la concurrence entre un GN défini fidèle et un GN défini totalement fidèle obtenu avec le maintien du modifieur. Autrement dit, Theissen a créé la nouvelle notion de l'hyper fidélité ou fidélité pleine dans ses travaux en se concentrant sur le maintien ou non de l'adjectif dans un SN défini anaphorique basique, c'est-à-dire un référent de type « *Un* + *modificateur* + *N* » dont N est un nom de base et le modificateur est un adjectif ou un participe.

Exemple :

(a) *Nous avons vu* ***un petit chien*** *dans le jardin du voisin.* ***Le petit chien*** *est vraiment adorable.*

(b) Nous avons vu ***un petit chien*** *dans le jardin du voisin.* ***Le chien*** *est vraiment adorable.*

La structure dans l'exemple (a), même si elle est moins fréquente, peut être utilisée comme une forme d'insistance, pour deux raisons :

- Le maintien de l'adjectif permet de rendre le référent plus saillant au moment de la reprise.
- L'adjectif est considéré comme un terme de base dans le cas où la fidélité se limite à la tête lexicale.

Avec le modifieur maintenu, la « subjectivité » joue aussi un rôle important dans la recherche des antécédents. Le plus souvent, soit c'est le nom choisi qui a lui-même le sens principal, soit c'est le modifieur qui a ce sens. Le deuxième est le plus fréquent en réalité. Par exemple :

Martin *s'est présenté au commissariat, (ce, le)* ***jeune artisan*** *a été cambriolé hier.* (Le Pesant, 2002b)

En fait, dans cet exemple, outre la subjectivité relevée par le jugement de l'adjectif *jeune*, nous constatons également que la relation entre le syntagme nominal défini (***le jeune artisan***) et le nom propre (***Martin)*** est assez faible. Dans le traitement des

anaphores nominales en général, la relation entre un GN défini et un nom propre est souvent faible, soit parce que celui-ci n'a pas encore été évoqué et qu'il n'est donc pas présent dans la conscience des interlocuteurs, soit parce que dans une situation donnée il ne s'impose pas à leur attention par ses propriétés perceptives, ou encore parce qu'il ne fait pas l'objet de connaissance ou de représentations conceptuelles supposées partagées (Neveu, 2004)

2.2.3. LES PARTICULARITES SYNTAXIQUES DES ANAPHORES NOMINALES

LA POSITION SYNTAXIQUE DES ANTÉCÉDENTS

En principe, toutes les GN précédant une anaphore sont initialement considérées comme candidats antécédents potentiels. La sélection du bon antécédent dépend de nombreux paramètres syntaxiques.

Concernant l'anaphore associative, Kleiber (Kleiber, 1999b) souligne qu'une anaphore associative peut relever de la dimension cotextuelle car le pontage peut reposer sur *la structuration interne du lexique*. En effet, l'accessibilité des anaphores associatives s'appuie sur une relation économe entre l'expression anaphorique et l'antécédent. Par défaut, l'accès se fait sur le **dernier élément pertinent disponible en mémoire** à partir de l'état actuel de la représentation linéaire du discours par le récepteur. L'entité anaphorique se réfère le plus souvent au dernier antécédent. Par exemple :

*Je suis entré dans **la pièce**. **Le plafond** était très haut.*

En revanche, si l'annotation des anaphores est fondée sur le principe de la chaîne (ce principe de la chaîne ne concerne pas l'anaphore associative), plusieurs contraintes et préférences sont proposées pour éliminer les candidats antécédents inappropriés.

Pour le cas des anaphores nominales fidèles, la plupart des linguistes exigent que les anaphores et leurs antécédents soient accordés en genre et en nombre, mais cette condition n'est pas indispensable pour les anaphores infidèles et associatives.

Pour faciliter la résolution des anaphores autrement dit pour filtrer les candidats inacceptables dans la recherche de l'antécédent, il faut aussi compter les contraintes et les préférences en théorie syntaxique. Ingria et Stallard (Ingria and Stallard, 1989) ont proposé certaines préférences, entre autres, le parallélisme syntaxique et la théorie du focus.

Dans le traitement des anaphores nominales, cette préférence serait applicable aux candidats antécédents GN **ayant la même fonction syntaxique** que l'expression anaphorique.

Par exemple :

(a) *The programmer successfully combined **Prolog** with C, but he had combined **it** with Pascal last time.*

(b) *The programmer successfully combined Prolog with **C**, but he had combined Pascal with **it** last time.*

Dans l'exemple (a) *Prolog* a été choisi de préférence, car dans :

*combined **Prolog** with C*

*combined **it** with Pascal*

Prolog et *it* sont les compléments d'objet direct du verbe *combine* de la structure « *combine A with B* », ils se trouvent juste après le verbe *combine*

Dans l'exemple (b), C a été choisi de préférence, car :

*combined Prolog with **C***

*combined Pascal with **it***

C et *it* sont les compléments d'objet indirect du verbe *combine* dans la structure *« combine A with B »*. Ils se trouvent après le mot *with.*

Le parallélisme syntaxique pourrait être utile dans la résolution des anaphores pronominales. En revanche, ce n'est pas le cas pour les anaphores nominales.

Bien que les critères syntaxiques et sémantiques pour la sélection d'un antécédent soient très puissants, ils ne sont pas toujours suffisants face à un ensemble de candidats possibles. En outre, ils servent plutôt à éliminer les candidats inaptes qu'à proposer le candidat le plus probable. En cas d'ambiguïté de choix des antécédents, on choisi généralement l'élément le plus saillant, c'est-à-dire l'élément qui est le plus répété. En TALN, cet élément le plus saillant est appelé « focus » ou « centre » (Grosz et al., 1983).

Par exemple, pour la phrase suivante, machines ou humains, nul ne sera en mesure de résoudre le pronom anaphorique "cet objet ":

Jean a mis la théière sur une assiette et il a cassé ***cet objet.***

Toutefois, si cette phrase fait partie d'un segment de discours qui permet de déterminer l'élément le plus saillant, nous pouvons en déduire :

Jean a trouvé une très belle théière dans une boutique la semaine dernière, il voulait l'acheter mais il n'avait pas assez d'argent avec lui. Le lendemain, il est retourné à la boutique pour acheter la théière tant convoitée. Pourtant, une fois à la maison, Jean a mis la théière sur une assiette et il a cassé cet objet.

Dans ce discours, "la théière" est l'entité la plus saillante car ce mot a été répété à plusieurs reprises auparavant et est devenu ainsi le centre de l'attention ou le « focus ».

Il est clair que très souvent, lorsque deux ou plusieurs candidats concourent pour l'antécédent, la tâche de résoudre l'anaphore revient à la tâche de trouver le focus. Cependant, il faut reconnaître la nature intuitive de la tâche.

LES CHAMPS DE RECHERCHE DES ANTÉCÉDENTS

Le champ de recherche est un élément nécessaire dans la résolution des anaphores nominales car il assure à la fois l'efficacité du choix et la faisabilité du système.

Il faut tout d'abord parler du problème de la délimitation de la phrase. À l'écrit, la limite habituelle de la phrase est un signe de ponctuation : le point, le point d'exclamation, le point d'interrogation, les trois points de suspension, le double point. Mais l'usage du moins en français ne se limite pas en fin de phrase : il faut prendre en compte les abréviations et les acronymes qui comportent des points sans pour autant délimiter une phrase. En outre, il peut arriver que ce cadre formel ne coïncide pas avec la syntaxe. Deux cas peuvent alors se présenter : soit la syntaxe déborde du cadre de la phrase, soit celle-ci contient plusieurs syntaxes indépendantes.

Souvent, la recherche des antécédents n'a lieu que dans la phrase courante ou précédente. Pourtant, un système idéal de résolution des anaphores devrait avoir un champ de recherche plus large, car il existe en réalité des antécédents qui se trouvent bien plus loin.

Supposons que le champ de recherche soit spécifié, que les GN qui précèdent l'anaphore dans ce champ de recherche soient identifiés comme candidats et que le nombre de facteurs de résolution d'anaphores soit la base pour choisir l'antécédent correct. Une mauvaise spécification du champ de recherche peut apporter trop de bruits ou de silences.

LES RÈGLES D'ÉTIQUETAGE DES FONCTIONS SYNTAXIQUES POUR LES GROUPES NOMINAUX

Etant donné que tous les GN précédant une anaphore sont initialement considérés comme candidats potentiels pour des antécédents, il est très important de connaître les étiquettes des GN. Dans un travail mené par (Liang and Lin, 2005) sur l'utilisation

du Tagger pour la résolution des anaphores nominales dans la langue anglaise, même si on ne peut pas connaître uniquement les informations syntaxiques des GN, on peut tout de même déduire les étiquettes de rôle des GN, par exemple *Oblique, Objet direct, Objet indirect,* ou *Sujet* en utilisant ces règles suivantes :

Règle 1 : Préposition GN (Oblique)

Règle 2 : Verbe GN (Objet direct)

Règle 3 : Verbe [GN] + GN (Objet indirect)

Règle 4 : GN (Sujet) [",[^Verbe], " |Prep GN]* Verbe

Règle 5 : GN1 Conjonction GN2 (avoir le même rôle que GN1) Conjonction

Règle 6 : [Conjonction] GN1 (avoir le même rôle que GN2) Conjonction GN2

L'IDENTIFICATION DES RELATIONS HYPERONYMIQUES AVEC LA SYNTAXE

Pour les anaphores associatives de type collection-membre, l'identification des hyperonymes et des hyponymes devient une tâche importante. Parmi les études sur ce sujet, il faut compter les travaux de (Hearst, 1992). Ces études visent à identifier les structures marquant la relation d'hyperonymie-hyponymie pour repérer automatiquement les hyponymes.

Par exemple, la structure suivante :

NP such as {NP_1, NP_2 ... (and /or)} NP_n

(… a*uthors such as Herrick, Goldsmith, and Shakespeare …)*

permet d'identifier les hyponymes suivants :

Hyponym (NP_1, NP)
Hyponym (NP_2, NP)
Hyponym (NP_n, NP)

dont NP_1, NP_2 et NP_n sont les hyponymes et NP est l'hyperonyme correspondant.

Fondé sur les études en anglais, (Hearst, 1992) a proposé six structures permettant d'identifier automatiquement des hyponymes suivants :

Such NP as {NP,}* {(or/and)} NP

… such authors as Herrick, Goldsmith, and Shakespeare

NP {, NP}* {,} or other NP

Bruises, wounds, broken bones or others injuries …

NP {, NP}* {,} and other NP

... temples, treasuries, and other important civic buildings.

NP {,} including {NP ,}* {(or/and)}NP

All common-law countries, including Canada and England ...

NP {,} especially {NP ,}* {or] and} NP

... most European countries, especially France, England, and Spain.

(Condamines, 2002) s'intéresse aux marqueurs de relations de toute nature (lexicaux, syntaxiques, typographiques) qui permettent de repérer la relation conceptuelle entre des éléments nominaux. Par exemple, le type de structure suivant est un marqueur d'hyperonymie :

[tous les N1] **sauf** [les N2]

2.3. POINT DE VUE DE LA SEMANTIQUE

2.3.1. LA RELATION ANAPHORIQUE

Lorsqu'une relation linguistique s'établit entre deux unités lexicales qui partagent le même référent, il y a une relation anaphorique entre elles. Dans une relation anaphorique, l'interprétation complète de l'expression anaphorique dépend d'une

expression précédente du discours, l'antécédent. L'identification des relations entre deux unités lexicales constitue une problématique qui relève principalement de la sémantique lexicale, c'est-à-dire la relation de sens entre deux entités lexicales, et très peu de la sémantique textuelle (au niveau discursif).

Plusieurs linguistes ont travaillé sur les différents types de relation qui peuvent exister entre un « antécédent » et son anaphorique.

Selon la définition d'anaphore de (Milner, 1982), 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, par exemple :

Les gens *ne sont pas très malins.* ***Certains*** *sont même très bêtes.*

L'antécédent *Les gens* est repris partiellement par le mot *certains*

(Lerat, 1981) propose deux relations principales : soit l'anaphore et son antécédent ont **l'équivalence sémantique**, soit il existe entre ces éléments une **inclusion lexicale** qui concerne les cas d'hyperonymie. Pour le cas de l'équivalence sémantique, il propose six cas suivants :

- La répétition lexicale :

  Un étudiant *est entré dans le bureau ;* ***l'étudiant*** *a demandé l'heure*

- La synonymie :

  Le spectacle *a de quoi étonner.* ***La représentation*** *était formidable.*

- La nominalisation :

  Des spécialistes de la traduction automatique se sont reconvertis dans la linguistique. Cette reconversion...

  Le GN *reconversion* entretient une relation de forme et de sens avec le verbe *reconvertir* dans le contexte précédent.

- Le supplétisme : le supplétisme est appelé « anaphores atypiques » par (Apothéloz, 1995b). Il s'agit des cas où la source n'est pas identifiable sous

la forme d'un nom.

Je ne vide pas le lave-vaisselle. C'est la seule tâche que j'ai réussi à inculquer à mes filles.

- La périphrase synonymique :

Le dollar continue à progresser. On le cotait à Francfort à 1,8875 DM (contre 1,8810). La devise allemande a perdu sa place.

- Le sigle anaphorique :

Je saute les écoles primaires et secondaires pour en arriver à l'École Nationale d'Administration. L'ENA est à la mode.

Le cas de l'inclusion lexicale concerne l'hyperonymie, en particulier pour le cas d'anaphore infidèle, par exemple :

*On pourrait même parler de **Mercure**, qui ressemble à la **Lune**, de **Jupiter** et de ses satellites, de **Saturne**, de **Vénus**, planète que l'on commence à connaître, mais nous ne traiterons que de **la Lune** et de **Mars**. Si la géomorphologie s'intéresse à **ces astres**, c'est qu'elle s'attache à tout milieu qui est de la terre.*

L'expression anaphorique hyperonymique *ces astres* inclut toutes les autres séquences hyponymiques comme *Lune, Jupiter, Saturne, Vénus, Mars*…

(Condamines, 2005) a proposé d'autres relations entre une anaphore et son antécédent, comme :

- Les dérivés d'un adjectif :

*Vus à la loupe, les grains ont un aspect **mat** mais on sait que de micros concrétions engendrent aussi cette **matité.***

- Les dérivés d'un nom :

*Les quatre « grands » sont les **super-acheteurs** mondiaux des bons du Trésor américains. Via ces **achats**, et des taux d'intérêt relativement élevés, ils ont donné le feu vert à Washington.*

Le nom *achats* est formellement et sémantiquement apparenté au nom *super-acheteurs.*

Dans la résolution de l'anaphore associative, la relation entre expression anaphorique et son antécédent repose sur la présence explicite d'un terme potentialisant un pontage entre l'anaphorique et son antécédent (Kleiber, 2001). La relation entre l'entité antécédente et l'entité nouvelle n'est plus une association uniquement discursive ou contextuelle mais relève d'un savoir conventionnel. (Kleiber, 2001) précise que le passage d'une entité à une autre est assuré par quatre types de relations :

- Relations méronymiques : Il s'agit d'une relation de type partie-tout entre un substantif qui est défini comme un holonyme par rapport à d'autres substantifs définis comme ses méronymes.

Exemple :

*Il s'abrita sous **un vieux tilleul**. **Le tronc** était tout craquelé*

*Paul aime sa **voiture** parce que **les sièges** sont confortables*

*Les policiers inspectèrent **la voiture**. **Les roues** étaient pleines de boue*

Le savoir partagé est nécessaire à l'interprétation de ces phrases. Par un savoir conventionnel, nous savons tous qu'une voiture a des sièges et des roues et qu'un arbre (le tilleul) a un tronc. Ontologiquement, l'expression anaphorique apparaît comme étant subordonnée à l'expression antécédente, elles sont comme composantes ou parties de l'entité antécédente (Kleiber, 2001). De plus, la relation de l'holonyme vers le méronyme est toujours à sens unique : le tout puis les parties, et moyennant une contiguïté sémantique, selon Irène Tamba (Tamba, 1991).

- Relations du type membre-collection : il s'agit d'une relation où l'entité de l'expression anaphorique se trouve reliée à l'antécédent par une relation qui unit les éléments ou membres à un ensemble collectif qui les rassemble.
  ***Le régiment** a été défait. **Les soldats** n'ont pas eu le temps de combattre*

Le régiment est l'ensemble qui rassemble ses membres : *les soldats.* Ayant les mêmes propriétés hiérarchiques que les relations méronymiques, les relations du type membre-collection requièrent également une hiérarchie à sens unique : la collection puis les membres (Kleiber, 2001)

- Relations locatives : le lien entre les entités peut également tenir plus largement à la structuration sémiotique du lexique, par exemple :

  *Nous entrâmes dans **un village**. **L'église** était située sur une hauteur*

Dans cet exemple, on parle d'un village. Un village possède généralement une église. Cette caractéristique intéresse non seulement l'emploi de l'article défini mais aussi la cohésion du discours. L'anaphore associative entre *le village* et *l'église* décrit également un rapport méronymique entre les objets référents des syntagmes.

- Relations actancielles :

  Paul coupa le pain et posa le couteau

Dans cet exemple, l'antécédent est un prédicat (le verbe *couper*) et l'expression anaphorique correspond à un de ses arguments ou actants (*le couteau*). La relation actantielle est une relation associant un prédicat et son actant.

- Relations fonctionnelles :

*Paul s'est inscrit dans un **club de foot**. **Le président** lui a fait signer une licence pour deux ans.*

La fonction du *président* est de diriger le *club de foot*, il s'agit d'un savoir partagé. La deuxième entité *président* garde une certaine fonction par rapport à la première entité *club de foot.*

En réalité, toutes ces quatre relations ne sont qu'une question de prototypie, de stéréotypie ou d'holonymie (Kleiber, 2001), (Kleiber, 1992).

Pour l'anaphore de type associatif, la relation entre l'expression anaphorique et l'antécédent n'est pas établie par le discours mais elle est préconstruite et se joue à un niveau lexical. Ainsi, le mode de sélection de l'antécédent doit respecter certaines régularités reposant sur des structures génériques et préétablies de la langue comme la stéréotypie ou la connotation.

Les travaux de Kleiber sont plus ou moins centrés sur la relation partie-tout, ce qui explique également nos motivations dans le choix du traitement des anaphores associatives de ce type. Du point de vue de la sémantique lexicale, notre intérêt porte sur le statut de la relation partie-tout parmi les différentes sortes de relations lexicales, notamment la relation hyperonymie-hyponymie et la relation de synonymie qui sont des relations fondamentales pour deux autres sortes d'anaphores nominales : l'anaphore fidèle et l'anaphore infidèle.

La relation hyperonymique est une relation d'inclusion : B est inclus dans A. A est l'hyperonyme de B si et seulement si l'ensemble des êtres décrits par A inclut dans l'ensemble des êtres décrits par B. Par exemple : *animal* est l'hyperonyme de *chien* qui est lui-même l'hyperonyme de *labrador*, et inversement *labrador* est un hyponyme de *chien* qui est lui-même un hyponyme d'*animal*. Si c'est un labrador alors c'est un chien, si c'est un chien alors c'est un animal.

La relation synonymique est une relation d'identité : A = B. A et B sont synonymes si et seulement si l'ensemble des êtres décrits par A est le même que l'ensemble des êtres décrits par B ; deux signifiants sont synonymes lorsqu'ils ont un sens identique, lorsqu'ils ont une même dénotation. Il s'agit d'une relation de double implication : si c'est A alors c'est B et inversement si c'est B alors c'est A.

Du point de vue du lexique mental, notre intérêt porte sur la représentation des propriétés sémantiques rattachées à la relation méronymique (partie-tout) de telle sorte qu'il est possible de tisser des liens entre monde sensible et univers mental. La dimension référentielle du langage est un concept incontournable pour préciser

comment notre savoir linguistique traite l'articulation entre les holonymes et leurs méronymes.

Du point de vue de l'analyse textuelle, notre intérêt porte sur la façon dont la relation entre un holonyme et son méronyme contribue à établir une chaîne de référence. Des conditions précises doivent être respectées pour qu'une relation partie-tout fonctionne comme une anaphore (par exemple, il faut que l'expression anaphorique soit un groupe nominal défini).

2.3.2. LA CATÉGORIE LOGICO-SÉMANTIQUE ET SÉMANTICO-ÉNONCIATIVE DU LEXIQUE

Du point de vue de la théorie des trois fonctions primaires, il existe trois niveaux linguistiques : niveau logico-sémantique, niveau énonciatif, et niveau interprétatif. Dans toute phrase, les relations logico-sémantique sont assumées par la structure prédicats-arguments. La fonction actualisatrice concerne l'actualisation des prédicats et des arguments par la présence des éléments qui inscrivent la phrase dans une situation d'énonciation : le temps, l'espace, la modalité… (Buvet, 2009). Ces fonctions relèvent de la situation de communication ou, autrement dit, du niveau énonciatif de la phrase. Un autre niveau concerne l'interprétation de la phrase en contexte et en situation de communication, ce qui permet de rendre le sens des phrases dans certaines situations d'ambigüité.

NIVEAU LOCICO-SÉMANTIQUE

La question de l'anaphore nominale relève non seulement de la sémantique formelle, mais aussi de la logique. Une étude sémantique des groupes nominaux réalisée par (Beust and Nicolle, 1997) dégage les propriétés des expressions anaphoriques du

point de vue de la sémantique nominale. Ces propriétés relèvent de la perception de l'objet par les actants, perception qui sous-tend une catégorisation grâce à deux dimensions : l'identité sortale et l'identité individuelle de l'objet.

Selon Beust & Nicolle (Beust and Nicolle, 1997), si **l'identité individuelle** de l'objet est l'identité qui lui préexiste, c'est-à-dire les critères qui différencient un objet des autres objets de sa catégorie, **l'identité sortale** est construite en apportant la réponse à la question : pourquoi un objet tombe sous une catégorie et ne dépend pas uniquement d'une description formelle de celle-ci ? L'identité sortale de l'objet inclut les traits qui représentent les attributs de sa classe ainsi que les valeurs de ces attributs qui lui sont propres. Ainsi, l'identité sortale est créée par les processus de catégorisation dont la mise en commun dans le dialogue.

L'identité individuelle est une simple instanciation de l'identité sortale, elle inclut une « *dimension temporelle et est introduite par la facette ontogénique de l'objet. Elle y est rattachée mais elle perdure même si l'identité sortale est rompue car les traits de la dimension ontogénétique se conservent quoi qu'il arrive* » (Beust and Nicolle, 1997)

Dans la représentation, l'identité individuelle permet de considérer les objets dans des situations de récit et non plus uniquement dans des situations de discours. Identifier la dépendance sémantique entre une anaphore et son antécédent revient à se poser un problème sur l'identification des phénomènes de dépendance nominale au terme de leur identité sortale et individuelle.

Exemple :

Un chasseur *s'est blessé ce matin.* ***Le jeune homme*** *a été emmené à l'hôpital.*

Dans cette phrase, l'expression anaphorique nominale *le jeune homme* se réfère à *un chasseur,* mais on ne peut pas dire qu'il existe une relation sémantique d'ordre prototypique entre un chasseur et un jeune homme. On ne peut donc pas se fonder sur l'identité individuelle pour déduire la relation entre ces deux séquences. On a introduit un nouveau substantif pour un même objet, ce n'est plus l'évocation d'un

objet à travers un autre. L'objet est identifié par de nouveaux prédicats sortaux, et il invoque ainsi un enrichissement de l'identité sortale du référent.

Pierre Beust & Anne Nicolle proposent des classes d'éléments, chaque élément de ces classes inclut des traits sémantiques qu'on peut catégoriser en fonction de leur caractère informatif en rapport à la classe. Chaque lexème est différencié par un ensemble de traits permettant de les opposer deux à deux, par exemple :

"chaise" = /avec dossier/ + /sur pieds/ + /pour une personne/ + /pour s'asseoir/

"fauteuil" = /avec dossier/ + /sur pieds/ + /pour une personne/ + /pour s'asseoir/ + /avec bras/

L'unique trait /avec bras/ différencie ces deux sémèmes (le sème est un ensemble de traits).

Définir un trait hors contexte revient à mettre en évidence un jeu d'oppositions des traits sémantiques (Pierre Beust & Anne Nicolle, 1997).

Les modèles actuels d'intelligence artificielle appliqués à la sémantique n'arrivent pas encore à résoudre parfaitement le phénomène de la polysémie ou la métaphore, même si ces phénomènes occupent une place centrale dans les processus d'interprétation, puisque leurs traits peuvent prendre plusieurs significations en fonction de ce à quoi on l'oppose.

NIVEAU ÉNONCIATIF

Au niveau énonciatif, l'analyse de la structure des phrases et l'analyse de la distribution des éléments lexicaux sont parfois nécessaires pour interpréter une phrase. Ainsi, les contraintes du niveau énonciatif rendent difficile l'attribution des unités lexicales dans des classes d'objets.

Par exemple, dans les phrases :

*Elles lui ont fait des **coupures** avec des lames de rasoir.*

*Il avait des **coupures** sur ses deux épaules et des lésions sur la main gauche.*

Le premier GN *coupure* a une interprétation processive et le second GN a une interprétation stative (Buvet, 2011).

Il est nécessaire ainsi de subdiviser les classes sémantiques en catégories notionnelles, et la catégorisation reste parfois arbitraire. Les travaux de P.A. Buvet proposent une trentaine de catégories notionnelles de ce type. Par exemple, les classes désignant un ressenti psychologique centré sur l'intériorité d'un individu comme <Joie>, <Peur>, <Tristesse> sont rattachées à la catégorie <Affect>. Cette catégorie est rapportée ensuite dans la catégorie DESCRIPTION_SUBJECTIVE qui stipule les sentiments en rapport avec l'intériorité d'un être humain. Deux autres grandes catégories proposées par Buvet sont la DESCRIPTION_INTERINDIVIDUELLE qui prend en charge la relation entre deux êtres humains et la DESCRIPTION_OBJECTIVE qui prend en charge tous les autres prédicats.

2.3.3. DES PARTICULARITÉS SEMANTIQUES DES RELATIONS ANAPHORIQUES

En discours il est bien souvent difficile d'identifier un réel antécédent pour une expression anaphorique (quand l'anaphore réfère à un « générique » ou à un « équivalent »), ou de déterminer la relation entre ces deux entités.

Par exemple :

*La pression atmosphérique avait **été évaluée** il y a une trentaine d'années à 1/12 de celle de l'atmosphère terrestre. On a réduit **cette appréciation** car Mariner 4 a trouvé qu'elle équivalait à 6 millibars […]*

Il est difficile de caractériser la nature de la relation qui existe entre l'anaphorique et son antécédent : synonymie ou hyperonymie (Condamines, 2005).

Pierre Beust et Anne Nicolle ont montré que la relation entre l'anaphore et son antécédent dans une anaphore associative peut être partielle, mais il est difficile de préciser quelle est la relation entre deux entités, comme dans l'exemple :

***Le mardi** était pluvieux, **le lendemain** était ensoleillé.*

Une autre particularité a été remarquée par (Mathilde Salles, 2010) concernant l'usage des termes désignant les parties du corps et la différence entre une anaphore possessive et une anaphore associative dans l'usage de ces termes.

Selon Mathilde Salles, les anaphores associatives ne semblent pas acceptables pour les parties du corps :

*(*)**Jacques** est tombé du premier étage. **Les pieds** sont cassés.*

*J'ai fait tomber **la petite table**. **Les pieds** sont cassés.*

***Jacques** est tombé du premier étage. **Ses pieds** sont cassés.*

Pour les rendre valides dans l'anaphore associative, soit il est nécessaire d'employer un marqueur supplémentaire tel le connecteur « En effet », « Du coup », « De ce fait », « Ainsi », « Donc », par exemple :

*Jean a été étranglé. **Le cou** est en effet tout couvert de bleus.* (Kleiber et al., 1994)

soit d'employer la séquence qui s'interprète en termes de cause-conséquence :

*Après la chute, il n'a pas pu lever la main. (En effet) Il s'est tordu **le poignet***

Ainsi, l'anaphore possessive garde plus de valeur explicative que l'anaphore associative.

Après la chute, il n'a pas pu lever la main, ***son poignet a été tordu.***

Une autre particularité des anaphores nominales concerne le parallélisme sémantique (Ingria and Stallard, 1989). Le parallélisme sémantique est une préférence utile (bien plus que la préférence parallélisme syntaxique), mais nécessite une analyse automatique de rôle sémantique, ou la fonction des groupes nominaux dans la phrase (sujet, objet…). Le parallélisme sémantique stipule que les GN qui ont le même rôle sémantique que l'anaphore sont favorisés pour être son antécédent.

Par exemple :

Vincent *a prêté un CD à Julie.* ***Le jeune homme*** *a aussi prêté un livre à Marc*

Vincent a prêté un CD à ***Julie. La jeune fille*** *a aussi demandé un livre*

Cette caractéristique demande ainsi une vérification en accord de genre et de nombre de l'antécédent potentiel et de son anaphore.

3. LES PROJETS TAL SUR LE TRAITEMENT DES ANAPHORES NOMINALES

Du point de vue du TALN, la résolution des anaphores nominales consiste en l'application des méthodes et outils informatiques au traitement du phénomène des anaphores nominales; il s'agit de la recherche automatique des antécédents pour les expressions anaphoriques, opérée par des ordinateurs programmés. Cette discipline fait partie de la linguistique appliquée, c'est à dire qu'on utilise des compétences à la fois informatiques et linguistiques pour la résolution.

La résolution des anaphores est une tâche difficile et complexe, qui fait intervenir plusieurs niveaux d'interprétation : la morphologie, la syntaxe, la sémantique, et les connaissances extralinguistiques. Il existe plusieurs méthodes différentes: méthodes symboliques qui incluent les systèmes à base de règles, le calcul de la similarité ou les

systèmes de préférence et de contraintes ; les méthodes statistiques basées sur le calcul de la probabilité ou par les arbres de décision ; les méthodes cognitives ainsi que des théories linguistiques.

Les premiers travaux de TALN datent du milieu des années 60. Ils se focalisent en partie sur les aspects syntaxiques de la langue sans véritablement intégrer l'aspect informationnel contenu dans ces textes. Par la suite, les chercheurs se sont intéressés à un problème qui est encore d'actualité à savoir : comment récupérer des informations utiles à partir des textes écrits en langue naturelle ? Ainsi, les travaux sur l'anaphore, qui constituent un processus important du traitement de la langue, sont étudiés dès les années 70 avant qu'on s'oriente vers l'extraction profonde d'information vers la fin des années 80.

Comme le langage naturel est très riche, les phénomènes anaphoriques peuvent s'exprimer de différentes manières en fonction du type d'antécédent ou du type de reprise. Par conséquent, il existe plusieurs types d'anaphores que l'on peut classer, selon que la reprise est un pronom personnel, relatif, démonstratif, possessif, GN indéfini, GN indéfini etc., ou en fonction de la position de l'antécédent (intraphrastique ou interphrastique). La résolution de l'anaphore revient à trouver l'antécédent associé à une anaphore. La résolution d'anaphore est considérée comme un des sujets fondamentaux du TAL, de la linguistique et des sciences cognitives.

La majorité des recherches sur la résolution des anaphores s'est d'abord focalisées sur les anaphores pronominales, ou anaphores déclenchées par un pronom personnel, qui sont d'une part les plus simples à distinguer, et qui, d'autre part, utilisent peu de connaissance sémantique. Ce type d'anaphore concerne tous les pronoms à la troisième personne, et vise plus particulièrement les pronoms personnels, démonstratifs, réflexifs, possessifs… La plupart des systèmes de résolution des anaphores traitent les antécédents qui sont des syntagmes nominaux car identifier des antécédents d'autres formes comme les groupes verbaux, les propositions, les phrases ou même les paragraphes etc. est très compliqué. Les travaux de traitement des anaphores pronominales ont été abordés par Hobbs

(Hobbs, 1978) puis bien présentés avec Hirst (Hirst, 1981). Les connaissances sur le sujet sont le fruit des travaux de (Lappin and Leass, 1994) (Weissenbacher and Nazarenko, 2007), (Kennedy and Boguraev, 1996) ; (Baldwin, 1997)…

Un autre phénomène fréquemment traité jusqu'à maintenant concerne les anaphores nominales ou anaphores définies, portées par des groupes nominaux définis au lieu des pronoms. Les anaphores nominales permettent de créer des liens sémantiques entre les antécédents et les anaphores, dont les liens de méronymie/holonymie ou les liens d'hyperonymie/hyponymie.

Les anaphores nominales provoquent un vif intérêt dans la communauté scientifique des linguistes informaticiens puisqu'on trouve des travaux dès la fin des années 70 (Sidner, 1979), (Grosz et al., 1983) etc. puis du début des années 90 jusqu'à l'heure actuelle avec (Connolly et al., 1994), (Aone and Bennett, 1995), (Vieira and Poesio, 2000) (Soon, W.M. et al., 2001), (Ng and Cadie, 2002), (Markert and Nissim, 2005),(Garera and Yarowsky, 2006).

3.1. APPROCHES EXISTANTES

Les travaux sur la résolution des anaphores qui s'inscrivent dans la perspective du TALN portent généralement sur la théorie de l'accessibilité (la linguistique cognitive), la théorie du centrage ou du focus (linguistique formelle) ou bien font intervenir des calculs du poids de saillance (linguistique statistique). En revanche, ces travaux n'exploitent pas les nombreuses études en langue française sur les anaphores bien qu'elles aient permis d'accumuler une somme considérable de connaissances sur le sujet. C'est le cas, notamment, de la typologie des anaphores nominales issue des travaux de (Kleiber, 2001, 2001).

Plusieurs approches provenant du TALN ont été proposées afin de résoudre le phénomène de l'anaphore en général. On remarque globalement trois types

d'approche : les approches riches en connaissances ; les approches pauvres en connaissances; et les approches purement statistiques.

Les approches riches en connaissances utilisent non seulement les connaissances au niveau grammatical, syntaxique avec des règles linguistiques définies empiriquement, mais elles y ajoutent des informations au niveau discursif et des connaissances du monde. Ainsi, les dictionnaires encyclopédiques ou les dictionnaires de synonymes sont souvent utilisés dans cette approche, avec l'intervention humaine avant ou après le prétraitement.

A l'opposé, les approches pauvres en connaissance, ou approches syntaxiques telles que (Baldwin, 1997), (Mitkov, 1996) et (Poesio et al., 2010) ont cherché à minimiser l'apport de connaissances extérieures. Les dictionnaires électroniques, les informations grammaticales et syntaxiques sont souvent utilisés pour traiter les anaphores. L'intérêt de ce type de méthode est d'obtenir un bon niveau de résultats avec des méthodes simples qui reposent sur un nombre de traitements linguistiques limités.

Par ailleurs, les spécialistes de résolution des anaphores appliquent également différentes techniques d'apprentissage automatique fondées sur corpus, surtout à partir de (Aone and Bennett, 1995) pour les anaphores nominales.

Parmi les algorithmes pour la résolution d'anaphores, celui de Jerry Hobbs, qui fonctionne uniquement sur la base de critères syntaxiques est un précurseurs. Plus récemment, Shalom Lappin et Herbert J. Leass ont formulé un algorithme qui exploite des informations morpho-syntaxiques et sémantiques. Cette approche hybride a eu un certain succès et reste une des références dans le domaine. Le troisième algorithme que nous présenterons est celui de Ruslan Mitkov, qui repose sur des données d'entrée assez pauvres.

3.1.1. APPROCHES RICHES EN CONNAISSANCES

Les travaux sur la résolution des anaphores, dans leur grande majorité, ont d'abord porté sur les anaphores pronominales qui sont d'une part les plus simples à distinguer, et qui, d'autre part, utilisent peu de connaissance sémantiques, contrairement aux anaphores nominales.

Puisque l'objet d'étude dans cette recherche concerne uniquement la résolution des anaphores nominales, les traitements des anaphores pronominales n'ont pas été abordés dans le cadre de cette thèse. En revanche, les lectures faites concernant ce sujet pourront se révéler très utiles aux travaux de recherche liés aux anaphores nominales.

La première approche linguistique riche en connaissance performante est connue vers la fin des années 70. Hobbs (Hobbs, 1978) a été parmi les premiers à se pencher sur la résolution des anaphores en utilisant les informations syntaxiques et sémantiques, produites à partir de phrases manuellement annotées. Ayant pour objectif de traiter les pronoms personnels, il a élaboré un algorithme sur la base de contraintes syntaxiques, sémantiques et morphologiques de l'anglais.

L'algorithme prend en entrée un arbre syntaxique complet et correct qu'il parcourt à la recherche d'antécédents en leur appliquant diverses contraintes imposées. L'algorithme qu'il a mis au point repose sur l'ordre selon lequel l'arbre syntaxique de la phrase est parcouru, ainsi que sur les critères de sélection en termes de la compatibilité sémantique. Ensuite, son système de résolution opère en deux temps.

Premièrement, des templates (des fichiers modèles contenant des informations sur les différents éléments, événements ou entités) sont créés grâce notamment à l'application de patrons.

Dans un second temps, il s'agit de fusionner les templates en regroupant trois critères : la structure interne des syntagmes nominaux désignant les entités ; la proximité des entités; et la compatibilité des templates.

L'idée principale de son travail était de sélectionner le meilleur antécédent possible en appliquant certaines contraintes sur le pronom à résoudre et son antécédent probable, par exemple : les accords en genre ou en nombre ; la nature des antécédents nominaux (animés ou inanimés) ; la proximité en nombre de phrases entre l'antécédent et l'anaphore…

Et lorsqu'un pronom anaphorique est reconnu, la recherche de l'antécédent se fait en parcourant l'arbre syntaxique de la phrase selon un ordre prédéterminé. Au niveau intraphrastique, l'algorithme consiste en un parcours en largeur de gauche à droite avec une préférence pour l'antécédent le plus proche de l'anaphore. Un parcours en largeur est aussi effectué au niveau interphrastique, avec une préférence pour les sujets comme antécédents. En effectuant le parcours, l'algorithme fait l'inventaire des antécédents possibles, qu'il vérifie ensuite en appliquant des contraintes d'accord morphologique – traits de genre et nombre. Il applique également des contraintes syntaxiques, qui sont fondées sur la théorie du liage, par exemple : un pronom non réfléchi et son antécédent ne peuvent pas apparaître dans la même phrase simple ; L'antécédent d'un pronom doit précéder ou c-commander le pronom…

En ce qui concerne la résolution des anaphores nominales avec une méthode linguistique riche en connaissance, il faut aussi compter les travaux de Poibeau (Poibeau, 2005). Fondée sur l'utilisation des patrons lexico-syntaxique couplés avec les ressources lexicales disponibles, sa méthode propose de rendre explicite le lien existant entre une entité de catégorie *Personne* et le concept de *Fonction*. Les informations extraites à partir du corpus sont ensuite analysées et présentées sous la forme d'une base de connaissances permettant de décrire les différents types de liens entre les entités de catégorie *Personne* et les autres types d'entités. Sa chaîne de traitement se décompose en trois temps :

- L'annotation des syntagmes nominaux accompagnés par un nom propre. Les formes visées pendant cette phase de traitement sont les syntagmes nominaux définis, les appositions et les structures prédicatives, qui constituent une source d'information sur les personnes.

- L'extraction du contenu des annotations produites à l'étape précédente, sa transformation et sa structuration sous la forme d'un fichier XML, qui sert de base de connaissance pour les étapes ultérieures.
- La mise en œuvre de la stratégie de la résolution partielle des anaphores. Elle débute par une autre phase d'annotation. Les formes visées dans cette étape sont, en plus de celles reconnues précédemment : les noms propres seuls et les syntagmes nominaux non accompagnés par un nom propre.

3.1.2. APPROCHES PAUVRES EN CONNAISSANCES

Après le développement des approches linguistiques riches en connaissances, d'autres approches reposant uniquement sur les informations syntaxiques ont été appliquées à plus large échelle, on les appelle aussi les approches syntaxiques.

Depuis les années 90, pour la détermination de la saillance d'un antécédent, les travaux sur la résolution des anaphores se caractérisent par une volonté de réduire au maximum les ressources linguistiques utilisées par différents systèmes de bas-niveau. On utilise des marques morphologiques et relations syntaxiques élémentaires, et les systèmes, et non plus les connaissances d'ordre sémantique ou pragmatique, analyse syntaxique complète.

Un des algorithmes les plus connus et les plus classiques utilisant ces connaissances de bas niveau est celui présenté par (Lappin and Leass, 1994). L'algorithme de Lappin & Leass reste aujourd'hui encore une référence dans le domaine. L'approche de Lappin et Leass est proche de celle de Hobbs car elle aussi nécessite une analyse syntaxique complète du texte source, mais ce n'est pas la façon de parcourir l'arbre syntaxique de la phrase qui détermine quel est l'élément anaphorisé. Le but de leur travail est de proposer un algorithme, implémenté en Prolog, pour identifier des antécédents nominaux de pronoms de troisième personne et d'anaphores réflexives

et réciproques dans un corpus de manuels informatiques. Cet algorithme applique les représentations syntaxiques générées par un parseur grammatical et la mesure de la saillance dérivée de la structure syntaxique et le modèle dynamique simple. Il contient :

- Un filtre syntaxique et intraphrastique pour exclure les dépendances anaphoriques d'un pronom ou un groupe pronominal dans un groupe de syntaxe.
- Un filtre morphologique pour exclure les dépendances anaphoriques d'un pronom ou d'un GN causé par le désaccord en personne, nombre ou genre.
- Une procédure pour identifier le pléonasme.
- Une procédure pour attribuer les valeurs de saillances des GN selon leur rôle grammatical, le parallélisme de rôle grammatical, la fréquence des mentions, la proximité…
- Une procédure pour identifier les GN liés aux anaphores comme une classe équivalente pour laquelle la valeur de saillance globale est comptée comme le total des valeurs de saillance de chaque élément.
- Une procédure de décision pour sélectionner l'élément préféré d'une liste des candidats antécédents pour un pronom.

Les connaissances linguistiques dans le modèle de Lappin et Leass sont simplifiées davantage que dans les travaux précédents dans la littérature. Il se fonde toujours sur le calcul de saillance mais au lieu d'utiliser un analyseur morpho-syntaxique coûteux en temps et en argent, il utilise les techniques de surface pour le prétraitement des syntagmes nominaux. Les fonctions syntaxiques ainsi que les étiquettes grammaticales sont connues grâce à la distribution des mots, par exemple :

- L'observation des prépositions ou de la position des noms par rapport aux verbes peut permettre de deviner la fonction syntaxique d'un syntagme nominal.

- L'observation des accords des déterminants ou des mots suivant un nom (adjectifs, participes passés) peut permettre de déterminer ses genre et nombre.

De son côté, (Baldwin, 1997) a utilisé un analyseur automatique simple pour son système de traitement des anaphores nommé CogNIAC. Le principe de son algorithme est de minimiser l'utilisation de données syntaxiques et sémantiques, qui sont assez coûteuses.

Lors du prétraitement, le système procède à un simple découpage du texte en phrases, à l'étiquetage des catégories grammaticales des mots, et à une reconnaissance simple des syntagmes nominaux.

Par la suite, la reconnaissance de l'antécédent des éléments anaphoriques est réalisée grâce à quelques critères de sélection simples, par exemple : s'il y a qu'un seul antécédent compatible en genre et en nombre, il est choisi.

Dans un travail sur la résolution d'anaphores chapeautées par un nom propre, (Boudreau and Kittredge, 2006) ont développé un algorithme simple utilisant des connaissances linguistiques limitées. Le but de leur travail est de partitionner les expressions référentielles d'un corpus codé en XML en chaînes de références distinctes.

Leur algorithme n'utilise pas de dictionnaire ni d'analyseur syntaxique ou d'autres outils du TALN. L'algorithme est divisé en quatre étapes, selon les niveaux organisationnels linguistiques :

- L'identification des mots utiles.
- La construction des syntagmes à partir des indices laissés par l'étape précédente. Par exemple : les suites de noms propres, les pronoms, les syntagmes nominaux débutant par un déterminant et contenant un nom commun spécifique au domaine ainsi que certains syntagmes complexes.

- L'attribution approximative des fonctions syntaxiques repose sur la présence de prépositions ou de signes de ponctuation particuliers et sur la position d'un syntagme référentiel par rapport aux autres référentiels de la phrase.
- La pondération de ces facteurs pour l'algorithme pour limiter le nombre de syntagmes référentiels à traiter. Quelques facteurs pouvant influencer le choix d'un antécédent dans un algorithme de résolution d'anaphores sont relevés : la nature du déterminant et de la tête du syntagme ; la compatibilité des traits syntaxiques et sémantiques ; la fonction syntaxique (sujet > COD > COI > autre) ; la distance entre l'anaphore et l'antécédent ; la présence de réitération lexicale ; l'appartenance de la tête nominale au vocabulaire du domaine.

Nous devons également citer la méthode de Rémi Lavalley (Lavalley, 2012) dont les étapes de traitements ont procédé différemment :

- En fonction du genre et du nombre de l'anaphore, on ne souhaite garder comme antécédents candidats que les mots correspondants : par exemple, si l'anaphore est « *il* » on gardera les mots masculins singuliers, si l'anaphore est « *elle* » on gardera les mots féminins singuliers.
- On recherche ces antécédents parmi tous les mots depuis le début du texte jusqu'à la fin de la phrase contenant l'anaphore.

De même, un autre système pour la résolution des anaphores pronominales proposé par Nasukawa (Nasukawa, 1994) appelé « Indépendant de la connaissance externe » a été étudié. En effet, son approche est fondée sur les préférences en l'existence des modèles de colocation identiques dans le texte, la fréquence des répétitions des phrases précédentes et la position syntaxique.

Nasukawa utilise un dictionnaire de synonymes et accepte que le synonyme d'un candidat soit éligible lui-même. Par ailleurs, Nasukawa constate que le nombre important des GN ayant le même lemme que le candidat dans les phrases

précédentes du syntagme nominal peut être une indication de préférence pour sélectionner un antécédent.

En ce qui concerne la position syntaxique, il prend en considération mieux le rôle des sujets plutôt que celui des objets. Cependant, comme cette préférence sur le sujet exige une analyse syntaxique préalable, Nasukawa prend seulement en compte les préférences syntaxiques dans sa mise en œuvre finale, par exemple les candidats qui se trouvent dans la phrase la plus proche et le candidat le plus proche à partir du début de la même phrase.

Finalement, le choix du meilleur candidat est réalisé par le calcul de la saillance, avec des facteurs : la colocation, la fréquence ou les préférences syntaxiques…

3.1.3. APPROCHE STATISTIQUE

De manière générale, la linguistique apporte au traitement du TALN, une approche beaucoup plus qualitative que quantitative. Car si le traitement de la langue est fortement lié à la linguistique, il est traditionnellement très lié aux disciplines à forte composante mathématique.

Bien qu'on parle de mathématisation des sciences humaines, les études linguistiques éprouvent quelques difficultés lorsqu'elles doivent effectuer une étude qualitative sur des corpus volumineux. Cependant, les méthodes mathématiques donnent des résultats « relativement bons » sans nécessiter d'importants travaux d'analyse linguistique. C'est pourquoi, les méthodes statistiques ont trouvé une place importante dans le TALN, on les appelle aussi les méthodes par apprentissage automatique, méthodes quantitatives, ou méthodes numériques.

Le ministère américain de la défense a organisé pendant plusieurs années des « Message Understanding Conferences » (MUC) dont l'objectif consiste à faire réaliser par différentes entreprises des logiciels capables d'accomplir un certain nombre de tâches spécifiées dont la détection d'entités et la résolution des coréférences, l'extraction d'information. Lorsque des campagnes telles que MUC-6 et MUC-7 permettaient d'accéder à une quantité importante de données annotées, les spécialistes de résolution des anaphores commencent différents travaux utilisant les méthodes par apprentissage automatique. L'usage de méthodes statistiques pour la résolution des anaphores s'est ainsi répandu au milieu des années 90, à partir de (Aone and Bennett, 1995) pour les anaphores nominales ou (Ge et al 1998) pour les anaphores pronominales.

L'approche statistique est fondée sur le même principe : à partir des textes étiquetés manuellement, l'automate apprend à repérer et étiqueter correctement les anaphores. Le module de l'apprentissage est sauvegardé pour être utilisé dans les modules de reconnaissance. Pourtant, les spécialistes peuvent procéder selon deux approches : supervisées (Connolly et al., 1994) (Soon, W.M. et al., 2001) ou non supervisées (Cardie and Wagstaff, 1999).

L'apprentissage supervisé est une technique d'apprentissage automatique où l'on cherche à produire automatiquement des règles à partir d'une base de données d'apprentissage contenant des « exemples » déjà traités et validés (un corpus annoté par exemple). Cette approche nécessite un corpus d'apprentissage représentatif du phénomène étudié, par exemple : apprentissage de règles de désambiguïsation à partir d'un corpus spécialisé. Cette approche requiert également une annotation manuelle d'un grand nombre de données, ce qui entraine un coût important. Par exemple, pour l'annotation des entités nommées, on peut déduire les règles suivantes :

- Le contexte autour du mot courant :

"à", "dans", "vers", etc. précèdent plutôt des <lieu>

"Madame", "Ministre", etc. précèdent plutôt des <personne>

- La casse du mot courant :
  les noms propres impliquent souvent la majuscule.

Au contraire, l'apprentissage non supervisé est une technique de classement des données hétérogènes en sous-groupes selon certains critères (cette approche est utilisée par exemple dans le regroupement de textes similaires dans un corpus (qu'on appelle aussi clustering).

Pour Ng (Ng, 2010), les approches supervisées pour la résolution d'anaphores se répartissent en trois grandes classes :

- Une classe qui se charge de comparer une anaphore avec chacun des antécédents potentiels se trouvant dans les phrases précédentes et déterminer si le couple (anaphore et antécédent potentiel) est en relation anaphorique. La décision de classification est prise pour deux entités données sans tenir compte des précédentes décisions.
- Une classe s'assurant que l'expression anaphorique soit compatible avec l'antécédent potentiel.
- Une classe d'évaluation de la probabilité d'un antécédent d'être en relation anaphorique avec l'anaphore, afin de trouver le meilleur antécédent. L'antécédent potentiel ayant le meilleur rang est conservé.

(Aone and Bennett, 1995) a aussi construit un système automatique de résolution des anaphores fondé sur la méthode statistique. Ce système utilise comme ressources un corpus d'articles de journaux en japonais afin d'étiqueter les liens anaphoriques du texte. L'arbre de décision que les auteurs forment se fonde sur des vecteurs de caractéristiques pour les couples : anaphore - son antécédent potentiel. Ces caractéristiques peuvent concerner soit une anaphore, soit un antécédent, soit la relation du couple anaphore – antécédent potentiel.

(Aone and Bennett, 1995) utilise 66 fonctions d'entraînement, telles que la fonction lexicale, syntaxique, sémantique et des caractéristiques de position, et différentes méthodes de formation à l'aide de trois paramètres : les chaînes anaphoriques, l'identification du type anaphorique et les facteurs de confiance.

Ils ont construit six machines d'apprentissage automatique pour la résolution des anaphores, deux de ces machines d'apprentissage automatique montrent une précision proche de 90% pour une évaluation qui porte sur 1359 anaphores.

Afin d'appliquer des techniques d'apprentissage de la machine du type *non supervisé* pour la sélection de l'antécédent à la référence anaphorique, (Connolly et al., 1994) affirment que le meilleur problème défini adapté aux algorithmes d'apprentissage est le problème de classification. L'approche adoptée dans leur recherche est de ranger les candidats dans deux classes distinctes. La classification est définie sur une paire de candidats et une anaphore, où les classes correspondant sont à choisir parmi les candidats en tant que "meilleur" antécédent de l'anaphore. Les candidats seront analysés par paires successives, le classificateur trie par la suite le meilleur candidat entre ces deux candidats.

La classification est effectuée en fonction des valeurs discrètes. Les instances sont présentées en tant que vecteurs de valeurs d'attribut, où les attributs décrivent les propriétés de l'anaphore et les deux candidats, et les relations entre ces trois éléments. Une fois le classificateur construit, il est utilisé par l'algorithme de résolution de référence pour sélectionner le meilleur candidat en utilisant la stratégie suivante :

- Une instance initiale est réalisée en prenant l'anaphore et les deux premiers candidats.
- Cette instance est fournie au classificateur qui indique quel candidat est le meilleur.
- Le candidat «rejeté» est éliminé, et une autre instance est élaborée en comparant le candidat «gagnant» avec le candidat suivant.

- La nouvelle instance est maintenant fournie au classificateur qui sélectionne un nouveau gagnant entre ces deux candidats.
- Ce processus se poursuit jusqu'à ce que chaque candidat ait été examiné et le dernier vainqueur est choisi comme antécédent.

(Dagan and Itai, 1990) utilisent aussi une approche statistique pour résoudre les références du pronom " il " dans les phrases choisies au hasard dans un corpus, afin de lever l'ambiguïté des pronoms. Il s'agit d'une solution alternative à l'implémentation coûteuse des contraintes de connaissances. Ils utilisent des patterns de cooccurrences, les candidats pour les antécédents sont substitués à l'anaphore et seuls les candidats disponibles dans les modèles de cooccurrences fréquentes sont approuvés, par exemple, dans la phrase suivante extraite d'un corpus plus long :

They knew full well that the companies held tax money aside for collection later on the basis that the government said ***it*** *was going to collect* ***it.***

Il existe deux occurrences de *it* dans la phrase précédente. Le premier est le sujet de «collect » et le deuxième est son objet de la structure : *A collect B*

La méthode statistique analyse les trois candidats : "money", "collection" et "government". La table suivante va lister les patterns obtenus après la substitution de chaque candidat avec l'anaphore, et le nombre de leur répétition dans le corpus :

sujet-verbe	collection	collect	0
sujet-verbe	money	collect	5
sujet-verbe	government	collect	198
verbe-objet	collect	collection	0
verbe-objet	collect	money	149

verbe-objet	collect	government	0

Tableau 2 : Nombre des répétitions des GN dans le corpus

Après l'analyse statistique, *government* est préféré comme antécédent du premier *it* et *money* du second *it*. Avec cette méthode, Dagan & Itai obtient un score de 87% pour la résolution du pronom anaphorique *it*.

3.1.4. APPROCHE HYBRIDE

En toute logique, la méthode hybride, qui est la combinaison de méthodes linguistiques et statistiques, tire parti des atouts de ces deux méthodes de base. Le système de traitement peut commencer par l'apprentissage des règles automatiques puis un expert va réviser les résultats, ou inversement, un linguiste élabore des règles puis le système procède à l'apprentissage automatique. Ruslan Mitkov est un des pionniers dans la combinaison de méthodes linguistiques traditionnelles avec une nouvelle approche statistique (Mitkov, 1996). Le but de son travail est de faciliter l'implémentation tout en assurant un bon taux de réussite sur le traitement de manuels techniques.

Son modèle intègre des modules contenant différents types de connaissances : syntaxique, sémantique, domaine, discours et heuristiques, mais son système de traitement ne nécessite ni analyse syntaxique ni analyse sémantique, il prend simplement en entrée la sortie d'un étiqueteur morpho-syntaxique.

L'algorithme se déroule de la façon suivante :

D'abord, les groupes nominaux apparaissant à une distance au plus de deux phrases de l'anaphore à résoudre sont identifiés. Viennent ensuite les modules :

- Le module syntaxique prend en compte l'équivalent en nombre, genre et personne entre l'anaphore et son antécédent. Il vérifie également les contraintes C – Commande. En cas de parallélisme syntaxique, le module sélectionne le GN ayant la même fonction syntaxique que l'anaphore comme l'antécédent le plus probable.
- Le module sémantique vérifie la cohérence sémantique entre l'anaphore et l'antécédent possible. Il filtre les candidats sémantiquement incompatibles suivant la sémantique des verbes courants ou le statut du candidat et donne la préférence aux candidats GN ayant le même rôle sémantique que l'anaphore.

Les modules syntaxiques et sémantiques filtrent généralement les candidats possibles et ne proposent pas qu'un seul antécédent, c'est pourquoi les modules qui traitent le domaine de connaissance, heuristique, et module de discours sont ajoutés afin de choisir le meilleur candidat antécédent :

- Le module traitant le domaine de connaissance est une base de connaissances des concepts du domaine considéré.
- Le module de connaissances discursives avec un moteur statistique bayésien pour proposer le centre du segment discursif le plus probable, en se fondant sur la répétition d'expressions, la distance référentielle, la topologie lexicale du texte…

A chacun des candidats, un score est attribué et contribue à une somme totale. Le candidat avec le meilleur score total est sélectionné comme antécédent.

Le programme a été testé dans deux modes :

- Avec l'activation des modules syntaxiques, sémantiques et cognitifs.
- Avec l'activation des modules syntaxiques, sémantiques, cognitifs et discursifs. Ce mode a enregistré un meilleur résultat de (91.6%), qui justifie la performance de la combinaison des approches statistiques et les approches sémantico-syntaxique.

Les critères d'attribution du poids de saillance de l'algorithme de Mitkov sont résumés par Boudreau (2004) ainsi :

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

3.2. ALGORITHMES & CONCEPTION DES SYSTEMES DE RESOLUTION

En général, un système de traitement d'information typique est illustrée par :

- Un système d'analyse, qui permet une représentation ou une classification de l'entrée, comme : extraction d'information, identification de thème etc.
- Un système qui produit le langage de sortie : traduction, génération, système de résumé automatique.
- Un système interactif qui permet aux utilisateurs d'échanger des informations avec le système via une interaction multiple.

Ces systèmes peuvent être divisés en plusieurs classes :

- La segmentation qui comprend plusieurs niveaux : mots, phrases, discours ou article.
- Le marquage à différents niveaux de la description linguistique, qui comprend le marquage de la partie du discours, analyse morphologique, marquage de nom, marquage du sens des mots etc.
- L'extraction d'information, qui extrait les classes d'information spécifiques…

Les algorithmes de résolutions des anaphores sont classés parmi les systèmes d'analyse.

L'algorithme de (Mitkov, 1998a) utilise en prétraitement un simple balisage des syntagmes nominaux identifiés à partir de l'étiquetage du texte en parties du discours. Le texte est ensuite parcouru linéairement et chaque pronom anaphorique rencontré est traité. Les antécédents possibles sont les syntagmes nominaux apparaissant dans une fenêtre de trois phrases (celle où apparaissent le pronom à résoudre et les deux phrases précédentes). A chaque antécédent est affectée une saillance selon sa conformité aux critères suivants :

- L'antécédent est compatible en genre et en nombre
- L'antécédent est un syntagme défini
- L'antécédent est le thème de la phrase précédente
- L'antécédent est un syntagme nominal suivant un verbe indicateur
- L'anaphore est une réitération lexicale de l'antécédent
- L'anaphore est une expression qui reprend une partie du titre de la section
- L'antécédent n'est pas un complément indirect
- L'antécédent partage le même contexte que le pronom
- L'anaphore est dans une construction intrinsèquement anaphorique
- L'anaphore est proche de l'antécédent
- L'antécédent est un terme appartenant au sujet du texte
- L'antécédent obtenant la saillance la plus importante est choisi

L'algorithme de Lappin & Leass (Lappin and Leass, 1994), implémenté en Prolog, utilise un modèle qui calcule dynamiquement la saillance d'un antécédent potentiel sur la base de différents facteurs. A chaque facteur est attribué un indice différent selon son utilité dans la procédure de résolution. L'algorithme se déroule en trois étapes :

1, Le prétraitement :

- Le texte est tout d'abord découpé en phrases avec étiquetage des catégories grammaticales des mots. Vient ensuite l'identification des anaphores et des antécédents potentiels, qui sont les syntagmes nominaux de la phrase contenant le pronom anaphorique et/ou des phrases précédentes.
- Tous les antécédents potentiels qui forment une chaîne anaphorique sont regroupés dans des classes d'équivalence, les éléments déjà identifiés comme étant anaphores appartiennent à la même classe.
- Les pronoms non-anaphoriques, comme les pronoms pléonastiques sont éliminés.
- Ensuite, un filtre de contraintes morphologiques et syntaxiques élimine les candidats par certains critères morphologiques (compatibilité en genre et en nombre) ou positionnels (Théorie du liage). Par exemple, un pronom ne peut pas renvoyer à un nom s'il se trouve dans le même syntagme nominal que ce dernier.

2, La mesure de la saillance :

Les facteurs de saillance utilisés dans l'algorithme sont principalement des propriétés structurales ou syntaxiques. Chaque facteur permet, selon sa pertinence, d'augmenter le score des antécédents potentiels. Le poids de saillance est attribué selon les critères positionnels :

- Le poids de l'antécédent est divisé par deux pour chaque phrase séparant le pronom anaphorique de l'anaphorisant.

- Plus un antécédent est proche de l'anaphore, plus il est saillant.
- Les antécédents intraphrastiques ont la priorité sur ceux qui sont interphrastiques.

Les valeurs de saillance sont aussi attribuées selon les critères morphologiques (ou le rôle grammatical des GN) :

- Un antécédent en position de sujet est plus saillant, puis en position d'objet direct, puis d'objet indirect.
- Les arguments des verbes sont plus saillants que les adjoints et objets des GN jouant la fonction d'adjoint aux verbes.
- Une tête du groupe nominal (tête « substantif ») est plus saillante que d'autres éléments du GN, par exemple les compléments des têtes « substantif ».
- Un antécédent en position d'emphase est plus saillant : « c'est … qui »
- Le poids des candidats occupant la même fonction syntaxique que le pronom anaphorique sera majoré, alors que le poids des candidats fortement enchâssés sera minoré, selon le critère du parallélisme des fonctions.

3, La décision :

La mesure de saillance est utilisée afin de classer les candidats potentiels pour déterminer une préférence. Tous les antécédents potentiels sont pondérés et dans une classe d'équivalence, l'antécédent potentiel ayant le poids le plus important (antécédent le plus saillant) est choisi comme le référent d'un pronom. Dans les situations où deux ou plusieurs antécédents ont la même valeur de saillance (deux candidats également saillants), c'est celui qui est le plus proche du pronom qui l'emporte.

A, L'algorithme de Lavalley

L'algorithme de Rémi Lavalley (2012) est réalisé en deux versions. La première version parcourait chaque mot du texte et allait chercher son genre dans un dictionnaire de noms communs en ligne, et on le supprimait de la liste des candidats si c'était un prénom de genre opposé à l'anaphore. Si le mot commençait par une majuscule, on allait le chercher dans un dictionnaire de prénoms. A la fin de cette phase, il nous restait comme candidats les mots identifiés comme noms communs du même genre de l'anaphore et les mots commençant par une majuscule sans être des prénoms du genre opposé (étant donné que le nombre de noms propres connus était très faible, cela permettait de garder tous les noms propres inconnus comme noms d'entreprises, de pays etc.).

Evidemment cette méthode comportait de nombreuses failles, dues à :

- Le phénomène d'homographes (pas de différence entre « est » du verbe être et la direction « est »).
- Le mélange entre un nom commun et un nom propre (un mot de début de phrase était un candidat nom propre potentiel).
- La difficulté de retrouver certains noms féminins (par exemple « institutrice » qui est rangé à « instituteur, institutrice ».
- La gestion manuelle des pluriels.

Sa seconde version indique la classe et le genre de chaque expression anaphorique et tous les syntagmes nominaux d'une phrase. Il suffisait alors de passer le texte au système (qui s'appelle LiaTagg) et de ne récupérer comme antécédents candidats que les mots intéressants (noms de ville, pays, prénoms, noms de famille, noms communs) du même genre que l'anaphore.

Le problème de cette méthode s'explique par l'imperfection de l'analyseur. En effet, il lui arrive de se tromper et d'étiqueter les mots dans une mauvaise catégorie : la confusion entre un nom commun et un nom propre ; noms communs étiquetés comme adjectifs et n'apparaissant donc pas dans la liste des candidats.

De plus, l'analyseur ne connaît pas tous les mots que nous employons et il en étiquette un certain nombre comme « mot inconnu ». On décide donc soit de garder tous les mots inconnus (et potentiellement beaucoup de bruit), soit de les écarter tous (beaucoup de silence).

Pour rattraper ces erreurs, le système de Rémi Lavalley (2012) a appliqué un stoplist qui empêche un certain nombre de mots de devenir candidats. Viennent par la suite les méthodes de calcul de saillance qui se fondent sur la méthode de Lappin & Leass.

B, L'algorithme de Chaumartin

Le système de résolution d'anaphores dans une encyclopédie en langue anglaise, proposé par (Chaumartin, 2007) utilise la méthode linguistique pauvre en connaissance. Son système de traitement d'anaphores vise à extraire des connaissances d'une encyclopédie en langue anglaise, puis de les représenter sous forme de graphes conceptuels. L'un des modules utilisés dans la chaîne de traitement concerne la résolution d'anaphores et l'identification de chaînes de référence, il tire parti de caractéristiques qui facilitent leur analyse automatique : ils sont généralement correctement écrits; ils relatent des faits, avec des temps de verbe le plus souvent au passé ; les anaphores sont fréquentes et sont majoritairement pronominales, et portent le plus souvent sur le titre de l'article, c'est-à-dire son sujet. L'algorithme qu'il propose est celui-ci :

- Analyse du texte (au choix : simple étiquetage morphosyntaxique ; analyse syntaxique superficielle ; analyse syntaxique en profondeur)
- Parcours du texte : détection des pronoms personnels et possessifs, détermination du caractère « anaphorique » du pronom, par élimination des *it* pléonastiques et impersonnels.
- Pour les pronoms retenus comme « anaphoriques », le système marque tout d'abord les différents antécédents candidats, puis procède à la vérification des contraintes syntaxiques (notamment de c-commande, l'accord en genre et en

nombre..). Ensuite, le système applique différentes heuristiques qui augmentent ou diminuent la saillance de chaque candidat. Celui présentant la saillance la plus élevée est retenu.

- Extraction des antécédents en chaîne anaphorique.

C, L'algorithme de Liang et Lin

Dans une étude concernant le traitement d'un corpus du domaine biomédical, Tyne Liang et al. (Liang and Lin, 2005) a proposé son algorithme avec des règles d'attribution de la saillance. Le poids de saillance est fixé arbitrairement dans ce tableau, en fonction des caractéristiques de l'antécédent :

Caractéristiques de l'antécédent		Poids de saillance
1	La proximité par rapport à l'anaphore 0, s'il se trouve à la distance de deux phrases 1, s'il se trouve à la distance d'une phrase 2, s'il est intraphrase	0-2
2	Fonction sujet et objet	1
3	Même fonction grammaticale que l'anaphore	1
4	Accord en nombre	1
5	La séquence sémantique la plus longue	0 ou - 3
6	Accord en type sémantique	-1 ou - 2
7	Appartenance au domaine thématique	-2 ou 2

Tableau 3 : Poids de saillance proposés par Tyne Liang

Le premier trait caractéristique représente la mesure de la distance entre une anaphore et ses candidats antécédents en nombre de phrases. D'après les statistiques des deux corpus, la plupart des antécédents et leurs anaphores correspondantes se trouvent dans deux phrases maximum, ainsi, une taille de recherche pour trouver des candidats antérieurs est fixée à deux phrases dans le système proposé.

Concernant les rôles grammaticaux que joue une anaphore dans une phrase, comme de nombreuses anaphores sont des sujets ou des objets, les antécédents aux mêmes étiquettes grammaticales sont préférés. En outre, les candidats antécédents recevront plus de points de saillance s'ils sont en accord en rôles grammaticaux ou en nombre avec leurs anaphores.

Les autres traits sont liés à l'association sémantique, qui concerne le cas où l'anaphore est une variante sémantique de son antécédent et vice versa.

Leur algorithme est proposé en pseudo-code suivant :

S'il existe une correspondance au niveau lexical et sémantique entre l'antécédent et son anaphore

Alors score de saillance = score de saillance + 3

Sinon si tous les composants de l'antécédent (à part sa tête) sont correspondants avec son anaphore

Alors score de saillance = score de saillance + 2

Sinon si tous les composants de l'antécédent correspondent avec l'hyponyme de son anaphore, fourni par une analyse par WordNet 2.0

Alors score de saillance = score de saillance + 1

D, L'algorithme de Nouioua

Un travail sur la résolution d'anaphores proposé par **Farid Nouioua** (Nouioua, 2007) a été réalisé à l'aide des textes d'accidents de la route. La résolution des anaphores

nominales dans ce genre de texte aide à établir des responsabilités des participants. Par exemple :

"*Je roulais sur la partie droite de la chaussée quand* ***un chauffeur*** *arrivant en face dans le virage a été complètement déporté. Serrant à droite au maximum, je n'ai pu éviter* ***la voiture*** *qui arrivait à grande vitesse.*"

Il a proposé son algorithme avec les principes suivants :

- Le système part d'abord à la recherche du sujet qui est l'agent actif (je = le véhicule)
- Le système détermine la liste des anaphores à résoudre en gardant le nom central (chauffeur = le chauffeur du véhicule).
- Le système affecte à chaque anaphore une référence dans une liste prédéfinie : auteur, véhicule A, véhicule B, nom de la personne 1, nom de la personne 2… Ces constantes seront substituées aux arguments correspondants.
- On procède ensuite à un traitement sémantique en associant une information sémantique aux référents : véhicule = vélo, voiture, moto, camion…
- Une résolution immédiate peut être effectuée : je = ma, mon + GN = auteur du véhicule A.
- Si la liste disponible de candidats est vide, une nouvelle référence peut être éventuellement intégrée (procédure de propagation des références) : Affecter une référence nouvellement calculée pour une anaphore X à une autre anaphore Y non encore résolue si X et Y présente certain liens. La propagation est fondée sur les indices suivants :

  Un, une + GN -> nouvel élément, candidat éventuel existe

  Le, la + GN + qui/participe présent : nouvel élément existe.

  Adversaire, voisin = nouvelle référence

  Autre, premier + GN = nouvelle référence

- Le choix du meilleur candidat : si la liste des candidats est vide, on peut ajouter la nouvelle référence. Si la liste contient une référence, on peut prendre cette référence comme candidat probable. Si la liste contient plus d'une référence : on doit choisir le candidat le plus proche de l'anaphore avec moins de priorité à la référence « auteur » car on considère qu'il y a peu de chance que l'auteur se désigne par un référent.

Son algorithme est présenté avec le pseudo-code suivant :

Entrée : *liste de référents*

Sortie : *liste de références associées*

Début

Résoudre les cas triviaux

Propager les références et les natures

Pour *(tout référent* ***X*** *non résolu)* ***Faire***

Si *(un indice d'ajout d'une nouvelle référence se présente)* ***Alors***

Ref(X) = *nouvelle référence*

Sinon

Construire la liste ***ANT**** *des antécédents compatibles (* antécédent)*

Si *(**card**(**ANT**)=**0**)* ***Alors*** ***Ref(X)*** *= nouvelle référence* ***FinSi***

Si *(**card**(**ANT**)=**1**)* ***Alors*** ***Ref(X)*** *= le seul élément de* ***ANT*** ***FinSi***

Si *(**card**(**ANT**) > **1**)* ***Alors***

Ref(X) *= la référence la plus proche avec moins de priorité à la référence* ***auteur****.*

Fin Si

Fin Pour

Fin

E, L'algorithme de Mouton

C. Mouton (Mouton, 2007) a proposé une amélioration de l'algorithme de Lappin & Leass pour la résolution des anaphores pronominales. Son algorithme est divisé en trois étapes :

1. Filtre des antécédents et des pronoms anaphoriques : à l'aide des graphes à état fini
2. Filtre de meilleur antécédent avec seuil de poids de saillance :

- Chacun des antécédents potentiels va se voir attribuer une pondération de saillance correspondant à sa probabilité d'être sélectionné : le poids de saillance final qui comprend le poids de saillance global et le poids de saillance local (le tableau ci-dessous est un exemple).
- Parmi les candidats, on élimine ceux dont le poids de saillance final est trop bas (50 par exemple), on ne garde que le candidat dont le poids final est le plus élevé.
- Au cas où deux ou plusieurs candidats ont le même poids, on choisit le candidat le plus proche du pronom anaphorique.

3. Filtre morphosyntaxique : on compare le candidat et le pronom anaphorique selon plusieurs critères (genre, nombre etc.) pour ne garder que les antécédents pertinents.

Type de candidat	**Poids de saillance attribué**
Candidat est sujet	85
Candidat est attribut	70

Candidat est COD	60
Candidat est COI, complément d'agent, complément circonstanciel	40
Candidat est dans la proposition subordonnée	-75
Candidat est apposé	-35
Candidat est dans la phrase se terminant par « : »	50
Candidat est dans la phrase introduite par « : »	30
Candidat se trouve dans la phrase où se trouve le pronom anaphorique (S)	50
Candidat se trouve dans la phrase qui précède le pronom anaphorique (S-1)	40
Candidat se trouve dans la phrase avant (S-2)	30

Tableau 4 : Poids de saillance proposés par C. Mouton

3.3. RESULTATS OBTENUS

De nos jours, les approches par apprentissage automatique ont des résultats globalement meilleurs pour les anaphores nominales par rapport aux approches linguistiques, qu'elles soient riches ou pauvres en connaissance. Le moyen le plus utilisé pour évaluer les résultats est le calcul des taux de précision, taux de rappel et le F-mesure (généralement F1).

Les avantages des approches riches en connaissances sont qu'elles obtiennent souvent un bon taux de précision.

Tout d'abord, l'algorithme de résolution des anaphores pronominales de Hobbs (Hobbs, 1978), qui traite les pronoms personnels et qui se fonde sur la méthode linguistique traditionnelle et riche en connaissance, a obtenu un taux de réussite globale assez élevé, de l'ordre de 88.3% et atteint jusqu'à 92% lorsqu'il est complété avec des contraintes sélectionnelles simples.

Son algorithme est fondé sur diverses contraintes syntaxiques pour la recherche dans l'arbre syntaxique. La recherche est effectuée dans un ordre optimal de manière que la tête du syntagme nominal soit considérée comme l'antécédent probable. Un autre avantage du système de Hobbs est que le système est non seulement lisible mais encore évolutif car on peut modifier les règles ou ajouter les entrées dans les dictionnaires.

Cependant, l'algorithme de Hobbs montre qu'un dictionnaire est loin d'être exhaustif. De plus, il n'est pas possible d'établir tous les synonymes pour un groupe nominal sans prendre en compte le contexte. Aussi est-il évident qu'il ne peut pas résoudre les pronoms dans certaines constructions, telles que la reprise d'éléments phrastiques. L'algorithme effectue aussi une recherche pour des antécédents de pronoms cataphoriques, qui apparaissent après le pronom.

L'inconvénient de cette méthode s'explique aussi par le fait que la performance du système est instable en fonction des niveaux de langage utilisés dans les corpus. En

outre, les ressources construites manuellement sont coûteuses à construire et entretenir.

Il existe d'autres méthodes plus sophistiquées comme (Baldwin, 1997) mais du point de vue mathématique, l'algorithme de Hobb est moins coûteux et elle donne de bons résultats.

Les travaux antérieurs sur les résolutions des anaphores ont utilisé les connaissances lexicales de différentes manières. (Poesio et al., 2010) se sont fondés sur une analyse de WordNet pour traiter notamment les anaphores indirectes et associatives, et leur système apporte un taux de rappel de 35% pour 12 cas de synonymie, 56 % pour les 13 cas d'hyperonymie, et 38% pour les 11 cas de méronymie. (Poesio et al.) ont également examiné la structure syntaxique pour l'acquisition de connaissances lexicales, et les meilleurs résultats ont été constatés pour le cas des méronymie, avec 66 % de rappel.

(Gasperin and Vieira, 2004) ont testé la résolution des anaphores indirectes avec l'utilisation de listes de mots similaires. Ils ont effectué deux expériences concernant les noms (33 % de rappel) et les noms propres (22 % de rappel).

(Markert and Nissim, 2005) ont présenté deux façons d'obtenir des connaissances lexicales pour la sélection des antécédents définitifs (anaphore directe et indirecte). Ils ont atteint 71 % de rappel en utilisant une méthode fondée sur le Web et 65 % en utilisant une méthode fondée sur WordNet.

Le système CogNIAC de (Baldwin, 1997) résout les liens anaphoriques avec une précision importante (supérieure à 90%) et un rappel bien moindre (environ 60%), ce qui peut s'avérer suffisant selon les applications visées.

(Lappin and Leass, 1994) proposent un algorithme en plusieurs étapes, nécessitant une analyse syntaxique. Leur système a obtenu une bonne couverture du système de résolution et revendique une précision de 86% sur un corpus technique en anglais (avec une analyse syntaxique corrigée manuellement). Leur système a été testé sur les

textes annotés manuellement contenant 360 occurrences de pronom, et cet algorithme arrive à les reconnaître avec un taux de réussite de 86%, 4% de plus que l'algorithme de Hobbs sur le même corpus.

Pourtant, l'inconvénient du système de Lappin et Leass est la présence de diverses données d'entrée et de nombreux niveaux d'analyse. Leur algorithme fonctionne avec une analyse syntaxique et morphologique, et nécessite aussi une analyse des rôles sémantiques des GN dans le texte. Cela rend une implémentation assez coûteuse en temps et en argent.

Nasukawa (Nasukawa, 1994) utilise, comme corpus d'évaluation, 1904 phrases consécutives, avec au total 112 pronoms de la troisième personne à partir de huit chapitres de deux manuels informatiques. Son algorithme prend en charge le pronom «*it*» et permet de choisir un antécédent correct dans 93,8 % des cas. Ses méthodes donnent un bon taux de rappel, même sur les textes bruités. Cependant, on observe trois inconvénients principaux dans sa méthode : la phase d'apprentissage requiert un important volume de textes étiquetés ; le système est peu lisible pour les non-spécialistes ; il n'est pas toujours possible d'intervenir sur les résultats après coup.

L'algorithme d'une approche « robuste et pauvre en connaissance » de Mitkov a été conçu pour contourner certains des problèmes cités (Mitkov, 1998b). Fondée sur plusieurs heuristiques qui combinent une analyse syntaxique superficielle et une implémentation statistique cette approche donne une précision de l'ordre de 90% à 91% sur un corpus de manuels techniques (en trois langues : anglais, polonais et arabe).

Pour conclure, le résultat obtenu est très variable en fonction de la méthode utilisée, de l'objet étudié, du type de corpus, de la langue, du type d'anaphore… Notre étude va se fonder sur le modèle de données du laboratoire LDI, la théorie des trois fonctions primaires, pour la résolution des anaphores nominales de deux types :

anaphores associatives et anaphores infidèles en utilisant des corpus en langue française aspirés sur l'internet, dans la perspective du TALN.

CHAPITRE 3. METHODOLOGIE

Dans ce chapitre, nous allons décrire nos stratégies adoptées pour la résolution des anaphores nominales. Précisément, nous présenterons d'abord la description globale de notre système de résolution des anaphores nominales, qui se compose de trois modules : module de prétraitement du corpus, module de résolution des anaphores nominales et le module d'évaluation. Ensuite, nous présenterons les principes de traitement du lexique pour notre étude, en nous fondant sur les analyses au niveau linguistique des anaphores nominales. Enfin, nous aborderons dans cette partie la présentation de la plateforme Unitex, une des principales ressources que nous utilisons pour notre recherche. Nous allons aussi expliquer pour quelle raison nous avons fait ce choix

1. NOTRE POSITION

Pour pouvoir mettre notre système de résolution des anaphores nominales en pratique, nous sommes amenés tout d'abord à avoir une vue globale sur ce phénomène, qui consiste en l'explicitation et à la formalisation du fonctionnement des anaphores nominales. Notre démarche scientifique consiste en différentes étapes :

1- Observation du phénomène des anaphores nominales

2- Hypothèses sur le fonctionnement et l'organisation des anaphores nominales

3- Test des différentes hypothèses pour calculer leurs implications respectives et pour comparer avec le phénomène observé.

Nos études introduisent de façon plus cruciale l'étape du test des hypothèses qui doit faire partie intégrante de la démarche scientifique. Elle permet en traitement automatique d'avoir des procédures très rigoureuses pour tester les hypothèses émises sur le fonctionnement et l'organisation des anaphores nominales.

Ce sont notre corpus et nos ressources à disposition qui ont conditionné le choix de notre méthode. Nous avons à notre disposition un corpus aspiré sur le web. Nous utilisons aussi un outil d'analyse syntaxique libre, qui est Unitex. Pour des raisons pratiques, nous avons choisi de suivre les méthodes pauvres en connaissances (*knowledge-poor*) en nous reposant sur des critères de saillance. La notion de saillance est fondée sur la distribution des éléments concernés et sur les informations syntaxiques et sémantiques de ces éléments. Après le calcul de la saillance, l'élément qui a plus de saillance sera mis à jour pour vérifier ou établir leur relation anaphorique avec un élément analysé comme expression anaphorique.

De plus, nos corpus étant assez volumineux, nous devons pour calculer ces scores diviser les corpus en plusieurs échantillons. Le calcul des scores dépend des résultats des échantillons.

Pour construire notre système de résolution automatique des anaphores nominales, nous suivrons une méthodologie d'investigation faisant appel à des compétences pluridisciplinaires :

- Analyse linguistique des corpus annotés
- Modélisation de l'anaphore nominale du type fidèle-infidèle et associatif.
- Développement des modèles informatisés de résolution des anaphores.
- Évaluation.

Après la constitution du corpus et l'annotation des GN à l'aide d'Unitex, notre système sera divisé en trois modules principaux : module de prétraitement du corpus, module de résolution des anaphores nominales et module d'évaluation :

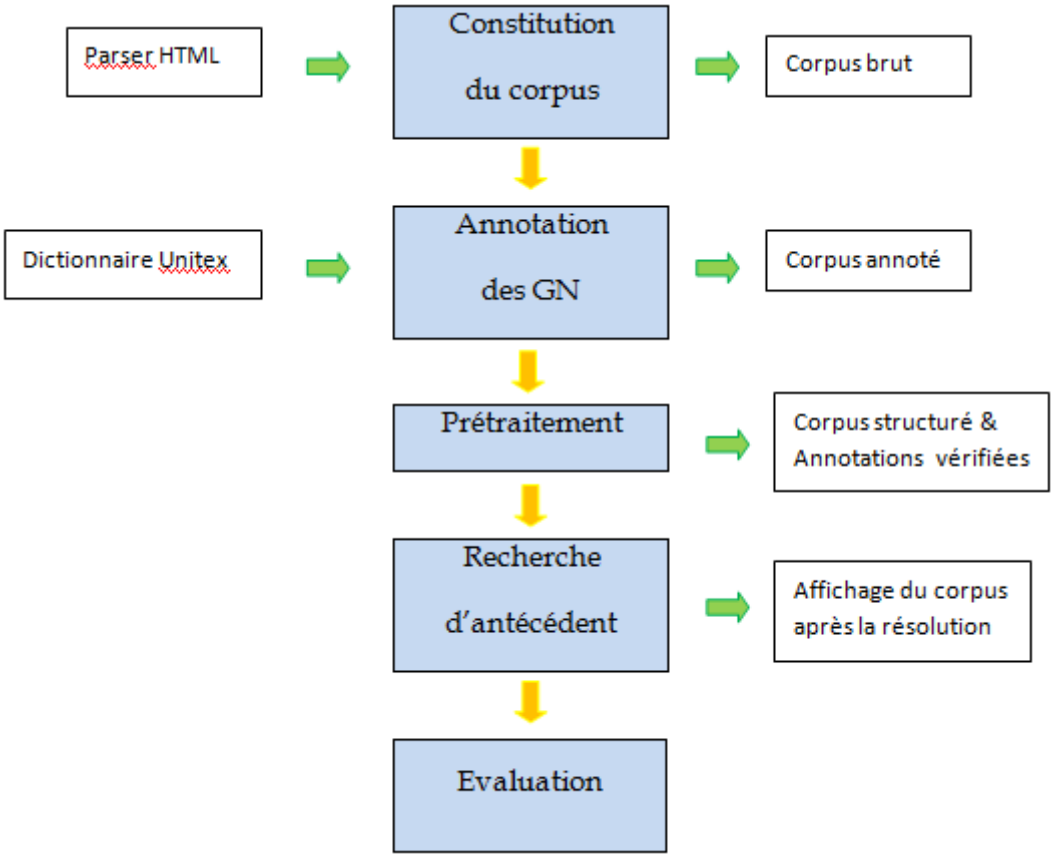


Figure 5 : Les modules de la résolution des anaphores

Nous expliquerons dans ce chapitre la méthode suivie pour l'annotation des GN (expressions anaphoriques et les antécédents candidats), puis les méthodes de sélection des meilleurs candidats, ou les stratégies pour apparier une anaphore à son antécédent. Nous mesurerons également l'impact de chaîne de référence sur la cohésion des textes. Nous modéliserons les résultats en termes de poids sémantique et nous étudierons les interactions entre les poids sémantiques alloués aux différents types d'anaphores nominales et ceux alloués à d'autres paramètres textuels afin de contribuer au calcul global du poids sémantique des textes. Ces modules de résolution d'anaphore sont développés afin d'expérimenter les conditions formelles que nous avions proposées. Nous évaluerons également les résultats fournis par le module de résolution d'anaphores. Nous proposerons un protocole qui comparera nos travaux à d'autres travaux. Nous examinerons exhaustivement les différences entre les résultats obtenus et nous préciserons quels sont les facteurs qui expliquent les différences observées.

1.1. MODULE DE PRÉTRAITEMENT

Le module de prétraitement utilise un corpus de texte brut non structuré en entrée. Après une étape de reconnaissance du lexique, le corpus sera traité pour devenir un corpus structuré.

Le module de prétraitement utilise des techniques variées :

- Aspiration des textes depuis les pages web en utilisant un parseur spécifique : Le parseur parcourt le contenu html, puis il ne prend que des informations qui nous intéressent pour notre étude : les informations métalinguistiques comme le nom de l'auteur, la date du texte, le titre etc. et le contenu textuel qui compose la partie principale du corpus : le corps de l'article s'il s'agit des articles journalistiques, les avis des internautes s'il s'agit des forums.

- Traitement du texte brut obtenu à l'étape précédente par Unitex : la segmentation en phrases (avec les étiquettes {S}) ; le repérage des groupes nominaux désignant les classes concernées à l'aide des dictionnaires Unitex et des graphes spécifiques. Les dictionnaires contiennent également toutes les informations concernant le genre, le nombre, les informations sémantiques équivalentes (classe, sous-classe, hyper-classe, domaines…)
- Vérification et correction éventuelles des annotations résultant de l'étape précédente en se fondant sur la distribution des GN dans le contexte phrastique, notamment avec la présence du déterminant (DET) précédant le GN.

1.1.1. CONSTITUTION DU CORPUS

Nous avons chosi de produire nous-mêmes des corpus provenant du web. Nous sommes alors face à une multitude de choix possible entre différents types de corpus : spécialisé, non spécialisé, moyennement spécialisé et entre différentes sources : les blogs, les articles de journal en ligne, les fiches de présentation de produits, les commentaires des internautes sur un sujet proposé, les modes d'emploi ou manuel d'utilisation d'un produit, des extraits littéraires, des textes de forum, des revues, les discours politiques…

Dans notre étude, nous utilisons comme support non seulement des textes standardisés comme des articles de journal, des présentations de produits mais aussi des supports non standardisés comme les commentaires extraits des forums en ligne. De plus, les corpus de ce genre sont souvent assez courts et riches en informations, le temps de traitement devient ainsi assez rapide par rapport au corpus annoté.

Nous avons choisi quatre corpus sur quatre thèmes différents : les échanges entre les acheteurs et les fournisseurs (après-vente) ; les présentations des produits innovants, les articles journalistiques sur les fait-divers et les avis des internautes sur une liste de produits de consommation. Ces sources sont toutes datées et classées selon un ordre chronologique déterminé (du plus récent au plus ancien), mais seuls le titre et les contenus des textes nous intéressent au niveau de l'analyse et du traitement, car ils contiennent les informations linguistiques.

Pour les anaphores nominales associatives méronymiques (relation du type partie-tout), nous choisissons de travailler sur un thème qui concerne l'avis des consommateurs sur des produits électroménagers ou des produits hi-tech, fort présent en veille technologique. Les produits objet des discussion sont des artefacts. Nous avons remarqué que la présence des paires : *artefacts-parties* est assez régulière dans ce type de corpus, la classe <artefact> est donc intéresssant.

Pour les anaphores nominales du type infidèle, nous avons remarqué que les faits-divers des journaux en ligne nous conviennent le plus… Les faits-divers concernent, le plus souvent, des affaires entre individues ou les enquêtes policières. Pour cela, nous avons élaboré une classe <personne>.

Après collecte des informations sur les noms des pages qui nous intéressent (URL), nous utilisons un parseur de Python (BeautifulSoup parser) pour parcourir le contenu des pages web et pour aspirer les informations qui nous intéressent comme le titre, le sous-titre, la date et les parties textuelles concernées.

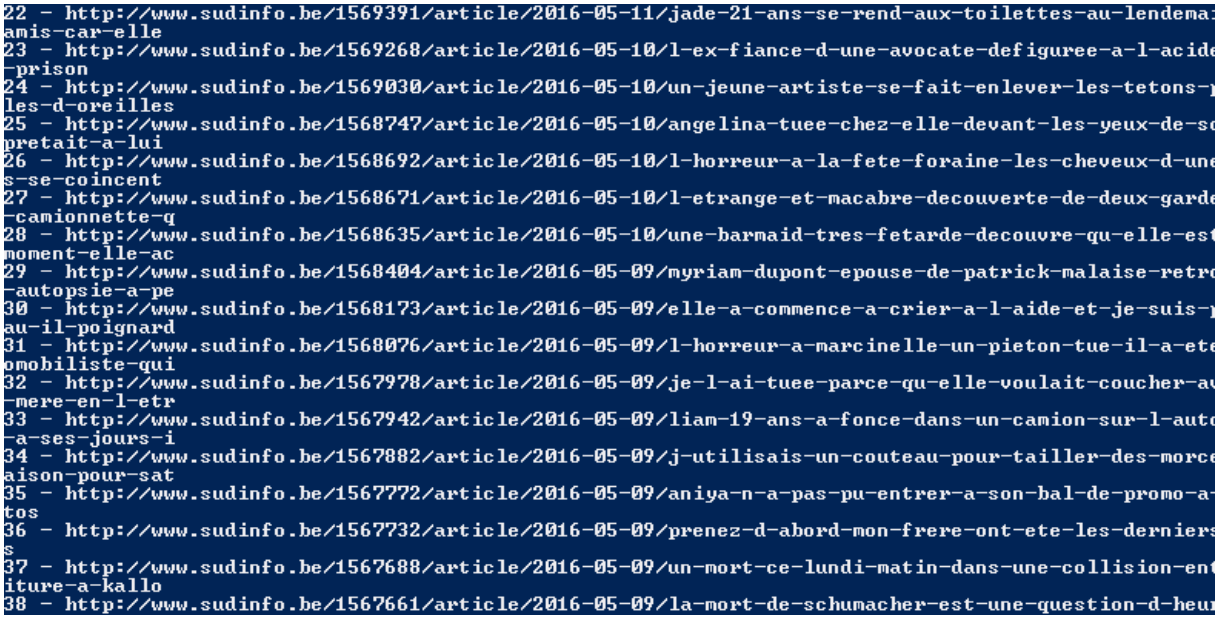


Figure 6 : Aspiration d'un corpus

Une fois le corpus aspiré, l'étape de normalisation du corpus intervient ensuite. A cette étape, le contenu aspiré (qui contient beaucoup d'espaces de trop ou beaucoup de caractères indésirables…) sera mis sous une forme plus standardisée, ou plus « propre » avec la vérification de l'encodage; la suppression des espaces ou retours de ligne de trop ; la suppression des doublons de texte, des publicités ; la séparation et le marquage visible des début et fin des articles, des parties des commentaires. Cette étape n'améliore pas le traitement du système de résolution des anaphores, mais elle facilite la vérification dans les étapes suivantes et l'affichage des résultats …

```
---------------------------
Titre : Trois « enfants sauvages » découverts à La Courneuve
Comment quatre enfants déclarés à l'état civil ont-ils pu passer à trave
Il a fallu la naissance du quatrième enfant d'un couple d'origine indien
Selon une source judiciaire, aucun signe de la présence d'enfant en bas
« PRIVATION DE SOINS PAR ASCENDANT »
Selon le conseil général de Seine-Saint-Denis, l'alerte a été donnée par
Les deux enfants les plus âgés, qui présentent des troubles majeurs du d
Fin janvier, la sûreté départementale est saisie de l'enquête. Mis en ex
28 000 NAISSANCES DÉCLARÉES EN SEINE-SAINT-DENIS
Jeudi, le Défenseur des droits, Dominique Baudis, s'est saisi de l'affai
Or une naissance dans une maternité accroît les chances de repérer les d
Chaque année, 28 000 naissances sont déclarées dans le département. Selo
« ON NE PARLE PAS UNE LANGUE QU'ON N'A PAS APPRISE »
Des cas tels que celui découvert à La Courneuve sont exceptionnels, même
Lire aussi :  Enfants maltraités : deux morts par jour
Si elle reste rare en France, la maltraitance des enfants est un phénomè
Le passé des parents joue également un rôle. « S'ils n'ont pas connu l'a
---------------------------
Titre : Suicide d'une cadre de La Poste sur son lieu de travail
Une cadre de la Poste s'est suicidée par pendaison sur son lieu de trava
« Nous venons d'apprendre qu'une postière, cadre supérieure de Coliposte
« C'est avec une profonde émotion que La Poste a appris hier le décès br
La femme, quinquagénaire, a été retrouvée pendue dans une partie non occ
---------------------------
Titre : Découverte de traces d'ADN des disparues de Perpignan
La piste criminelle s'est étoffée dans l'affaire des disparues de Perpig
Si les sources interrogées par l'AFP appelaient lundi à la prudence, des
ADN DANS UN LAVE-LINGE
```

Figure 7 : Exemple d'un corpus normalisé

A ce stade, la constitution des corpus nous a permis de recueillir quatre corpus dont les informations concrètes sont résumées dans le tableau suivant :

Type Anaphore	Anaphore Possessive	Anaphore infidèle (synonymie)	A. Associative membre-collection partie-tout	Anaphore infidèle (synonymie)
Type de corpus	Forum - Réclamations des consommateurs	Journal en ligne - Articles de fait-divers	Commentaires - Avis des usagers de voitures	Publicité - Présentation de produits
Thème	Artéfacts - hitech	Enquêtes policiers	Equipement voiture	Produits innovants
Taille (lignes)	1250 lignes	3000 lignes	2000 lignes	4300 lignes
Taille (page)	120 pages	130 pages	120 pages	170 pages
Taille (autres)	190 avis	294 articles	370 commentaires 70.000 mots	280 produits
Source	http://www.60millions-mag.com/forum/commerce-en-ligne	http://www.sudinfo.be/12/actualite/faits-divers	http://www.caradisiac.com/voiture--citadine/avis/	http://www.lsa-conso.fr/univers-produits/innovations

Figure 8 : Les corpus utilisés

1.1.2. ANNOTATION DES UNITÉS LEXICALES

Après avoir obtenu les corpus bruts normalisés, nous allons procéder à l'étape concernant l'annotation des unités lexicales de ce corpus. Les unités lexicales qui nous intéressent ici sont les GN.

Dans l'historique des approches de traitement des anaphores nominales, les ressources lexicales « artisanales » (ou manuelles) sont devenues indispensables. Notre système de résolution d'anaphores nominales dépend aussi fortement d'une quantité importante des connaissances manuelles.

Notre état de l'art montre que les études utilisant des analyseurs morpho-syntaxiques existants sont nombreuses. Les analyseurs morpho-syntaxiques gratuits pour l'annotation de texte avec les étiquettes morphosyntaxe et lemmalisation sont nombreux : à part les outils indépendants, dont les plus connus pour le traitement des corpus en français sont Treetagger et Standford, nous pouvons avoir aussi des outils intégrés comme la bibliothèque NLTK de Python. Pourtant, nous n'avons choisi aucun outil pour les annotations des unités lexicales de nos corpus pour plusieurs raisons.

Tout d'abord, les informations de partie du discours fournies par Treetagger ou Standford pour la langue française sont assez limitées. En effet, avec Treetagger et Standford, les étiquettes GN (qui nous intéressent le plus) sont simplement annotées en *NOM*, elles ne nous permettent pas d'obtenir d'autres informations plus détaillées comme le genre, le nombre du GN que notre travail requiert.

De plus, notre algorithme ne demande pas une annotation complète de toutes les unités lexicales (qui vont alourdir le programme), mais nous privilégions surtout les GN qui sont en relation avec notre thème privilégié.

La dernière raison : l'outil Unitex que nous allons utiliser contient aussi un analyseur morphosyntaxique. Même s'il a aussi ses défauts, comme tous les autres analyseurs,

Unitex nous permet, en plus, d'inclure les informations sémantiques au moyen des dictionnaires que nous pouvons construire nous-mêmes.

Code	Signification
A	adjectif
ADV	adverbe
CONJC	conjonction de coordination
CONJS	conjonction de subordination
DET	déterminant
INTJ	interjection
N	nom
PREP	préposition
PRO	pronom
V	verbe

Figure 9 : Les codes grammaticaux usuels d'Unitex

```
Nous  PRO:PER     nous
avons VER:pres    avoir
découvert   VER:pper    découvrir
deux  NUM   deux
corps NOM   corps
dans  PRP   dans
le    DET:ART     le
domicile    NOM   domicile
familial    ADJ   familial
.     SENT  .
Il    PRO:PER     il
semblerait VER:cond    sembler
que   KON   que
le    DET:ART     le
mari  NOM   mari
ait   VER:subp    avoir
tiré  VER:pper    tirer
sur   PRP   sur
sa    DET:POS     son
femme NOM   femme
avec  PRP   avec
un    DET:ART     un
fusil NOM   fusil
```

```
J'    NAM   <unknown>
ai    VER:pres    avoir
acheté      VER:pper    acheter
une   DET:ART     un
caméra      NOM   caméra
go-pro      NOM   <unknown>
le    DET:ART     le
22    NUM   @card@
octobre     NOM   octobre
2015.J'envoie     VER:subp    <unknown>
ma    DET:POS     mon
camera      VER:futu    camer
GO    NAM   <unknown>
PRO   NAM   <unknown>
en    PRP   en
réparation NOM   réparation
.     SENT  .
On    PRO:PER     on
me    PRO:PER     me
tel   PRO:DEM     tel
et    KON   et
me    PRO:PER     me
propose     VER:pres    proposer
un    DET:ART     un
```

Figure 10 : Analyse morphosyntaxique des corpus avec Treetagger

La reconnaissance des GN est réalisée à l'aide des graphes à état fini d'Unitex, qui font ensuite appel aux dictionnaires internes de l'outil et aux dictionnaires que nous construisons. Nos dictionnaires construits portent sur différentes classes selon le thème traité, par exemple les classes de <personne> (ou <humain>) pour traiter les

anaphores nominales infidèles, les classes d'<artefact> pour traiter les anaphores nominales associatives du type partie-tout… Pour chaque classe, nous élaborons des sous-classes correspondantes. Par exemple, pour la classe <artefact>, nous avons subdivisé en huit sous classes, et pour la classe <humain>, nous avons subdivisé en 16 sous-classes, en fonction du type de noms, du genre et du nombre des noms de chaque classe.

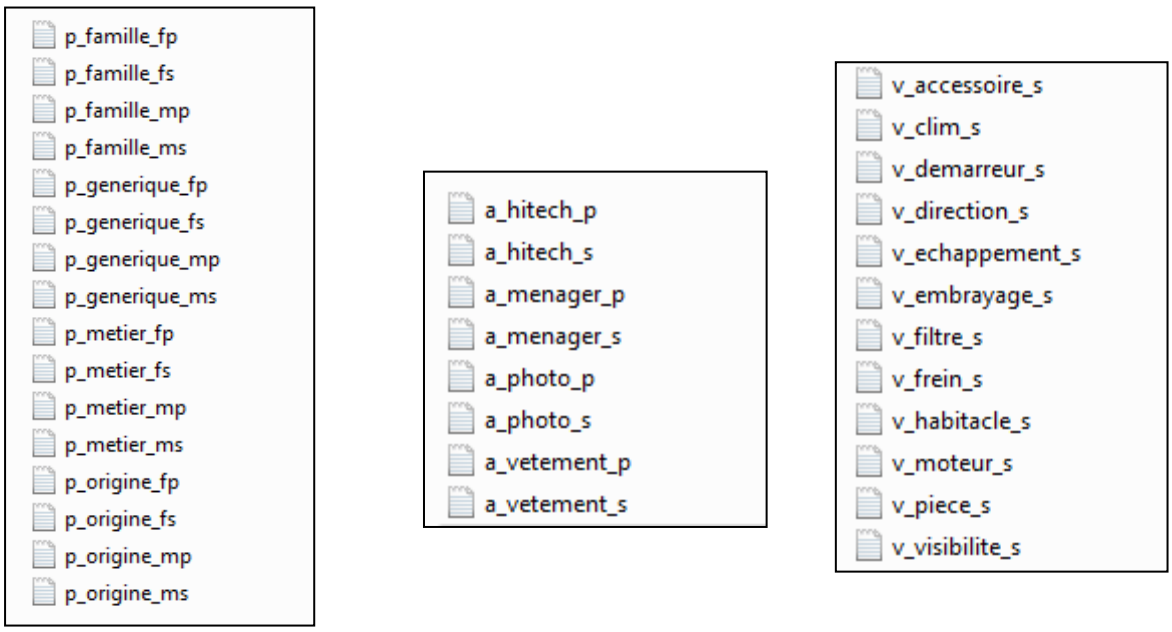


Figure 11 : Les dictionnaires utilisés

LA CONSTITUTION DES DICTIONNAIRES

Pour construire nos dictionnaires, nous cherchons tout d'abord à aspirer automatiquement des noms communs de base pour les différentes classes visées, de la même technique d'aspiration du corpus. Par exemple, pour la classe d'<artefact>, nous allons sur les sites de vente en ligne pour trouver les noms des appareils. Nous n'avons pas besoin, ou plus précisément, nous n'avons pas de moyen d'acquérir la liste exhaustive d'<artefact>.

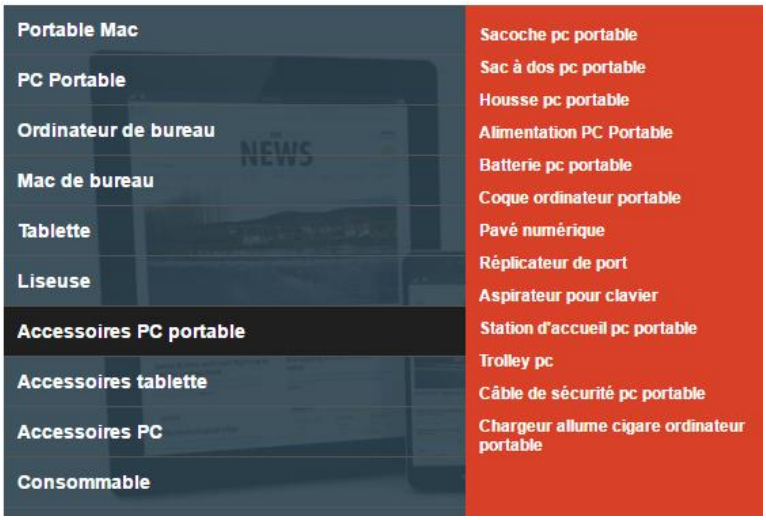


Figure 12 : Aspiration des noms d'<artefact>

Après l'acquisition de la liste de base des classes telle <artefact>, <humain>, …, nous procédons à une reconnaissance des unités lexicales en créant un dictionnaire et des graphes à états finis de l'outil Unitex.

Notre dictionnaire électronique contient des entrées dont la structure est :

montre,.N+ARTEFACT+H=APPAREIL+C=APPAREIL_SIGNALISATION+D1=TECHNIQUES+D2=VIE_QUOTIDIENNE+D3=_

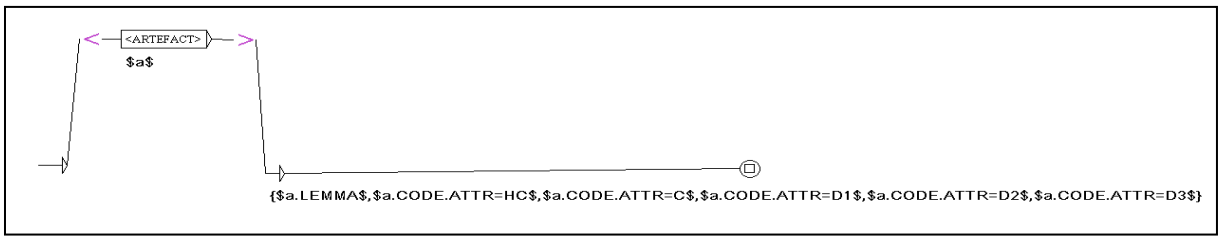


Figure 13 : Graphe pour annoter des étiquettes grammaticales et sémantiques des GN

La liste des noms communs sera enrichie au fur et à mesure à cette étape. Pour enrichir un dictionnaire des noms communs, nous nous appuyons sur la méthode supervisée pour acquérir automatiquement d'autres noms en utilisant des patrons syntaxiques, produits après une analyse de la distribution des unités lexicales dans le corpus (contexte droite et gauche des unités lexicales repérées).

Par exemple, la partie suivante :

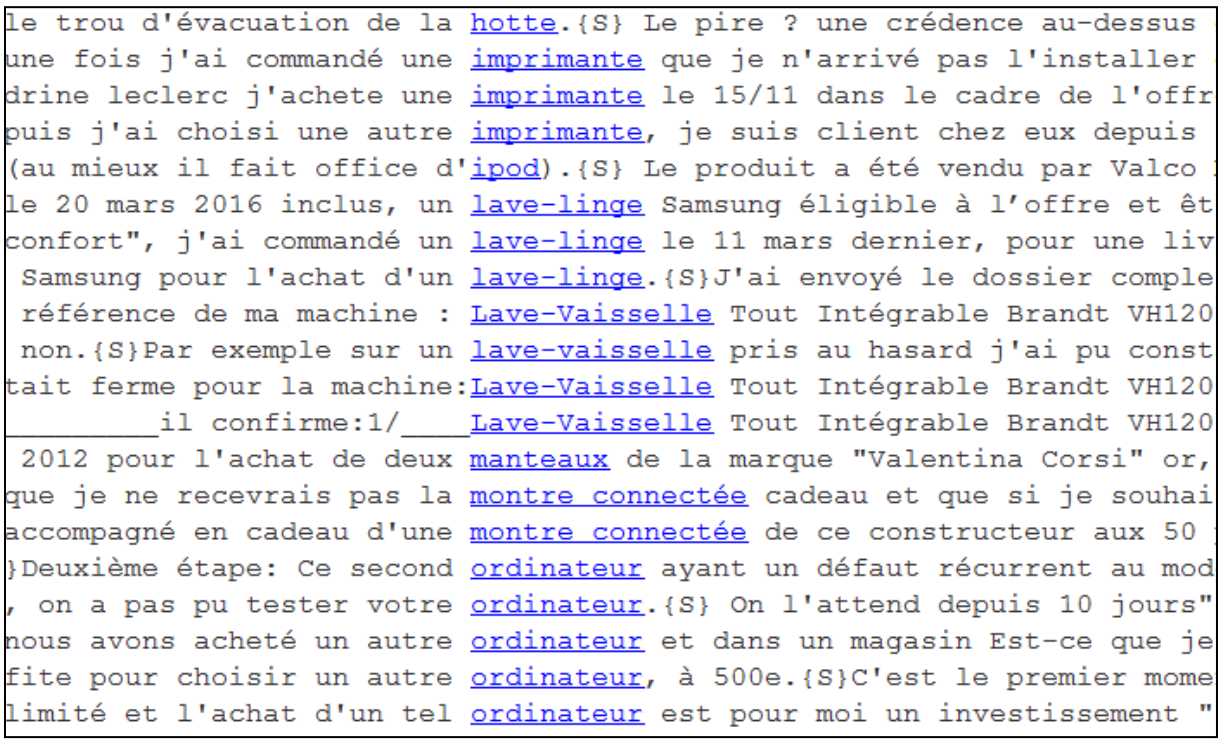


Figure 14 : Analyse de la distribution des unités lexicales dans le corpus

nous permet de construire les patrons comme :

Commander un <artefact>

Acheter un <artefact> (la forme nominale est *Achat d'un <artefact>*)

Choisir un <artefact>

Recevoir un <artefact>

Tester un <artefact>

<artefact> de marque + nom de marque

…

Ces patrons syntaxiques permettent d'acquérir de nouvelles entrées lexicales dans les corpus existants pour enrichir le dictionnaire.

Par exemple, pour le patron *commander un <artefact>*, nous obtenons une nouvelle liste d'<artefact> :

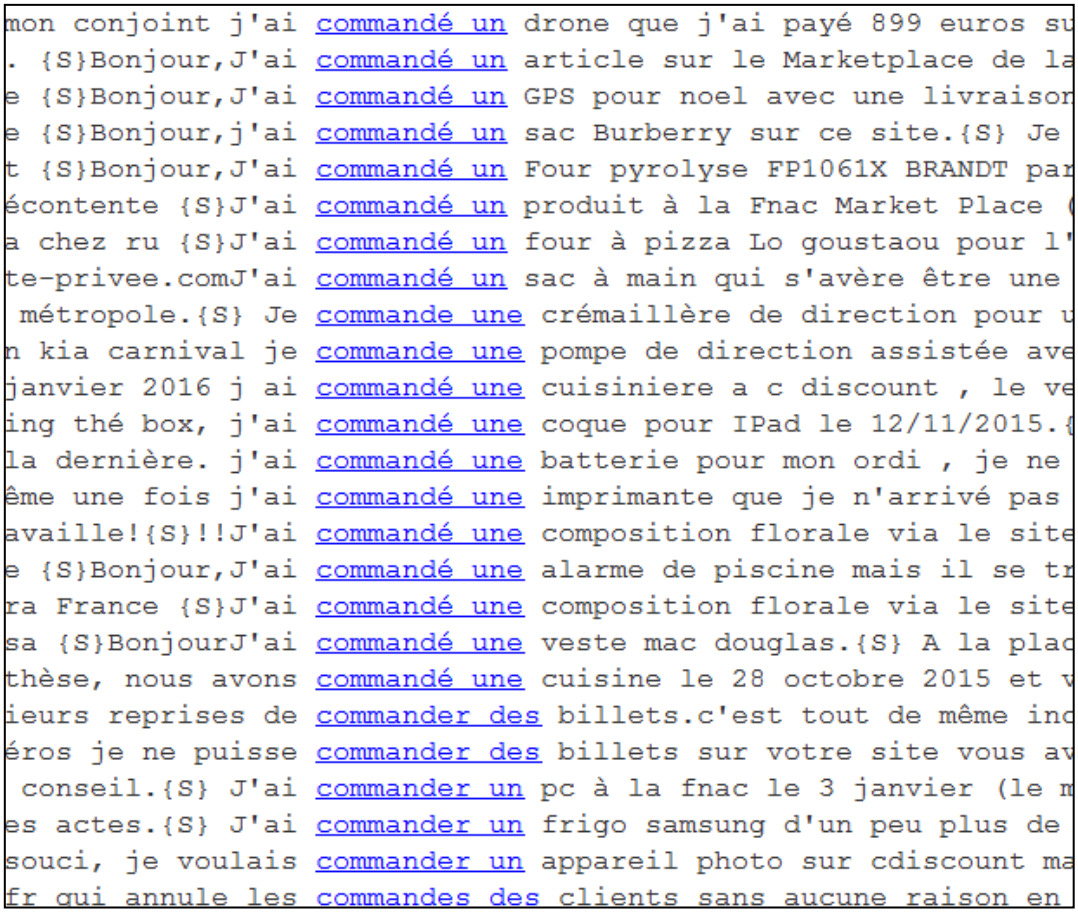


Figure 15 : Application du patron « commander un <artefact> » pour chercher de nouvelles entrées

Pour obtenir la liste de ces nouvelles entrées, nous utilisons un module qui traite des expressions régulières de Python, ainsi, la liste des nouvelles entrées est générée automatiquement. Avant d'être insérées dans les dictionnaires équivalents, ces entrées nécessitent une analyse complète, qui requiert une intervention humaine, et le classement se fait :

- Par famille de produit, par exemple *médecin* sera mis dans le dictionnaire des noms de <métier> et non pas le dictionnaire de <famille>

- Par les informations concernant le nombre et le genre : les entrées dans les dictionnaires sont caractérisées par leur nombre et leur genre. Ainsi, le mot *médecins* sera mis avec les noms pluriel et masculin. A ce stade, les noms communs dont le genre masculin et féminin qui sont écrits de façon identique (par exemple le mot *enfant)* sont classés masculins par défaut (le fonctionnement d'Unitex oblige à faire un choix)

Les étapes de constitution et d'enrichissement des dictionnaires sont réalisées les unes après les autres.

Une fois terminée la constitution des dictionnaires nous utilisons encore les graphes à états finis d'Unitex pour la reconnaissance des unités lexicales.

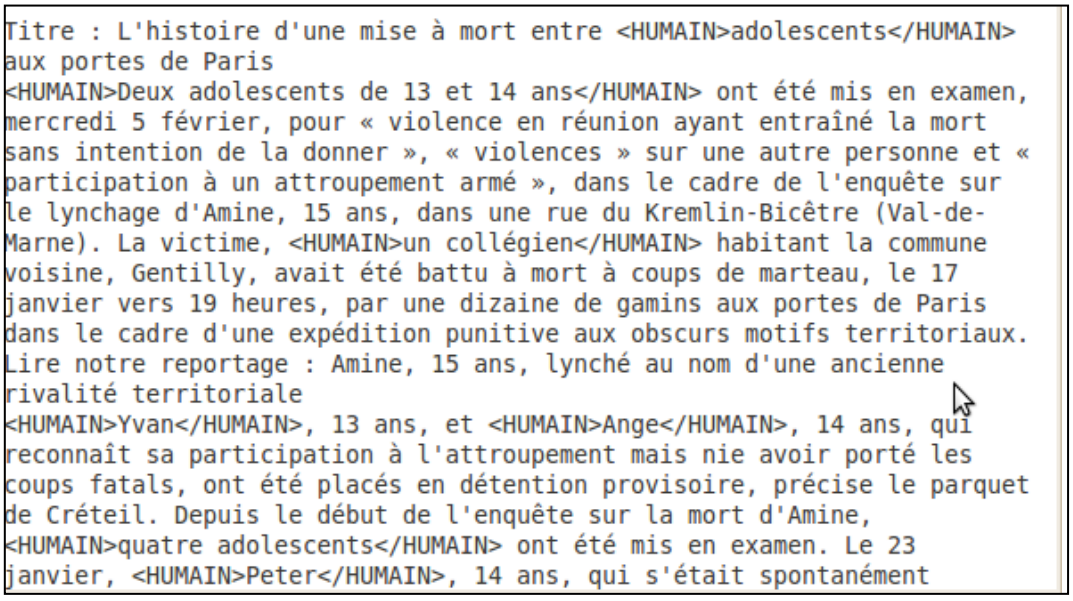


Figure 16 : La reconnaissance de la classe <humain> par les graphes d'Unitex

Figure 17 : Exemple d'un texte annoté

VÉRIFICATION ET CORRECTION DES ANNOTATIONS

La vérification et la correction éventuelle des annotations à l'étape précédente se fondent sur la distribution des GN dans le contexte phrastique, notamment avec la présence du déterminant (DET) précédant le GN. C'est une étape importante et indispensable dans le prétraitement de nos corpus.

En effet, si l'usage des dictionnaires Unitex s'avère suffisant pour une annotation des groupes nominaux hors contexte, il n'est pas fiable si le GN est placé en contexte, d'où la nécessité de vérifier ces informations. Par exemple : un GN peut être défini comme du nombre singulier dans le dictionnaire mais il est du nombre pluriel en contexte. Par exemple : tous les GN finissant par un « *s* » comme *français* par exemple. Ils peuvent être singuliers ou pluriels en contexte, mais le fonctionnement de l'Unitex oblige un choix par défaut du dictionnaire : soit le GN est considéré comme un nom au singulier, soit il est considéré comme un nom au pluriel.

1.1.3. TRANSFORMATION EN CORPUS STRUCTURÉ

Après avoir obtenu des corpus annotés avec des étiquettes de GN bien vérifiées et corrigées, nous mettons au format XML, format qui permet de structurer les informations de façon intelligente et qui sert de base de connaissance pour les étapes ultérieures. Ce choix est inspiré des travaux de Poibeau (Poibeau, 2005). Le corpus au format XML obtenu à cette étape sera utilisé en entrée dans l'étape suivante : le calcul de la saillance et la résolution des anaphores nominales.

Cette transformation est réalisée avec Python, la sortie donnera un corpus structuré au format XML, avec la tokenization en phrases et le repérage des groupes nominaux.

```
<article id="2">
    <titre>Titre : Trois « <person class='generique' g='m' n='p'>enfants</person> sauvages »
        découverts à La Courneuve </titre>
    <phrase id="1">Comment quatre <person class='generique' g='m' n='p'>enfants</person>
        déclarés à l'état civil ont-ils pu passer à travers les mailles du filet de toutes les
        institutions pendant plusieurs années ?</phrase>
    <phrase id="2"> C'est la principale question qui se pose depuis que RTL a révélé, jeudi 20
        mars, le cas d'une fratrie qui vivait recluse, dans une barre HLM de la cité des 4 000, à La
        Courneuve (Seine-Saint-Denis). </phrase>
    <phrase id="3">Il a fallu la naissance du quatrième <person class='generique' g='m'
        n='s'>enfant</person> d'un couple d'origine indienne pour que la police découvre ce qui
        se cachait depuis 2008 au 7e étage de la tour du Mail de Fontenay.</phrase>
    <phrase id="4"> Trois <person class='generique' g='m' n='p'>enfants</person>, des <person
        class='generique' g='m' n='p'>garçons</person> de 2, 5 et 6 ans, privés de soins, d'école
        et de tout suivi médical depuis leur naissance, partageaient dans le dénuement le plus
        complet une pièce d'un appartement de 85 m2. </phrase>
    <phrase id="5">Selon une source judiciaire, aucun signe de la présence d'<person
        class='generique' g='m' n='s'>enfant</person> en bas âge dans l'appartement, à
        l'exception d'un lit pliant de <person class='generique' g='m'
        n='s'>bébé</person>.</phrase>
    <phrase id="6"> Aucun jouet.</phrase>
    <phrase id="7"> La baignoire ne semble pas avoir servi depuis des mois.</phrase>
    <phrase id="8"> Pendant plus de cinq ans, aucun des occupants des 301 logements de
        l'immeuble n'avait jamais vu les deux aînés. « PRIVATION DE SOINS PAR <person
        class='famille' g='m' n='s'>ASCENDANT</person> » </phrase>
    <phrase id="9">Selon le conseil général de Seine-Saint-Denis, l'alerte a été donnée par
        l'hôpital Delafontaine à Saint-Denis où la <person class='generique' g='f'
        n='s'>mère</person> se présente le 1er janvier pour accoucher de son quatrième
        <person class='generique' g='m' n='s'>enfant</person>, une <person class='generique'
        g='f' n='s'>fille</person>.</phrase>
```

Figure 18 : XMLisation du corpus

Au format XML, le corpus (balise <document>) est structuré à différents niveaux : un <document> contient plusieurs <article>. Un <article> a un <titre>, éventuellement un <soustitre>, et des <phrase>. Les annotations des GN de la classe <personne> (ou <humain>) sont incluses dans les <titres>, <soustitre> ou <phrase>.

Les informations en provenance des dictionnaires concernant les GN (classe, genre, nombre…) sont toutes annotées en tant que valeurs d'attributs de l'élément <personne>. Pour faciliter la vérification et l'évaluation de l'annotation, le corpus annoté au XML peut également être affiché au format html.

```
<article id="3">
<p id="0">Titre : Suicide d'une <span style="color: red;" class='metier' g='m' n='s'>cadre</span> de La Poste sur s
<p id="1">Une <span style="color: red;" class='metier' g='m' n='s'>cadre</span> de la Poste s'est suicidée par pend
(Seine-Saint-Denis), où son corps a été retrouvé jeudi, a-t-on appris vendredi de source syndicale et auprès de la
qu'une <span style="color: red;" class='metier' g='f' n='s'>postière</span>, <span style="color: red;" class='metie
Coliposte [l'opérateur colis du groupe] en fonction à Noisy-le-Grand, a mis fin à ses jours sur son lieu de travail
avec une profonde émotion que La Poste a appris hier le décès brutal d'une collaboratrice, <span style="color: red;
son activité colis », a indiqué de son côté le groupe. « La Poste a immédiatement pris contact avec sa famille, éta
psychologique auprès de ses collègues et s'est mise à la disposition des autorités <span style="color: red;" class=
<p id="2"> Un CHSCT extraordinaire se tient actuellement », a ajouté l'entreprise. </p>
<p id="3">La <span style="color: red;" class='generique' g='f' n='s'>femme</span>, <span style="color: red;" class=
a été retrouvée pendue dans une partie non occupée du bâtiment.</p>
<p id="4"> Thierry Roux, <span style="color: red;" class='metier' g='m' n='s'>responsable</span> syndical FO chez C
eu lieu le drame reste inconnu, car la <span style="color: red;" class='metier' g='m' n='s'>cadre</span> « a été vu
c'est entre mercredi soir et jeudi après-midi, où on a retrouvé le corps ».</p>
<p id="5"> Selon le syndicaliste, outre le CHSCT extraordinaire qui était en cours vendredi matin, une enquête de p
</article>
```

Figure 19 : Le corpus transformé au html

Titre : Suicide d'une cadre de La Poste sur son lieu de travail

Une cadre de la Poste s'est suicidée par pendaison sur son lieu de travail à Noisy-le-Grand (Sei
direction du groupe. « Nous venons d'apprendre qu'une postière, cadre supérieure de Coliposte
indiqué FO dans un communiqué. « C'est avec une profonde émotion que La Poste a appris hie
immédiatement pris contact avec sa famille, établi un dispositif d'écoute et de soutien psycholo

Un CHSCT extraordinaire se tient actuellement », a ajouté l'entreprise.

La femme, quinquagénaire, a été retrouvée pendue dans une partie non occupée du bâtiment.

Thierry Roux, responsable syndical FO chez Coliposte a indiqué que le moment précis où a eu
soir et jeudi après-midi, où on a retrouvé le corps ».

Selon le syndicaliste, outre le CHSCT extraordinaire qui était en cours vendredi matin, une enq

Figure 20 : L'affichage du corpus par un navigateur Web

1.2. MODULE DE RESOLUTION DES ANAPHORES NOMINALES

Pour parvenir à notre objectif, nous tiendrons compte de différents paramètres syntactico-sémantiques comme l'opposition entre les noms élémentaires et les noms prédicatifs, comme l'opposition entre l'article défini et le déterminant démonstratif pour ce qui est du mode d'identification de l'antécédent d'une reprise et la notion de classe sémantique. A partir d'un état de l'art sur la résolution automatique des anaphores, nous identifions et empruntons d'autres paramètres à des modèles linguistiques différents de celui que nous envisageons d'adopter.

Notre algorithme, intégré en Python, applique les mesures de la saillance fondées sur certains paramètres syntactico-sémantiques que nous avons validés. La procédure de décision se fonde sur un seuil de poids sémantique que nous décidons arbitrairement.

1.2.1. L'ALGORITHME DU SYSTEME DE RÉSOLUTION DES ANAPHORES NOMINALES.

Nous avons fondé notre implémentation sur les heuristiques de Lappin & Leass, en y ajoutant certains critères sémantiques que nous avons pu obtenir avec nos ressources lexicales.

Notre algorithme de résolution se base sur le calcul du poids de saillance d'un GN par rapport à un autre GN, c'est-à-dire : quel est le score de ressemblance entre ces deux GN ? Il se compose de deux classes.

La première classe est réservée à :

- La création d'un objet GN avec tous ces propriétés comme ses étiquettes annotées (HC, C, SC, D1, D2, D3, genre, nombre, texte, fonction grammatical), sa position (il se trouve dans quelle phrase, quelle article, dans le titre de l'article) et son texte.
- La création d'une fonction calcule_score qui se charge de comparer 2 objets GN1 et GN2, puis elle calcule différent score comme :

  score_proximite : s'ils se trouvent dans une même phrase ou plus loin

  scores_semantiques : si leurs HC, C, domaine sont les mêmes

  scores_morphosyntaxiques : s'ils sont du même genre, du même nombre, si l'antécédent est apposition, si l'antécédent est sujet, objet ou COD.

Cette fonction renvoie le score_final = total des scores/nombre de facteurs

La deuxième classe est la classe principale. Elle contient une fonction qui permet de lire le fichier xml et stocker les GN dans une liste selon l'ordre d'apparition des GN. A cette étape, les informations comme l'identifiant de l'article, l'identifiant de la phrase, le texte et tous les attributs des balises <GN>. Par la suite, cette fonction prend le dernier élément de la liste (GN courant) et on calcule son score de saillance avec tous les éléments précédents. On assure qu'ils sont dans le même article ; On assure que l'antécédent ne se trouve pas trop loin du GN ; On utilise la fonction calcule_score pour calculer le score entre le GN courant et l'antécédent. Puis on afficher le résultat.

Pour représenter l'algorithme de notre système de résolution des anaphores nominales, nous pouvons utiliser un langage de modélisation d'objet qui est UML (Unified Modeling Language). Notre système est modélisé par des diagrammes suivants :

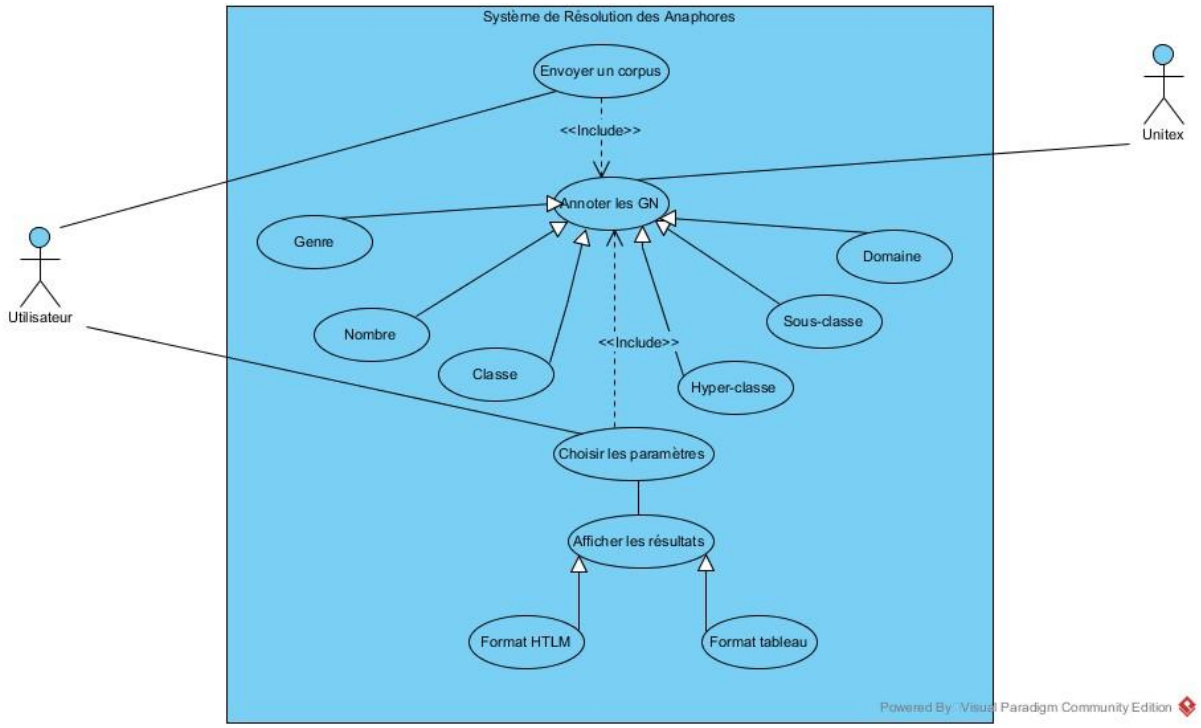


Figure 21 : Algorithme de résolution des anaphores nominales – diagramme de cas d'utilisation

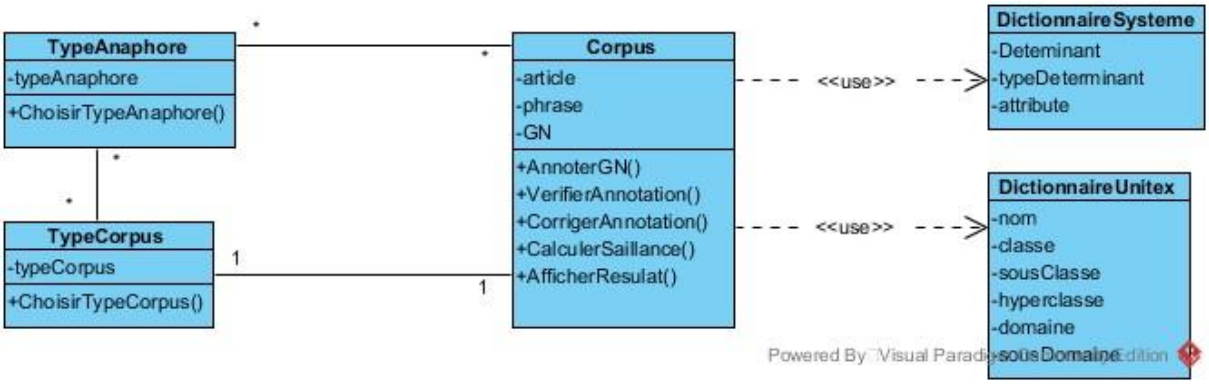


Figure 22 : Algorithme de résolution des anaphores nominales – diagramme de classes

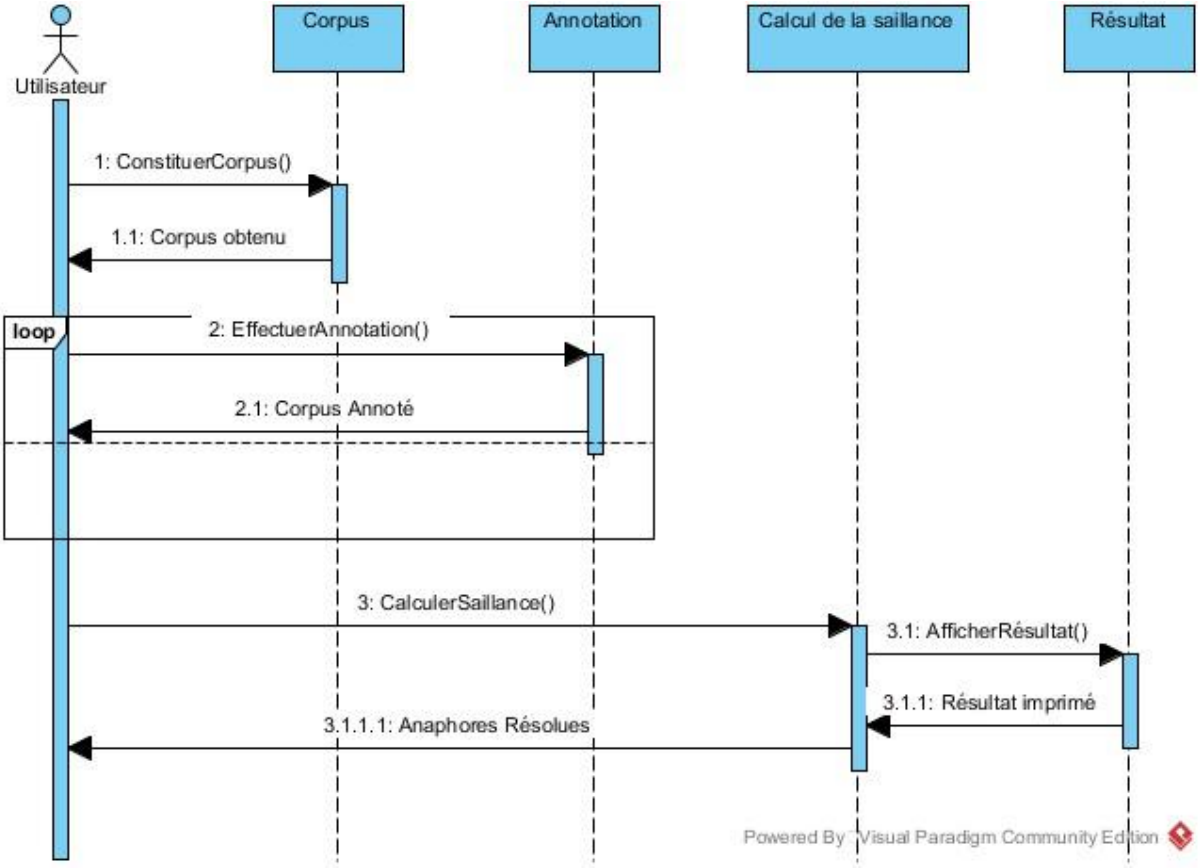


Figure 23 : Algorithme de résolution des anaphores nominales – diagramme de séquences

Les paramètres que nous pouvons choisir pour calculer le poids de saillance sont optionnels et fonction du corpus utilisé et du type d'anaphore traité. Nous pouvons choisir tous les paramètres ou juste certains paramètres en question, et les résultats varieront en fonction des paramètres choisis.

1.2.2. L'ATTRIBUTION DE LA SAILLANCE & PROCÉDURE DE DÉCISION

Pour effectuer le calcul de la saillance, nous devons nous fonder sur des paramètres lexicaux, syntaxiques et sémantiques des GN et la relation entre eux.

PARAMÈTRES LEXICAUX

Nous avons certaines règles concernant les paramètres lexicaux. En effet :

- Les GN composés, les GN du type sigles et les GN de noms propres sont traités de la même façon que les GN simples, car les entrées de ces GN particuliers sont aussi prises en compte dans nos dictionnaires Unitex.
- Le GN récursif est considéré comme un GN global. On privilégie la longueur maximale des GN pour qu'ainsi, les GN composants ne soient pas pris en compte.
- Pour les GN précédés par des déterminants spéciaux comme *le dit* + *N; le même* + *N* ; *un autre* + *N* ; *le second* + *N* ; *le premier* + *N* …, nous pouvons accorder une importance pour les candidats ayant le même genre et le même nombre que le GN.
- Le N + modifieur : le traitement des modifieurs est le cas le plus difficile dans la résolution des anaphores nominales car il demande d'autres ressources lexicales et le traitement devient plus complexe. A ce stade, nous n'avons pas pu traiter les N + modifieur.

PARAMÈTRES SYNTACTICO-SÉMANTIQUES

Concernant les paramètres syntactico-sémantiques, les règles sont :

- L'antécédent candidat ayant la même fonction syntaxique que le GN en question sera privilégié.
- Le repérage de la relation hyperonyme - hyponyme se fera selon des structures spécifiques.
- Le repérage de la relation méronymique (tout/partie) se fera selon des structures spécifiques, par exemple, le patron : *GN1 dont GN2* nous donne l'interprétation : GN1 est le tout et GN2 est une partie du tout.

- Sur certains corpus, les GN se trouvant dans le titre sont privilégiés.

Le calcul du poids de saillance se fonde sur la comparaison entre deux objets principaux : le GN en question et un GN antécédent candidat. En effet, chacun des antécédents potentiels va se voir attribuer une pondération de saillance équivalent à sa probabilité d'être remarqué et remémoré en premier. Le calcul de la saillance dépend des facteurs saillants. Le poids total de la saillance est associé au total des poids de chaque facteur.

Cette procédure emploie des valeurs telles que la proximité des GN, leurs traits lexicaux définis par l'hyperclasse, la classe, la sous-classe et les domaines. Les antécédents candidats doivent avoir un ou plusieurs attributs ressemblant à l'attribut de la reprise (l'hyperclasse, la classe, les domaines etc.), et certains GN peuvent devenir candidats potentiels grâce à ces attributs.

Par exemple, si le genre du GN antécédent et le genre du GN en question sont les mêmes, nous ajoutons 1 dans le poids *genre* (poids local)… De la même façon, nous obtenons les différents poids pour le *nombre*, la *classe*, la *sous-classe,* l'*hyperclasse*, le *domaine*… et un autre poids pour la *proximité* (si le GN antécédent candidat est relativement « proche » au GN en question, la saillance des antécédents les plus éloignés diminue). Si les antécédents candidats plus proches sont plus saillants que les GN plus loin, les antécédents qui ne se trouvent pas dans la bonne fenêtre de sélection seront, logiquement, rejetés. Cette décision est inspirée des travaux de Kleiber (Kleiber, 1999b), selon lesquels Kleiber a montré que l'accessibilité des anaphores associatives s'appuie sur une relation économe entre l'expression anaphorique et l'antécédent. Par défaut, l'accès se fait sur *le dernier élément pertinent disponible en mémoire* à partir de l'état actuel de la représentation linéaire du discours par le récepteur.

Le poids de saillance total correspond au poids moyen de tous les poids locaux. En Python, il est intégré avec cette formule :

Score_total = float(total de scores élémentaires)/ nombre de scores élémentaires

Parmi les candidats restants, on élimine ceux dont le poids final (global + local) est trop bas (au dessous de 0.5 par exemple), ce qui permet de dire que l'antécédent est inconnu.

Si plus d'un candidat est retenu, on va choisir le candidat ayant le poids de saillance le plus important. S'ils ont le même poids, on va choisir le plus proche de l'anaphore.

Pour la sortie, le programme affiche l'ensemble des antécédents candidats avec leurs scores pour chacun des GN du document.

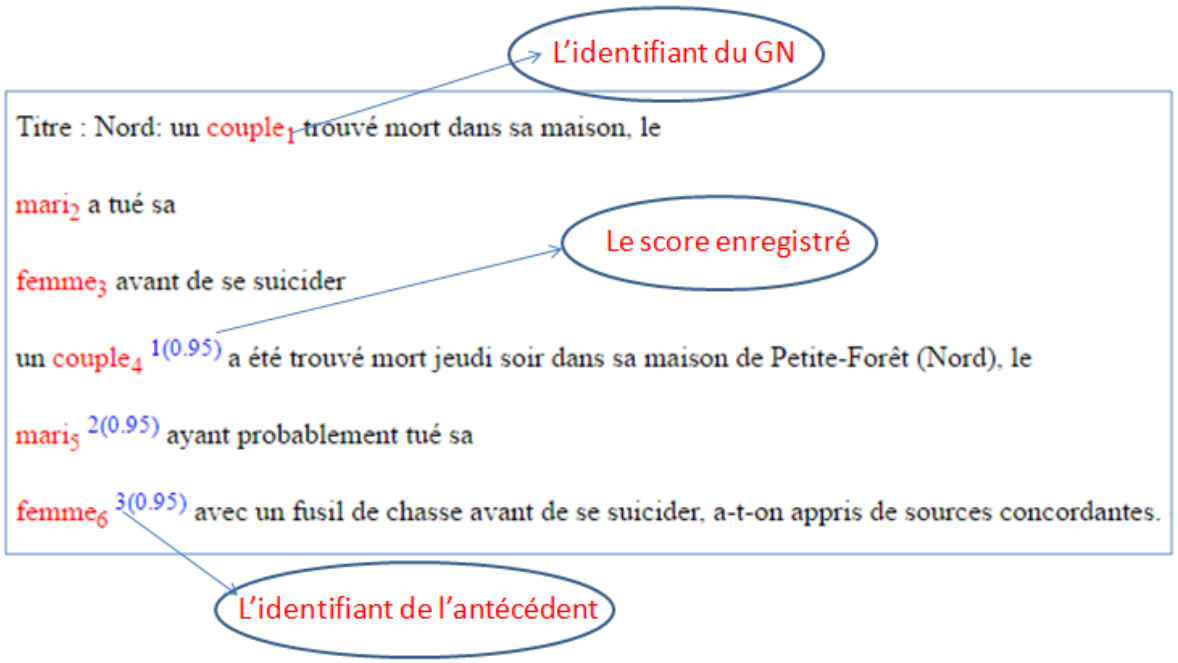


Figure 24 : Exemple d'une sortie

1.3. MODULE D'ÉVALUATION

Après avoir obtenu le résultat d'un traitement automatique, nous classons les séquences étiquetées selon différents cas. Le premier cas concerne les termes (ou les

appariements pour la résolution des anaphores) reconnus comme on le souhaite, alors cet étiquetage est correct. Le deuxième cas concerne les termes trouvés par le système mais qu'on ne devrait pas reconnaître (étiquetage incorrect). Il s'agit des « bruits » ou ensemble d'étiquettes non pertinentes trouvées lors de l'étiquetage d'un document. Le troisième cas représente les termes qui devraient être reconnus mais ne le sont pas, autrement dit, ils sont manquants dans la sortie (étiquetage manquant). Il s'agit du taux de silence, ou ensemble d'étiquettes pertinentes non spécifiées lors de l'étiquetage.

L'analyse de la performance d'un système de TALN se fonde généralement sur les mesures de rappel et de précision. Cette méthode d'évaluation est la plus utilisée dans ce domaine. L'évaluation d'un système doit se fonder sur deux critères : la fiabilité (la sortie contient-elle moins d'erreur ?) et le taux de couverture (la sortie couvre-t-elle la totalité des cas que l'on doit traiter ?).

La précision est la fiabilité du système, qui mesure la capacité du système à refuser les solutions non-pertinentes tandis que le taux de rappel est la capacité du système à donner toutes les solutions pertinentes.

$$\text{Précision} = \frac{\text{Nombre d'étiquettes correctes}}{\text{Nombre d'étiquettes correctes + Nombre d'étiquettes incorrect}}$$

Autrement dit, le taux de précision est la mesure de l'efficacité d'un système d'étiquetage établie à partir du ratio entre le nombre d'informations pertinentes trouvées lors de l'étiquetage d'un document et le nombre total d'étiquettes fournies par le système. C'est un indicateur de mesure de bruit car plus le taux de précision est faible, plus il y a de bruit.

$$\text{Rappel} = \frac{\text{Nombre d'étiquettes correctes}}{\text{Nombre d'étiquettes correctes + Nombre d'étiquettes manquantes}}$$

Le rappel est le taux de couverture que le système influence sur un document. Autrement dit, le taux de rappel correspond à l'ensemble d'étiquettes pertinentes non spécifiées lors de l'étiquetage établi à partir du ratio entre le nombre d'étiquettes pertinentes lors de l'étiquetage d'un document et le nombre total d'étiquettes pertinentes du document. Ainsi, plus le taux de rappel est faible, plus il y a de silence :

$$\text{Rappel} = \frac{\text{Nombre d'étiquettes pertinentes spécifiés}}{\text{Nombre d'étiquettes pertinentes existantes}}$$

La F-mesure est la moyenne harmonique du rappel et de la précision.

$$\text{F-score} = \frac{2 * \text{Précision} * \text{Rappel}}{\text{Précision+Rappel}}$$

Notre module d'évaluation se fonde principalement sur le calcul des taux de précision, de rappel et le F-score (ou F1).

Pour évaluer notre travail, nous avons choisi de segmenter chaque corpus en 3 sous-corpus :

- Un corpus d'entraînement, ou corpus de travail
- Un corpus de test
- Un corpus d'évaluation

Le corpus de travail et le corpus de test occupent chacun 40% et 50% respectivement du corpus global, et le corpus d'évaluation de 10% du corpus.

Pour pouvoir faire l'évaluation, les annotations manuelles des expressions anaphoriques et leurs référents sont indispensables. Pourtant, l'annotation des anaphores nominales reste une tâche très compliquée. Dans la résolution des anaphores pronominales ou des chaînes de références, l'usage des outils d'annotation existants est réalisable car le nombre des référents sont souvent limités.

Pourtant, le nombre élevé des expressions anaphoriques nominales dans nos traitements empêche l'usage des outils d'annotation habituels, car un tel travail demande beaucoup de temps.

Pour notre travail de thèse, nous avons divisé le corpus d'évaluation en échantillons. Le résultat de l'évaluation se base sur le résultat de nos échantillons et la taille des échantillons varie selon le corpus ou la méthode utilisée.

Nous pouvons également avoir une évaluation sur l'extraction des GN des différentes classes, comme <Personne>, <Humain> ou <Artefact>, en utilisant le format .doc (word) avec une couleur choisie (rouge). La sortie de notre programme sera un corpus au format html. Lorsque la sortie est interprétée par un navigateur web, nous pouvons observer une ressemblance avec le format .doc.

Par exemple, pour la classe <Personne>, nous avons cette annotation manuelle, au format .doc :

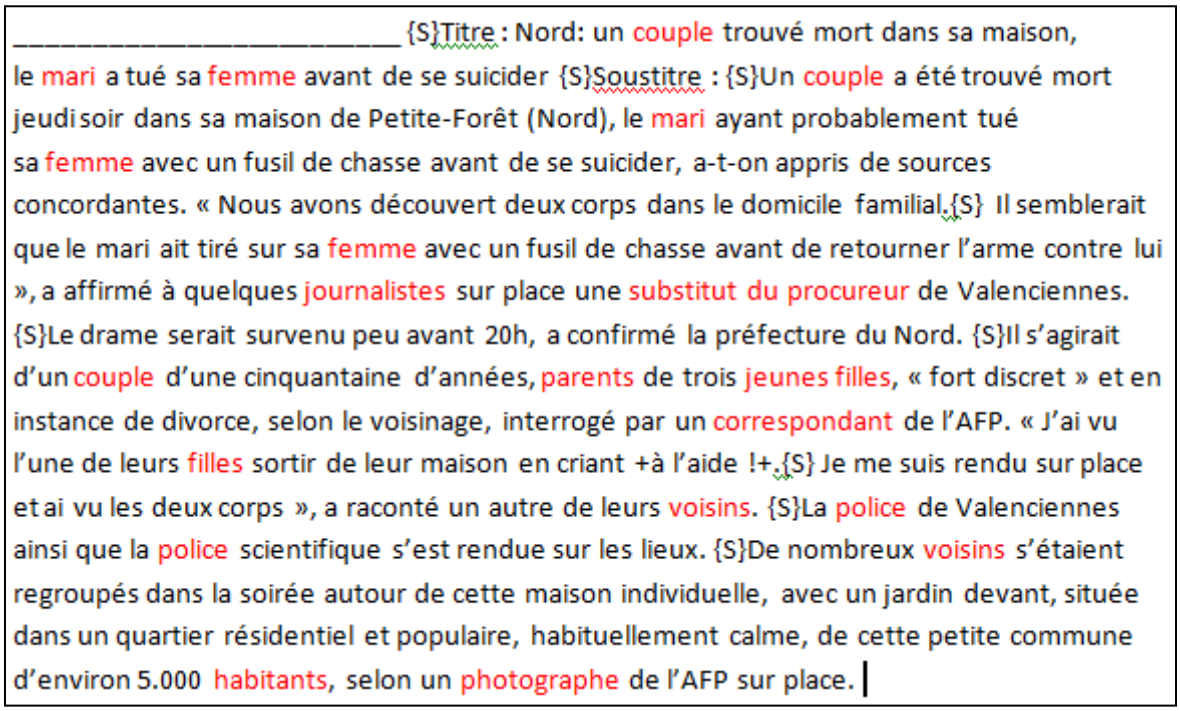

_____ {S}Titre : Nord: un couple trouvé mort dans sa maison, le mari a tué sa femme avant de se suicider {S}Soustitre : {S}Un couple a été trouvé mort jeudi soir dans sa maison de Petite-Forêt (Nord), le mari ayant probablement tué sa femme avec un fusil de chasse avant de se suicider, a-t-on appris de sources concordantes. « Nous avons découvert deux corps dans le domicile familial.{S} Il semblerait que le mari ait tiré sur sa femme avec un fusil de chasse avant de retourner l'arme contre lui », a affirmé à quelques journalistes sur place une substitut du procureur de Valenciennes. {S}Le drame serait survenu peu avant 20h, a confirmé la préfecture du Nord. {S}Il s'agirait d'un couple d'une cinquantaine d'années, parents de trois jeunes filles, « fort discret » et en instance de divorce, selon le voisinage, interrogé par un correspondant de l'AFP. « J'ai vu l'une de leurs filles sortir de leur maison en criant +à l'aide !+.{S} Je me suis rendu sur place et ai vu les deux corps », a raconté un autre de leurs voisins. {S}La police de Valenciennes ainsi que la police scientifique s'est rendue sur les lieux. {S}De nombreux voisins s'étaient regroupés dans la soirée autour de cette maison individuelle, avec un jardin devant, située dans un quartier résidentiel et populaire, habituellement calme, de cette petite commune d'environ 5.000 habitants, selon un photographe de l'AFP sur place.

Figure 25 : Annotations manuelles des syntagmes nominaux

La première étape de notre module d'évaluation consiste à évaluer le bruit et le silence à partir des corpus d'investigation, qui nous permettra d'identifier les phénomènes linguistiques qui seront pris en compte dans la résolution des anaphores nominales. Compte tenu de ces phénomènes et de leur nombre, les corpus d'investigation peuvent nous permettre de savoir si les ressources ou les méthodes choisies sont appropriées ou non à notre travail.

2. LES RESSOURCES LEXICALES UTILISEES

Notre sujet de thèse, qui concerne les anaphores nominales, demande une quantité importante et diversifiée de connaissances lexicales et encyclopédiques. L'état de l'art montre que des systèmes de résolution des anaphores dépendent fortement des ressources « artisanales » dont la première base de ressource sémantique à vocation universelle est WordNet. Les unités lexicales et les relations dans WordNet se fondent sur ses douze grandes classes nominales et les synonymies correspondantes (synsets), et chaque entrée lexicale est associée aux informations morphologiques et syntaxiques.

Pourtant, utiliser WordNet présente plusieurs inconvénients majeurs. Tout d'abord, les sens et les relations lexicales de nombreuses expressions sont absents de la base de données. De plus, l'information souhaitée peut ne pas être facile à récupérer, dû à sa hiérarchie structurelle. Par exemple, dans WordNet, «plancher» est encodé dans le cadre de la «construction», mais pas dans le cadre de «appartement», bien que «appartement» soit lui-même un méronyme de «immeuble», qui à son tour est un hyponyme de «bâtiment» (Vieira and Poesio, 2000). C'est une ressource peuplée manuellement et par conséquent majoritairement correcte mais incomplète, et on y trouve très peu d'informations sur les contextes dans lesquels les mots apparaissent. Elle ne sera significativement exploitée en résolution des anaphores qu'à partir de Vieira & Poesio.

Les ressources lexicales existantes peuvent être utilisées via des corpus annotés dont le FrenchTreebank. Le jeu d'étiquettes pour le FTB comporte 14 catégories lexicales qui sont les suivantes : A (adjectif), Adv (adverbe), D (déterminant), CC (conj de coordination), CL (pronom personnel clitique), CS (conj de subordination), NC (nom commun), NP (nom propre), P (préposition), PRO (pronom non clitique), V (verbe), I (interjection), ET (mot étranger dont on ne peut deviner la catégorie en contexte), PONCT (ponctuation).

Il existe également des informations sur les fonctions syntaxiques de chaque groupe :

SUJ : fonction sujet

OBJ : fonction objet

MOD : fonction modale

Bien que l'utilisation du FTB soit assez coûteuse en temps de traitement car les informations sont très riches, ce corpus n'est pas vraiment efficace pour le traitement des anaphores nominales, car les relations anaphoriques, pour un corpus tellement volumineux, sont assez rares.

C'est pour ces raisons que nous cherchons à élaborer nos propres ressources lexicales et nous limitons le domaine de traitement en utilisant les corpus spécifiques. Nos ressources lexicales se fondent sur la théorie des trois fonctions primaires, développée au laboratoire LDI (Buvet, 2009).

2.1. DESCRIPTION DES DICTIONNAIRES ELECTRONIQUES

Les ressources lexicales que nous utilisons dans notre recherche sont élaborées principalement en se fondant sur la théorie des classes d'objets étudiée par G. Gross ; Le Pesant ; Michel Mathieu-Colas (Mathieu-Colas, 1998). En effet, nous avons

choisi de créer nous-même les dictionnaires électroniques en fonction du thème de notre corpus et du type d'anaphore traité.

Les dictionnaires électroniques que nous utilisons dans le cadre du TALN sont de deux sortes :

- Des dictionnaires syntactico-sémantiques des prédicats qui permettent la catégorisation grammaticale de certains mots d'un texte. L'élaboration d'une classe de prédicats s'effectue selon le domaine d'arguments, afin de traiter les problèmes d'ambiguïtés liés aux prédicats. La constitution d'une classe de prédicats s'effectue aussi sur la base de la synonymie, elle est définie à partir d'un prédicat exemplaire. Il s'agit du prédicat le plus représentatif de la classe, les autres prédicats étant considérés comme périphériques (Buvet and Grezka, 2007)
- Des dictionnaires des arguments, qui sont en réalité des noms élémentaires de différentes catégories. Les classes sémantiques des arguments sont constituées de noms élémentaires, c'est-à-dire les noms qui fonctionnent uniquement comme des arguments. Par exemple, les noms d'artefact sont recensés et décrits dans un dictionnaire, et qui se trouvent dans une nomenclature qui comporte des noms simples (*alambic*) ou des noms construits du type suffixé (*boîtier*) ou du type composé (*boîte de vitesse*). Ils constituent la macrostructure du dictionnaire. Les variations de forme d'un nom donné constituent autant d'entrées du dictionnaire ; *bracelet cuir* et *bracelet en cuir* sont deux entrées différentes. La microstructure est constituée de la vedette et de cinq descripteurs (Buvet, 2015) :

    1) hyperclasse ;

    2) classe;

    3) domaine1;

    4) domaine2 ;

5) domaine 3.

Le concept du dictionnaire électronique, proposé par P.A Buvet (Buvet, 2015), décrit 18 hyperclasses, ou 18 catégories d'artefact suivant :

ALIMENT	*yaourt*
APPAREIL	*bouveteuse*
CONTENANT	*bassine*
COSMETIQUE	*lait de maquillage*
DISPOSITIF	
DOCUMENT	
GENERIQUE	*Produit*
INSTRUMENT	*guitare*
MACHINE	*tablette tactile*
MOYEN_TRANSPORT	*bicyclette*
ORGANE	*bouton*
OUVERTURE	
PARTIE_GENERIQUE	*sommet*
PARTIE_VETEMENT	*manche*
REVETEMENT	
SUPPORT	
USTENSILE	
VETEMENT	

Tableau 5 : Les hyperclasses proposées par P-A. Buvet

Le descripteur classe rassemble des catégories dénommées selon le principe suivant :

NOM D'HYPERCLASSE_NOM DE FONCTION D'ARTEFACT

Pour le français, plus de 20 000 noms d'artefact ont été recensés et près de 140 noms de fonction (classes) ont été associés à des noms d'hyperclasse. Par exemple, l'hyperclasse OUVERTURE subsume les classes :

OUVERTURE_CONNEXION

OUVERTURE_DIFFUSION

OUVERTURE_ EVACUATION

OUVERTURE_FIXATION

OUVERTURE_GENERIQUE

OUVERTURE_PASSAGE

OUVERTURE_TRANSMISSION

Le choix d'étudier des anaphores associatives de type partie-tout impliquant des noms d'artefact est une conséquence des propriétés remarquables de ces substantifs.

Les trois derniers descripteurs (domaine 1, 2 et 3) sont de nature pragmatique. Ils portent tous sur le domaine d'emploi des noms d'artefact. Seul le premier descripteur est obligatoirement spécifié, les deux autres le sont si le nom d'artefact relève de plus d'un domaine.

2.2. ANALYSE LEXICALE ET MORPHO-SYNTAXIQUE

Les connaissances de nature syntaxique, sémantique et pragmatique sont indispensables pour une bonne analyse du phénomène des anaphores nominales.

Le sujet de notre recherche concerne notamment les anaphores nominales, comme :

Les anaphores fidèles :

*Il a acheté **une limousine. La limousine** est très belle*

Les anaphores infidèles :

*Il a acheté **une limousine. Cette voiture** est très belle*

Les anaphores associatives du type méronymique :

*J'ai commandé **une machine à café. Le filtre** est rénové.*

Nous souhaitons tout d'abord procéder à l'analyse linguistique au niveau des mots. En effet, les groupes nominaux seront au cœur de notre travail. Au niveau du mot, l'analyse linguistique est primordiale car le mot est une unité linguistique dotée d'un

signifiant, qui désigne l'image acoustique du mot, et d'un signifiant qui désigne le concept, la représentation mentale d'une chose (Saussure, 1967). Les mots doivent être étudiés en contexte phrastique puisque la phrase est constituée de plusieurs unités linguistiques qui entretiennent des relations syntaxiques et sémantiques entre autres. Comme la plupart des mots sont polysémiques, le sens des mots doit être défini en contexte. Or, il y a peu de moyens automatiques pour préciser le contexte, nous choisissons de limiter les corpus par thème.

Une analyse morpho-syntaxique des groupes nominaux est constituée des étapes suivantes :

- Découpage du corpus en phrases
- Découpage des phrases en unités lexicales en formes complexes et formes simples. Les formes complexes, qui sont aussi appelées les locutions (par exemple : directeur de commerce, belle-mère…) requièrent une analyse morphologique compositionnelle. Le sujet des anaphores nominales nous restreint à l'analyse des locutions nominales.
- Reconnaissance des unités lexicales, notamment des unités complexes comme les groupes nominaux composés ; les groupes nominaux simples ; les déterminants ; et aussi les mots inconnus.
- Attribution d'information morpho-syntaxique aux mots. A ce niveau, notre étude privilège la reconnaissance des déterminants de type défini pour les anaphores associatives, et les déterminants possessifs pour les anaphores possessives. En fonction du type d'anaphore étudié, du corpus étudié, nous privilégions certaines informations liées aux groupes nominaux comme le genre, le nombre, la catégorie sémantique…

Une analyse des fonctions syntaxiques de la phrase est constituée des étapes suivantes :

- Identification des groupes de mots
- Identification de la structure syntaxique

- Identification des syntagmes et de leur tête
- Identification des relations entre syntagmes

Par exemple, le groupe sujet peut se trouver sous quatre formes :

Le golfeur joue sur le Drill.

Sujet = Un groupe nominal

Claude joue au golf.

Sujet = Un nom propre

Il joue au golf.

Sujet = Un pronom

Jouer au golf, est dynamisant.

Sujet = Un verbe

2.3. ANALYSE DE LA STRUCTURE PRÉDICAT-ARGUMENT

Du point de vue de la théorie des trois fonctions primaires, la phrase est toujours constituée d'un prédicat et souvent d'arguments et d'actualisateurs. Le besoin de traitement automatique a fait en sorte qu'on définisse la phrase en tant que prédicat suivi éventuellement des arguments et des actualisateurs.

En français, la description de l'ordre des constituants dans la phrase est fondée sur la hiérarchisation des fonctions grammaticales, plus précisément sur l'opposition entre fonctions essentielles et fonctions non essentielles (ou accessoires). Les fonctions essentielles recouvrent le sujet, l'attribut et les objets (direct, indirect, prépositionnel) et les fonctions accessoires englobent les différents types de compléments circonstanciels (ou circonstants) (Chébouti, 2014).

Par exemple, l'analyse des syntagmes et de leur tête dans la phrase :

Pierre demande une aide.

[Pierre] [demande] [une aide]

SN V SN

La relation entre les syntagmes de cette phrase est :

Le SN [Pierre] est le sujet du V [demande]

Le SN [une aide] est l'objet direct du V [demande]

La phrase française adopte typiquement l'ordre sujet-verbe-compléments (SVO) Cette séquence servant de structure de base, les compléments circonstanciels se présentent à l'initiale ou en finale de phrase. Le verbe forme avec le sujet le noeud sémantique de la phrase verbale, et c'est le verbe seul qui en constitue le noyau syntaxique. Il s'agit de la relation prédicats-arguments.

Une analyse de la structure prédicat-argument de la phrase permet une analyse de sens des unités lexicales.

Par exemple :

Paul a distribué du pain aux oiseaux.

Le sens des GN *pain* et *oiseaux* dans la phrase revient à l'analyse des structures prédicatives suivantes :

[arg1] distribuer [arg2] à [arg3]

[arg1] distribuer [arg2]

Puisque les arguments représentent différents degrés de généralité dans la description, une analyse des relations prédicats-arguments consiste en la détection des arguments de la phrase.

La détection des arguments de la phrase nécessite l'identification des types sémantiques et des rôles sémantiques.

Les types sémantiques sont les classes sémantiques de l'ensemble des expressions susceptibles d'instancier les arguments d'un prédicat. Le type sémantique d'une expression indique la nature de sa dénotation (un individu, un prédicat, un objet physique…).

Par exemple :

Jean coupa le pain avec un couteau

Les types sémantiques de la structure prédicative *X ouvrir Y avec Z* sont représentés ainsi : [*Individu*] couper [*ObjetPhysique*] avec [*ObjetPhysique*]

Les rôles sémantiques (ou rôle thématique) sont les relations sémantiques qu'un prédicat entretient avec ses arguments (Saint-Dizier, 2006)

Les rôles sémantiques de la structure prédicative *X ouvrir Y avec Z* sont représentés ainsi : *[Agent]* couper *[Patient]* avec *[Moyen]*

[Agent] est celui qui fait l'action volontairement

[Patient] est celui qui est affecté par l'action et subit un changement

De la même façon, nous avons d'autres rôles sémantiques comme :

[Expérienceur] est celui qui perçoit l'action mais ne la contrôle pas

[Bénéficiaire] est celui pour qui l'action est faite

[Thème] qui caractérise une entité déplacée, conséquence de l'action dénotée par le prédicat …

3. LA PLATEFORME UNITEX

Pour procéder à l'étiquetage des textes, aussi bien l'étiquetage morphosyntaxique que l'étiquetage syntactico-sémantique, nous utilisons le logiciel UNITEX développé à l'Université de Marne-la-Vallée à partir des travaux de Maurice Gross (Gross, 1997, 1989) en informatique et en linguistique. Ce logiciel exploite deux sortes de ressources linguistiques, à savoir des dictionnaires électroniques et des grammaires. Il permet aux utilisateurs non seulement d'implémenter les expressions régulières dans la recherche des informations, mais aussi de construire, de vérifier et d'appliquer les dictionnaires électroniques personnalisés.

3.1. LES DICTIONNAIRES D'UNITEX

On distingue deux sortes de dictionnaires électroniques : Les dictionnaires internes d'Unitex et les dictionnaires externes que nous pouvons créer nous-même.

Les dictionnaires internes ont deux types : le dictionnaire de formes fléchies et non fléchies. Le type que l'on utilise le plus couramment est le dictionnaire de formes fléchies, appelé DELAF (DELA de formes Fléchies) ou encore DELACF (DELA de formes Composées Fléchies) lorsqu'il s'agit d'un dictionnaire de mots composés. Le second type est le dictionnaire de formes non fléchies appelé DELAS (DELA de formes Simples) ou DELAC (DELA de formes Composées). Le format des DELAS est très similaire à celui des DELAF. La différence est qu'on ne mentionne qu'une forme canonique suivie de codes grammaticaux ou sémantiques. La forme canonique est séparée des différents codes par une virgule (Paumier, 2015).

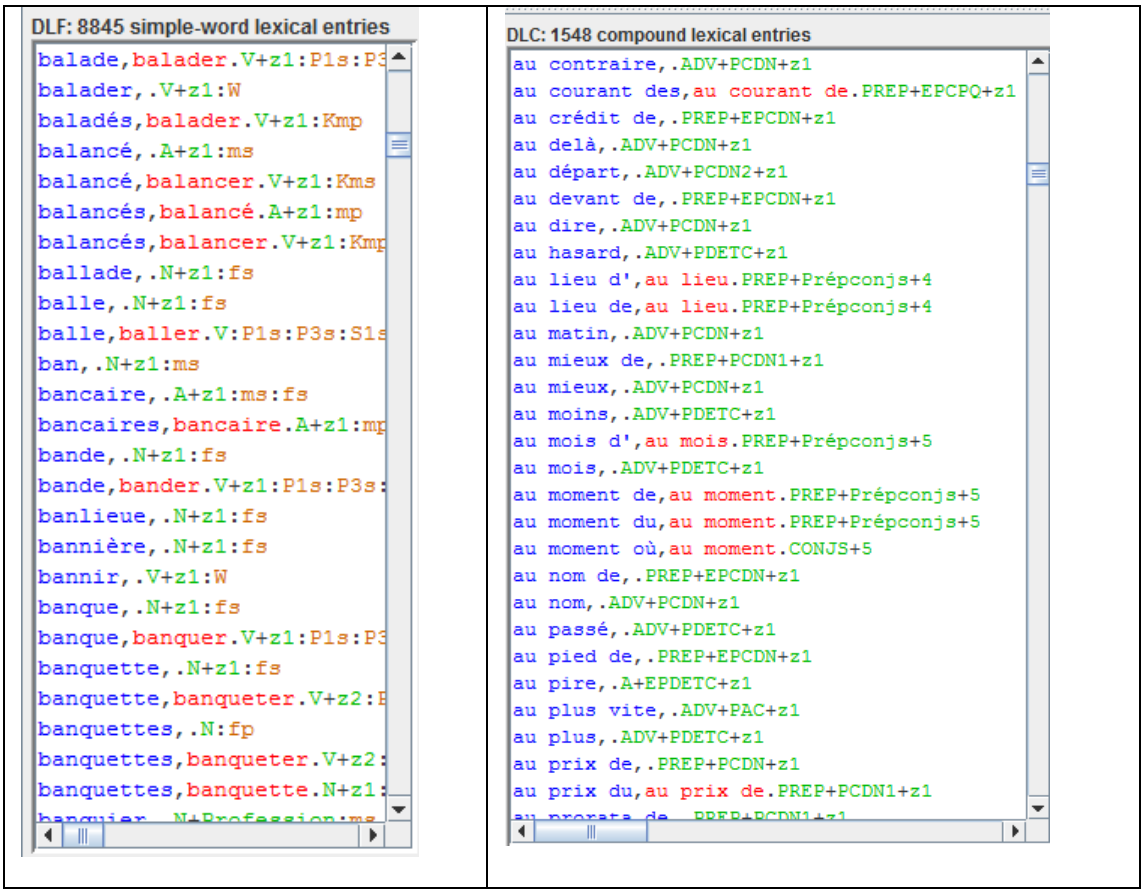


Figure 26 : Les formats DELAF

Nous pouvons aussi créer nos propres dictionnaires (dictionnaires externes) en suivant certaines normes d'écriture :

Forme fléchie, Forme canonique. CodeGrammaticale+CodeSémantique :GenreNombre

La forme fléchie est obligatoire. La forme canonique correspond au lemme de l'entrée. Pour les noms et les adjectifs, il s'agit en général de la forme au masculin singulier ; pour les verbes, la forme canonique est l'infinitif. Cette information est facultative. Dans le cas où on veut omettre la forme canonique, cela signifie alors que la forme canonique est identique à la forme fléchie. La forme canonique est séparée de la forme fléchie par une virgule, et le code devient :

Forme fléchie,. CodeGrammaticale+CodeSémantique :GenreNombre

Toute entrée doit comporter au moins un code grammatical ou sémantique, séparé de la forme canonique par un point. S'il y a plusieurs codes, ceux-ci doivent être séparés par le caractère +.

Certains mots composés comme grand-mère peuvent s'écrire avec des espaces ou avec des tirets. Pour éviter de devoir dédoubler toutes les entrées, il est possible d'utiliser le caractère =. Ainsi, l'entrée suivante : *grand=mères,grand=mère.N:fp* est remplacée par les deux lignes suivantes :

grand mères,grand mère.N:fp

grand-mères,grand-mère.N:fp

3.2. LES GRAPHES D'UNITEX

Unitex est un logiciel d'analyse morphologique, qui a pour but de définir des automates à états finis (AEF), genres de grammaires très proches des grammaires formelles : on part de l'axiome, on a un ensemble de règles, un vocabulaire non terminal (ensemble des symboles S, NP, VP, N, V, etc., vocabulaire du métalangage), et un vocabulaire terminal (lexique, vocabulaire de la langue naturelle). Un AEF comprend plusieurs états successifs dont le premier est l'état initial, les symboles (appartenant au vocabulaire terminal) permettant de passer d'un état à un autre jusqu'à l'état final.

Un automate à états finis est défini par 4 éléments :

- Un vocabulaire : ensemble des symboles trouvés sur les arcs reliant les différents états entre eux
- L'ensemble des états
- L'état initial
- L'ensemble des états finaux

La fonction de transition correspond à l'ensemble des passages, des arcs reliant les états les uns aux autres.

Pour un meilleur affichage des graphes, nous avons utilisé la version 3.0 d'Unitex car cette version comporte des améliorations appréciables au niveau graphique, la création de graphes est ainsi facilitée.

Les graphes syntaxiques de l'Unitex sont les grammaires locales, régulières utilisées par l'analyseur syntaxique. Ils représentent la classe des grammaires les plus simples dans la hiérarchie de Chomsky.

Le logiciel donne également la possibilité de créer des transducteurs qui convertissent les entrées en représentant des grammaires locales sous la forme d'un automate à états finis. L'automate à états finis est un graphe dans lesquelles les sommets sont appelés états et les flèches représentent les transitions. Un transducteur fini est un automate à états finis avec sorties. Il opère sur les entrées textuelles.

Ses transducteurs font appel aux dictionnaires, aux contextes et utilisent toute la palette des ressources internes.

Exemple d'un graphe :

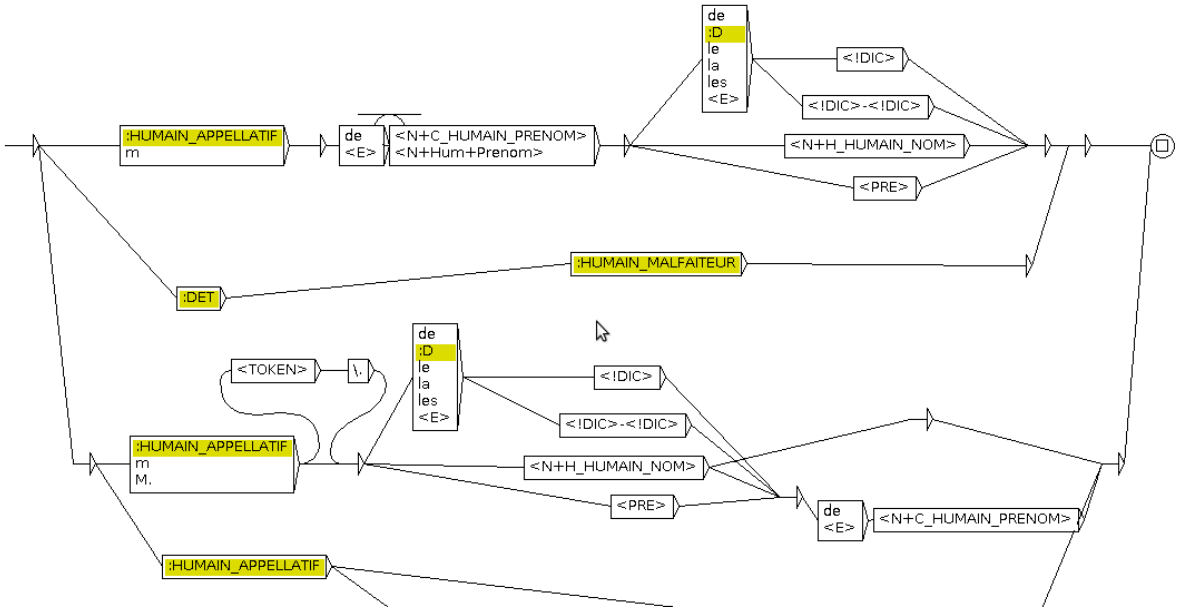


Figure 27 : Le graphe d'extraction de la classe <Humain>

Les graphes sont susceptibles également de générer une représentation en sortie. Les sorties peuvent être stockées dans les variables qui sont utilisées dans les graphes.

Par exemple, avec le graphe suivant, nous pouvons obtenir un texte avec les étiquettes de classe <humain> annotées :

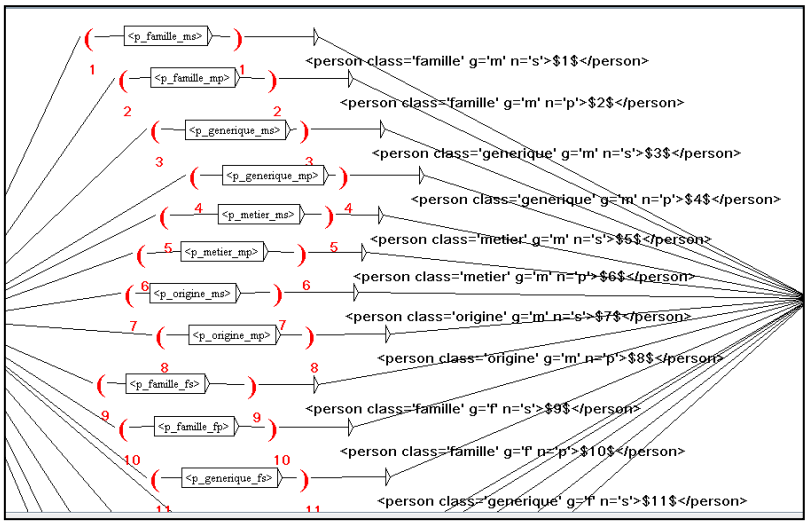
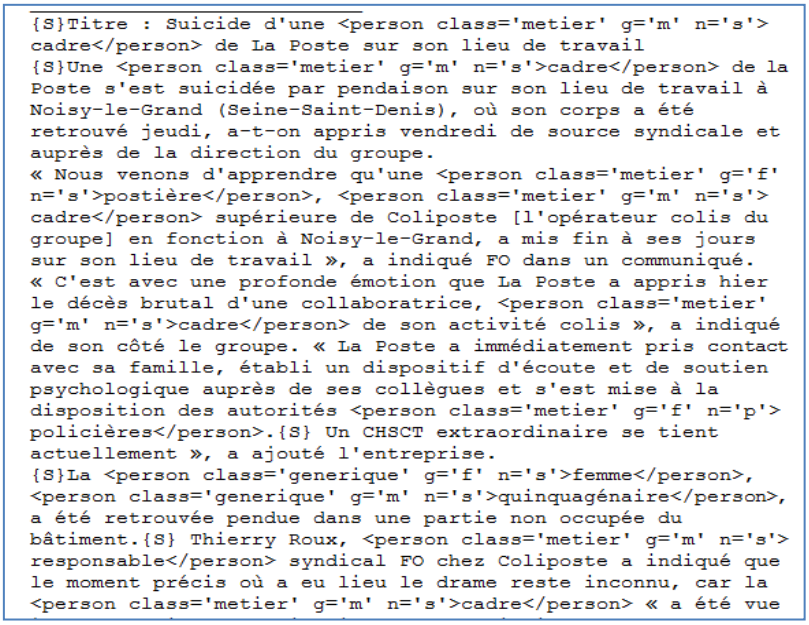


Figure 28 : Attribution des étiquettes de la classe <Humain>



```
{S}Titre : Suicide d'une <person class='metier' g='m' n='s'>
cadre</person> de La Poste sur son lieu de travail
{S}Une <person class='metier' g='m' n='s'>cadre</person> de la
Poste s'est suicidée par pendaison sur son lieu de travail à
Noisy-le-Grand (Seine-Saint-Denis), où son corps a été
retrouvé jeudi, a-t-on appris vendredi de source syndicale et
auprès de la direction du groupe.
« Nous venons d'apprendre qu'une <person class='metier' g='f'
n='s'>postière</person>, <person class='metier' g='m' n='s'>
cadre</person> supérieure de Coliposte [l'opérateur colis du
groupe] en fonction à Noisy-le-Grand, a mis fin à ses jours
sur son lieu de travail », a indiqué FO dans un communiqué.
« C'est avec une profonde émotion que La Poste a appris hier
le décès brutal d'une collaboratrice, <person class='metier'
g='m' n='s'>cadre</person> de son activité colis », a indiqué
de son côté le groupe. « La Poste a immédiatement pris contact
avec sa famille, établi un dispositif d'écoute et de soutien
psychologique auprès de ses collègues et s'est mise à la
disposition des autorités <person class='metier' g='f' n='p'>
policières</person>.{S} Un CHSCT extraordinaire se tient
actuellement », a ajouté l'entreprise.
{S}La <person class='generique' g='f' n='s'>femme</person>,
<person class='generique' g='m' n='s'>quinquagénaire</person>,
a été retrouvée pendue dans une partie non occupée du
bâtiment.{S} Thierry Roux, <person class='metier' g='m' n='s'>
responsable</person> syndical FO chez Coliposte a indiqué que
le moment précis où a eu lieu le drame reste inconnu, car la
<person class='metier' g='m' n='s'>cadre</person> « a été vue
```

Figure 29 : Le corpus analysé morpho-syntaxique

3.3. L'ÉTIQUETAGE DES GROUPES NOMINAUX

L'étiquetage des groupes nominaux est une condition nécessaire mais non suffisante au repérage des anaphores associatives. Pour cet étiquetage, nous exploitons un dictionnaire morphosyntaxique et des grammaires locales. L'étiquetage fait appel à trois opérations:

- La première opération est la tokenisation. Unitex utilise le programme *Tokenize* pour découper le texte en phrases {S} (à l'aide des délimiteurs comme les ponctuations, les retours à la ligne, la première lettre du premier mot de la phrase en majuscule …), ou en mot (un token est une suite de lettres encadrées de séparateurs, qui peuvent être des espaces ou des signes de ponctuation). Le module *Tokenise* permet aussi d'afficher la liste des unités lexicales dans l'ordre où elles ont été trouvées dans le texte. Après la tokenisation, les mots considérés comme inconnus sont aussi stockés dans un fichier à part. En raison de certains critères stricts, la segmentation ne peut pas être produite dans certains textes provenant des commentaires/avis des

internautes, si ces derniers ne sont pas bien écrits (par exemple : manque de ponctuation, de majuscule …)

```
_____________________________
{S}Titre : quelle garantie pour le smartphone de mon fils
{S}Bonjour à tous,je viens poster ici pour une question concernant la garan
un Iphone reconditionné sur le site Cdiscount.{S}Il s'est rapidement apperç
{S} Nous avons renvoyé l `appareil et ils nous ont fait un retour 1mois 1/2
, cet appareil a de nouveau un problème avec son écran qui n'affiche plus r
avait égaré.{S} Il m'a renvoyé le document par mail , avec précisé dessus "
afin d'un retour pour réparation ou changement.{S} La seule réponse que j'a
!!! ) était dépassée donc pas de prise en garantie.{S}Je précise que l'acha
garantie européenne de 2 ans ?{S} S'applique t elle que sur les objets neuf
{S}Bonjour,Pouvez-vous m'indiquer votre numéro de commande afin que je puis
{S}Bien sur.{S}Commande n° 15062120081RFQU du 21/06/2015Cdt
{S}Merci Tomislav, je vérifie auprès du Service Clients.{S} Marc
_____________________________
{S}Titre : Commande non expedié chez c discount
{S}Bonjour, J'ai eu le malheur de commander chez c discount un canape d'ang
et 22/04.{S} Le 23/04 toujours rien, j'ai appele plusieurs fois leur servic
commande avec un conseiller commercial au telephone .{S} En guise de dedomm
qu'entre le 6 et 10/05 je vais recevoir le colis.{S} Or je recois de nouvea
sera prolonge !{S} J'en ai plus qu'assez de leur site et leur service clien
demander un remboursement et la apparemment d'apres ce que j'ai lu c'est un
produit rapidemment et pourrais je reclamer un dédommagement ?
```

Figure 30 : Texte après la segmentation en phrases

```
_____________________________
Titre : Lybra 2.4 JTD LX (1999)
bonjours a tous je suis l'heureux propriétaire d'une lancia lybra 2.4 jtd 8cv est j'e
confortable aussi avec une consomation plus que raisonnable en moyenne 6.5 l au 100 k

_____________________________
Titre : Lybra 1.8 LX (2000)
bonjour à tousc'est une excellente voiture, confortable, maniable, avec du caractère
c'est presque une citadine…(en respectant les limitations de vitesses bien sure). et

_____________________________
Titre : Lybra 1.9 JTD LX (1999)
{S}Je possède cette voiture depuis deux ans et j'ai parcouru 50000Km sur une totalité
{S}Cette voiture de 135000Km acheté pour une véritable bouchée de pain offre le confo
chez Peugeot une 206 hdi 70 de société avec 200000Km...
{S}Certes le design est sobre mais plein de charme.{S}Le confort et l'équipement est
lombaire électrique. ordinateur de bord écran couleur. <voiture style="color: red;" c
{S}La finition est top, meilleur qu'une 307,406 ou megane de la même époque.
{S}Le <voiture style="color: red;" class='moteur' n='s'>moteur</voiture> fiable a du
plein a craquer.{S} Préférez le 2.4l JTD
{S}Voiture superbement entretenue et seulement une rotule et un triangle avant a chan
{S}Aucun regret, que du bonheur en plus elle fait tourner les têtes : on aime ou pas
```

Figure 31 : Erreur de segmentation pour les phrases commençant par une minuscule

- La deuxième opération consiste à enrichir le texte de catégories grammaticales (à ce niveau les dictionnaires de mots simples et de mots composés sont projetés sur le texte tokenisé pour reconnaître notamment les noms simples (*bateau*) et les noms composés (*bateau à voile*). Quelques grammaires locales sont également appliquées pour traiter des problèmes d'ambiguïté catégorielle récurrents.

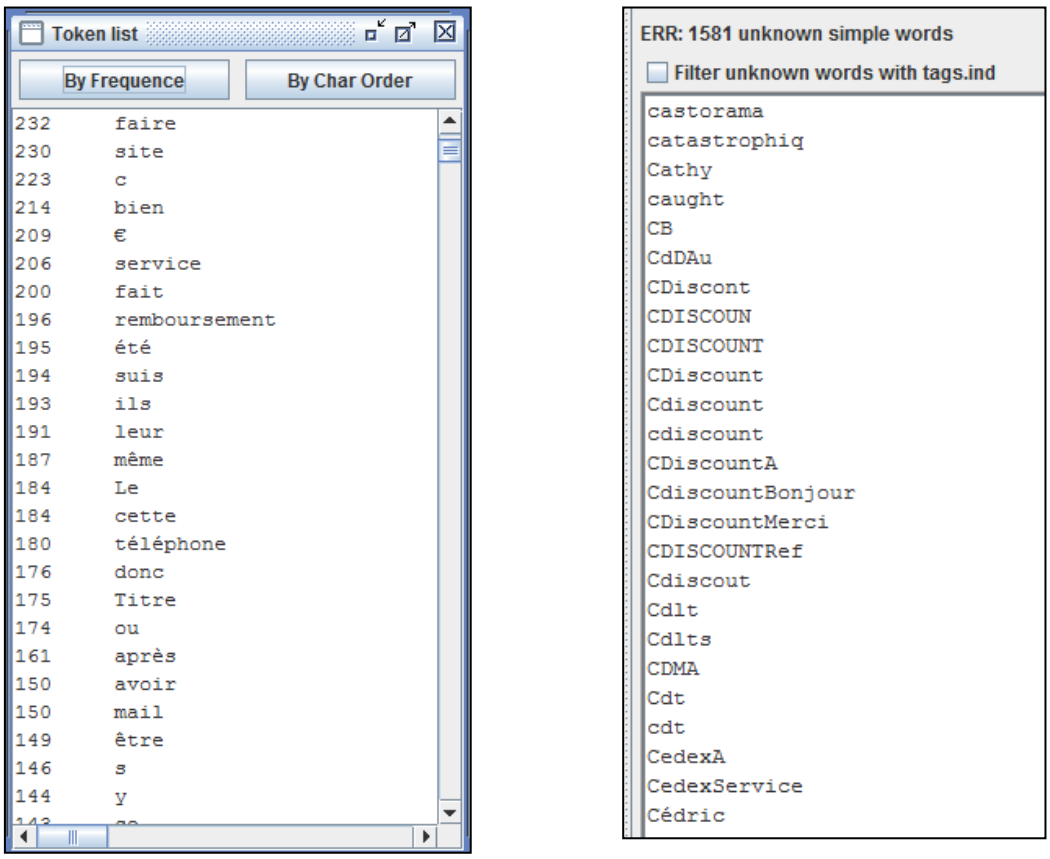


Figure 32 : Tokenisation et filtre des mots inconnus

- La troisième opération est l'application de la grammaire locale qui identifie les groupes nominaux formés d'un déterminant et d'un nom et qui leur associe une étiquette correspondante.

PARTIE 2. SYSTEME DE RÉSOLUTION DES ANAPHORES NOMINALES

Chapitre 4 : Résolution des anaphores nominales du type infidèles – sans l'identification des syntagmes verbaux

Chapitre 5 : Résolution des anaphores nominales du type infidèles – avec l'identification des syntagmes verbaux

Chapitre 6 : Résolution des anaphores associatives

CHAPITRE 4. ANAPHORES NOMINALES DU TYPE INFIDELE - SANS L'IDENTIFICATION DE SYNTAGMES VERBAUX

Nous présentons dans ce chapitre notre système de résolution des anaphores nominales du type infidèle. Nous avons essayé de traiter ce type d'anaphore avec deux méthodes différentes : une méthode simple (sans repérage des syntagmes verbaux) et une méthode plus élaborée (avec le repérage des syntagmes verbaux) qui sera présentée dans le chapitre suivant.

Nous présenterons le corpus utilisé pour cette méthode, les règles d'appariement, les stratégies de décision et le mode d'implémentation du système.

1. DESCRIPTION DU CORPUS

Notre système de résolution des anaphores nominales infidèles a été réalisé dans le cadre d'une recherche simple des antécédents pour les expressions anaphoriques repérées. Nous avons choisi un corpus qui se compose des articles de la rubrique faits-divers dans un journal en ligne.

Une anaphore du type infidèle est une anaphore où le nom de la forme de rappel est différent de celui de la forme introductrice. Le plus souvent, la différence est créée grâce à la synonymie ou l'hyperonymie. Le déterminant de la forme de rappel peut être différent de celui de la forme introductrice. C'est cette caractéristique de ce type d'anaphore qui nous a guidés à choisir un thème : les formes différentes de la classe <personne> dans les textes. Il s'agit d'un thème largement abordé, notamment dans les articles de journaux. C'est pour cette raison que nous avons décidé de travailler avec des corpus de faits-divers. La rubrique faits-divers a été choisie pour une autre raison : les articles sont souvent courts et riches en informations.

Notre corpus rassemble environs 300 articles sur le site http://www.sudinfo.be

Après avoir étés aspirés, les articles seront mis dans un fichier de texte avec leur titre, leur sous-titre éventuel, et leur contenu qui constitue le corps de l'article. Nous pouvons également obtenir les informations concernant l'auteur et la date de création de l'article, mais ces informations ne seront pas exploitées dans notre travail. Les articles sont séparés par un trait de séparation.

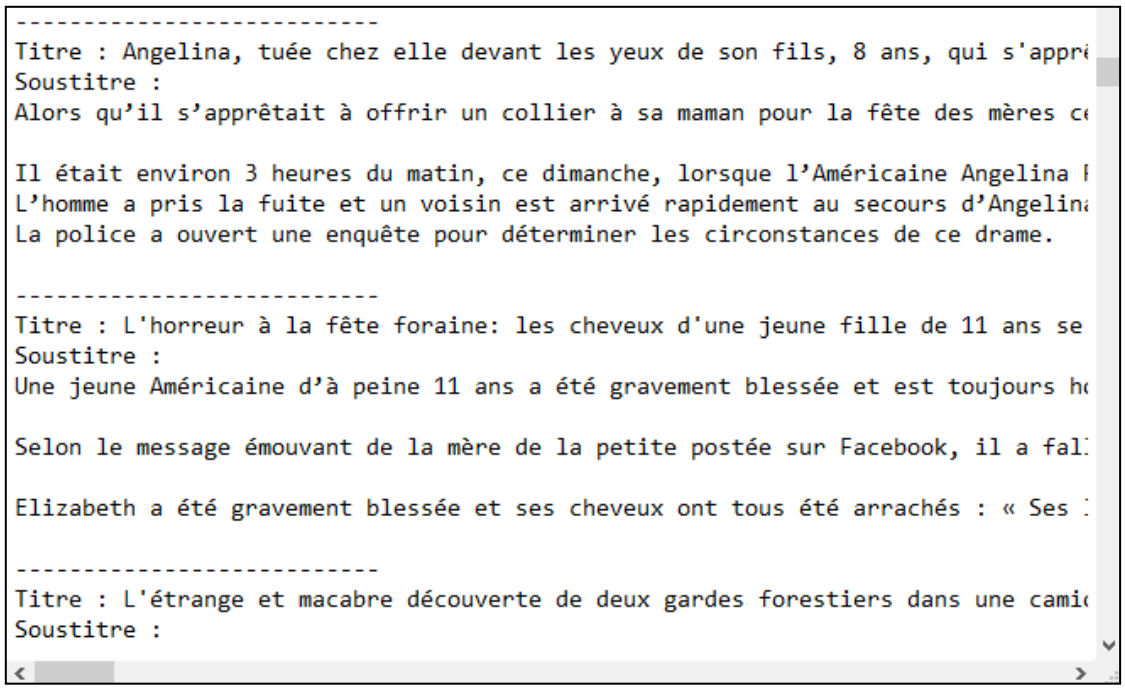


Figure 33 : Exemple d'un extrait du corpus

Appartenant au style journalistique, les articles aspirés sont correctement écrits, il n'y pas ou très peu de fautes d'orthographes.

2. REGLES D'APPARIEMENT

2.1. LE CALCUL DE LA SAILLANCE

Pour la méthode choisie, nous partons sur le principe général : trouver le GN antécédent relevant de la classe <Personne> pour un autre GN relevant de la même classe. Il faut que ces GN désignent la ou les mêmes personnes. Techniquement, nous identifions à la première étape tous les GN de la classe <Personne>. A la deuxième étape, nous associons les GN en identifiant des antécédents probables de chaque GN, grâce à certaines caractéristiques comme points de ressemblances ; leur distribution…

Pour pouvoir identifier les GN de la classe <Personne> et identifier leur ou leurs antécédents probables, nous devons trouver pour chaque GN une liste de GN antécédents candidats et choisir le meilleur candidat de la liste, en nous appuyant sur le calcul de la probabilité d'être retenu de chacun des éléments de la liste. La liste des GN est composée des GN qui apparaissent avant le GN analysé, à une distance (proximité) qui est généralement de 2 ou 3 phrases. La sélection du ou des meilleurs candidats dépend du seuil de sélection que nous avons fixé. Le seuil de sélection est modifiable. Nous appelons cette méthode : le calcul du poids de saillance des antécédents candidats.

Notre programme peut se diviser en deux étapes : le prétraitement et la résolution des anaphores nominales.

Le prétraitement

Le prétraitement est une étape importante dans cette expérience, qui demande l'élaboration de nombreux dictionnaires.

A cette étape, nous essayons d'identifier les GN de la classe <Personne>, ou de faire simplement une annotation du corpus en gardant toutes les informations concernant le GN : son genre, son nombre, son type de déterminant, son domaine (par exemple : il s'agit d'un nom de métier, d'une fonction professionnelle, d'une nationalité…). Cette étape contient aussi une procédure de désambigüisation au niveau morphologique et sémantique de certains GN en répondant à certaines questions : *Est-ce que l'annotation concerne un adjectif ou un nom ? Si c'est un GN, est-ce qu'il appartient vraiment à la classe <Personne> ? …*

La résolution

Cette étape est réservée à la comparaison du poids de saillance total de chaque élément d'une liste après avoir obtenu toutes les informations concernant les GN puis à la sélection de l'élément dont le poids de saillance est le plus élevé. En effet, pour chaque facteur caractéristique du GN, nous attribuons un poids de saillance

élémentaire. La saillance totale est la moyenne pondérée de tous les poids de saillance élémentaires de chaque GN. Ces facteurs, dont le poids est fixé arbitrairement, peuvent être :

- La proximité des antécédents candidats par rapport au GN analysé (Les antécédents candidats plus proches sont plus saillants)
- La distribution des antécédents candidats (les candidats qui se trouvent dans le titre de l'article sont plus saillants)
- L'accord en genre et en nombre des antécédents candidats et l'élément analysé. Si l'antécédent candidat et le GN analysé sont en accord en genre et en nombre, le GN analysé obtient des poids de saillance supplémentaires.

2.2. PROCEDURE DE DECISION

Pour sélectionner le ou les meilleurs candidats antécédents, nous éliminons les candidats dont le poids de saillance final est trop bas. Pour ce corpus, nous obtenons le meilleur résultat lorsque le poids de saillance est fixé à 0.8

Si aucun des éléments de la liste n'obtient ce seuil, nous pouvons dire que l'antécédent est inconnu ou silencieux.

Si plus d'un candidat est retenu, nous allons choisir le candidat ayant le poids de saillance le plus élevé.

Si deux éléments de la liste ont le même poids, nous choisissons l'élément le plus proche du GN analysé.

Pour la sortie, le programme affiche le GN analysé, puis l'ensemble des antécédents candidats avec leurs poids totals, classés dans l'ordre du plus saillant au moins saillant.

Nous avons testé différents tableaux de pois de saillance proposé, et chaque tableau donne un résultat différent. Les meilleurs résultats sont enregistrés dans le tableau suivant :

Indices	Poids de saillance proposé
Proximité = 0 (le GN et l'antécédent candidat se trouvent dans une même phrase)	0.8
Proximité = 1 (le GN et l'antécédent candidat se trouvent dans deux phrases contiguës)	0.6
Antécédent se trouvant dans le Titre	0.5
Accord en nombre entre GN et l'antécédent candidat	0.8
Accord en genre entre GN et l'antécédent candidat	0.7
Antécédent est en apposition avec GN	0.6
Antécédent et anaphore partagent les mêmes traits sémantiques	0.9
La séquence sémantique la plus longue	0.6

Tableau 6 : Les poids de saillance pour la résolution des anaphores infidèles - sans analyse de syntagmes vervaux

3. MODE D'IMPLEMENTATION

- L'extraction des GN

Le module d'extraction des GN concerne le balisage qui a lieu lors du premier parcours du corpus. Les GN sont étiquetés à l'aide de l'outil d'analyse syntaxique Unitex et des dictionnaires externes (les dictionnaires externes sont ceux qui ne sont pas fourni par l'outil Unitex que l'on doit intégrer).

La première étape de notre travail concerne l'annotation des GN à l'aide des graphes à état fini d'Unitex et nos dictionnaires de la classe <Personne> que nous avons élaboré préalablement. Nos dictionnaires permettent d'attribuer un champ sémantique à chaque GN. Nous disposons de 4 champs : les personne_famille *(*qui désigne le rôle des personnes dans une famille comme *père, mère, tante, femme, mari…)* ; les personne_metier *(*qui désigne le métier d'une personne comme *conducteur, étudiant…) ;* les personne_origine *(*qui désigne la nationalité d'une personne comme *français, américain…)* et les personne_generique *(*qui rassemble toutes les autres expressions désignant une personne, et qui ont, le plus souvent, une caractère générique comme *homme, femme, fille, garçon, jeune homme…)*

Par exemple :

> *Trois hommes âgés d'une vingtaine d'années ont été blessés par balle*

Le mot « hommes » est annoté comme un groupe nominal de personne, appartient à la classe « générique », du genre masculin, et du nombre pluriel, et sera annoté :

> *Trois <person class='generique' g='m' n='p'>hommes</person> âgés d'une vingtaine d'années ont été blessés par balle*

La désambigüisation est une étape importante dans notre travail, qui sera implémentée dans l'étape suivante.

L'annotation du corpus nous permet de stocker toutes les informations concernant le GN : son genre, son nombre, son type de déterminant, son champ sémantique…

L'extraction des GN implique aussi la **règle de la longueur maximale** du GN (les GN les plus longs seront repérés). Nous avons deux cas :

- Le premier cas concerne la structure GN + GN ou ADJ + GN, comme : *La jeune avocate avait été défigurée*. Dans ce cas, les informations syntaxiques du groupe de mot correspondent à celles du deuxième élément (*avocate*).

- Le deuxième cas concerne la structure récursive *GN de GN*, comme *La mère de la fillette* dans l'exemple : *La mère de la fillette a été arrêtée …* Dans ce cas, les informations syntaxiques du groupe *GN de GN* correspondent à celles du premier GN.

Si nous ne prenons pas en compte la règle de la longueur maximale du GN, la première phrase est analysée ainsi :

La <person style='color: red;' class='generique' g='m' n='s'>jeune</person> <person style='color: red;' class='metier' g='f' n='s'>avocate</person> avait été défigurée

Le module d'extraction des GN impliquant la règle de la longueur maximale du GN nous permet d'obtenir ce résultat :

la <person style='color: red;' class='metier' g='f' n='s'>jeune avocate</person> avait été défigurée

De même, si nous ne prenons pas en compte la règle de la longueur maximale du GN, la deuxième phrase pourrait être analysée ainsi :

La <person style='color: red;' class='famille' g='f' n='s'>mère</person> de la <person style='color: red;' class='generique' g='f' n='s'>fillette</person> a été arrêtée …

Mais avec la règle de la longueur maximale du GN, le résultat est devenu ainsi :

La <person style='color: red;' class='famille' g='f' n='s'>mère</person> de la fillette a été arrêtée …

- Le module de vérification et correction des informations morphosyntaxiques

Vient ensuite l'étape de vérification et de correction des annotations réalisées à l'étape précédente. Cette étape implique une procédure de désambigüisation au niveau morphologique et sémantique de certains GN.

Hors contexte, l'annotation des groupes nominaux peut se faire efficacement avec l'utilisation d'Unitex. Placée en contexte, l'analyse syntaxique devient, en revanche,

moins fiable car Unitex ne possède pas, en l'état actuel, d'un outil de désambigüisation. C'est pourquoi, nous reconnaissons la nécessité de vérifier et de corriger les erreurs de l'annotation automatique.

Certains GN peuvent avoir la même forme au singulier ou au pluriel (*français* par exemple), au masculin ou au féminin (par exemple : *enfant*) mais le fonctionnement de l'Unitex oblige un seul choix du dictionnaire. Par défaut, le GN est considéré comme un nom au singulier, ou au masculin. C'est la présence du déterminant (DET) précédant le GN qui nous permet de faire la distinction et la correction.

Le module de correction des informations morphosyntaxiques nous permet de distinguer les cas où le mot *français* est utilisé en tant qu'ADJ ou en tant que GN. Par exemple :

"*Il devrait prochainement regagner le territoire <person style='color: red;' class='origine' g='m' n='s'>français</person> pour y purger sa peine", a indiqué une source proche du dossier.*

L'absence du déterminant devant le mot *français* montre qu'il est utilisé dans cette phrase comme un ADJ et non pas comme un GN. Le module de correction automatique ne devrait attribuer aucune information morphosyntaxique au mot *français* dans ce cas :

Il devrait prochainement regagner le territoire français pour y purger sa peine", a indiqué une source proche du dossier.

Au cas où le déterminant existe, nous procédons à deux démarches :

- Définir préalablement le genre et le nombre du GN par défaut avec un dictionnaire d'Unitex.
- Corriger le genre et le nombre des GN à l'aide du déterminant, si nécessaire.

Dans le dictionnaire, le mot « *français* » est défini comme un GN de la classe <personne> au genre masculin et au nombre singulier. Pourtant « français » peut

aussi être au nombre pluriel. Nous devons nous fonder sur la distribution du mot ou le déterminant du GN de la classe <personne> pour vérifier le genre (et le nombre éventuellement) du GN. Nous procédons postérieurement à une correction automatique s'il s'agit d'une erreur dans l'annotation par défaut.

Par exemple, pour le mot « français », les dictionnaires Unitex nous ont fait apparaître des erreurs comme :

{S}Trois croix gammées, les inscriptions "La France aux <person style='color: red;' class='origine' g='m' n='s'>Français</person>" […] ont été tracées à la peinture noire sur la mosquée Al-Fath

Mais la présence du déterminant *aux* a permis à notre module de corriger l'annotation initiale des informations syntaxiques du GN *français.*

{S}Trois croix gammées, les inscriptions "La France aux <person det='def' g='m' nb='p' style='color: red;' class='origine'>Français</person>", […] ont été tracées à la peinture noire sur la mosquée Al-Fath

De la même façon, les autres déterminants nous ont permis également une correction automatique :

{S}Titre : Quatre morts en montagne, dont deux <person det='ind' g='m' nb='p' style='color: red;' class='origine'>Français</person> dans le massif des Ecrins

{S} Il a ajouté qu'il ne faisait "pas du tout confiance aux <person det='def' g='m' nb='p' style='color: red;' class='origine'>Français</person>"

Le mot *fils*, terminant par un 's', possède les mêmes caractéristiques :

L'un de ses <person det='poss' g='m' nb='p' style='color: red;' class='famille'>fils</person>, Anthony Moretti, mis en examen dans le <person det='def' g='m' nb='s' style='color: red;' class='metier'>cadre</person> de deux affaires criminelles, avait échappé à une tentative d'assassinat à Sartène en juin dernier

À l'intérieur du véhicule se trouvaient le <person det='def' g='m' nb='s' style='color: red;' class='famille'>mari</person> de Louise, ses deux <person det='ind' g='m' nb='p' style='color: red;' class='famille'>fils</person> âgés de 12 et 8 ans

L'un de ses <person det='poss' g='m' nb='p' style='color: red;' class='famille'>fils</person>, Anthony Moretti, mis en examen dans le <person det='def' g='m' nb='s' style='color: red;' class='metier'>cadre</person> de deux affaires criminelles, avait échappé à une tentative d'assassinat à Sartène en juin dernier

De plus, certains GN appartiennent à différents champs sémantiques (le plus souvent c'est personne_famille et personne_generique), nous devons privilégier un champ Par exemple, le nom *femme* appartient à la fois à la classe personne_famille et à la classe personne_generique. Notre algorithme de décision est réalisé ainsi :

- Dans une distance de 2 phrases en amont et en aval du GN en question, s'il existe d'autres GN de la classe personne_famille (par exemple : *mari*), cette classe sera choisie. Sinon, la classe personne_generique sera choisie par défaut.

- Attribution du poids de saillance

La troisième étape concerne l'attribution du poids de saillance pour chaque GN repéré. Mais avant d'attribuer un poids de saillance, nous distinguons deux types de GN : logiquement, le premier GN d'un article n'a pas d'antécédent, il ne sera pas traité comme expression anaphorique. Les autres GN de l'article sont capables de recevoir un antécédent possible, donc ils peuvent être une expression anaphorique. Nous devons procéder à la recherche de leurs antécédents probables.

Le processus de recherche d'antécédent est réalisé avec l'algorithme suivant :

- La recherche d'antécédent doit être réalisée pour tous les GN de l'article sauf le premier de l'article. Nous avons donc besoin d'établir la liste des GN concernés. La longueur de ces listes change en fonction de la distance choisie. La distance (ou la notion de proximité) est une notion paramétrable dans notre

système. Par exemple, si la distance entre l'expression anaphorique et l'antécédent candidat est fixée à 2 phrases, nous devons établir toutes les listes de GN contenus dans tous les espaces de 2 phrases, en n-gramme.

(Phrase1 (Phrase2 (Phrase3) Phrase4) Phrase5)

Liste GN 1 = les GN contenus dans Phrase 1 + Phrase 2 + Phrase 3

Liste GN 2 = les GN contenus dans Phrase 2 + Phrase 3 + Phrase 4

Liste GN 3 = les GN contenus dans Phrase 3 + Phrase 4 + Phrase 5

Et ainsi de suite.

- Par la suite, chaque élément d'une liste sera analysé comme expression anaphorique possible. Nous procédons alors à une nouvelle étape visant à lui trouver un antécédent possible en nous basant sur le calcul du poids de saillance de tous les autres GN de la liste. En termes d'algorithme, cette étape est réalisée par ordre inversé : le dernier GN d'une liste sera traité en premier.
- Les GN d'une liste, sauf celui en phase d'analyse, vont recevoir des pois de saillance équivalents à sa probabilité d'être retenu en antécédent. Le tableau des poids de saillance contient les facteurs de saillance élémentaires. Le poids de saillance final d'un GN correspond à la moyenne pondérée de tous ses poids de saillance élémentaires.

- Le calcul de score

Comme entrée du module de calcul de score, on prend la sortie du module d'extraction. Son but est d'extraire des couples formés d'un GN analysé comme une expression anaphorique et d'un autre GN retenu comme son antécédent potentiel et d'afficher son meilleur score.

Nous devons, à cette étape, comparer le poids de saillance total de chaque élément de la liste des GN en question. Pourtant, nous n'affichons pas tous les GN avec leur poids de saillance final, mais seulement les meilleurs candidats qui ont un score plus élevé que le seuil fixé.

Le module fournit en sortie une nouvelle structure de données : pour chaque GN analysé, nous affichons la liste des GN antécédent potentiels dont le score dépasse le seuil.

- Affichage des résultats

Après avoir trouvé les meilleurs antécédents candidats pour chaque GN, il nous reste à afficher le résultat du traitement. Pour une question pratique, nous préférons afficher les résultats dans un tableau, mais nous avons également choisi d'afficher les résultats dans le corpus-même pour obtenir une bonne visibilité et un meilleur suivi des erreurs.

Dans l'affichage des résultats, nous énumérons tous les GN d'un article en couleur rouge, cette énumération permet l'identifiant de chaque GN. Si nous trouvons des antécédents pour ce GN, nous affichons les informations concernant les antécédents en bleu. Ces informations se composent du numéro d'identifiant de ce GN et son poids, par ordre de privilège descendant. Au cas où nous n'avons pas trouvé l'antécédent pour un GN, nous n'affichons aucune information.

Les paramètres sont au nombre de 5, au total :

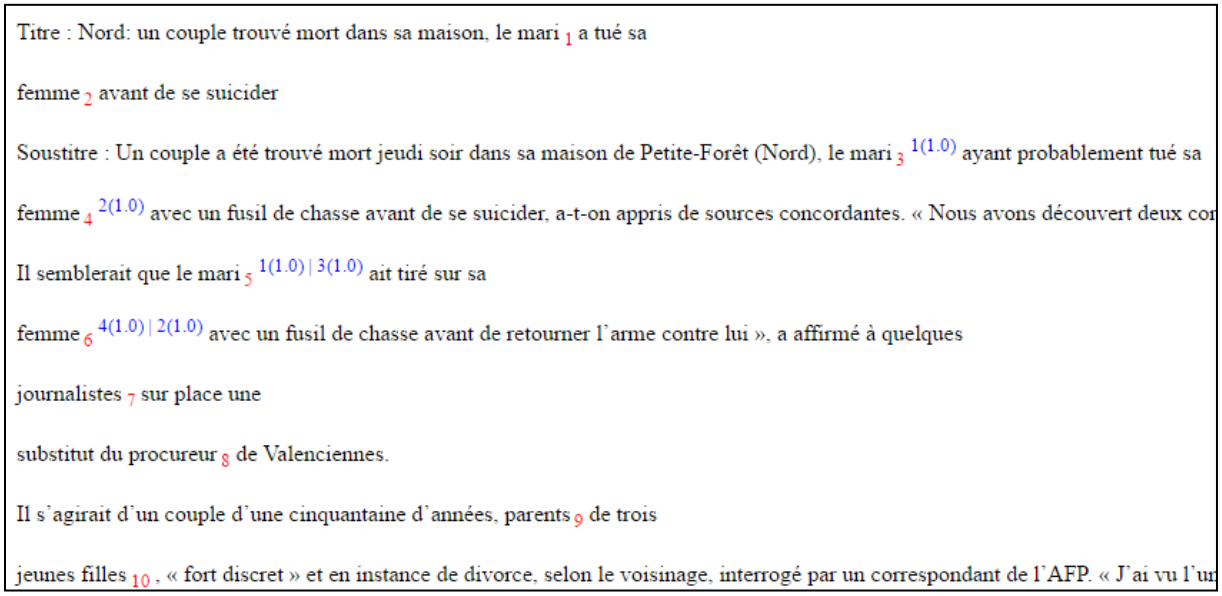
La classe du GN

Le genre du GN

Le nombre du GN

La distance entre un GN et son antécédent potentiel

Le type de déterminant du GN


Titre : Nord: un couple trouvé mort dans sa maison, le mari 1 a tué sa

femme 2 avant de se suicider

Soustitre : Un couple a été trouvé mort jeudi soir dans sa maison de Petite-Forêt (Nord), le mari 3 1(1.0) ayant probablement tué sa

femme 4 2(1.0) avec un fusil de chasse avant de se suicider, a-t-on appris de sources concordantes. « Nous avons découvert deux cor

Il semblerait que le mari 5 1(1.0) | 3(1.0) ait tiré sur sa

femme 6 4(1.0) | 2(1.0) avec un fusil de chasse avant de retourner l'arme contre lui », a affirmé à quelques

journalistes 7 sur place une

substitut du procureur 8 de Valenciennes.

Il s'agirait d'un couple d'une cinquantaine d'années, parents 9 de trois

jeunes filles 10 , « fort discret » et en instance de divorce, selon le voisinage, interrogé par un correspondant de l'AFP. « J'ai vu l'ur

Figure 34 : L'affichage du résultat au format html.

CHAPITRE 5. ANAPHORES DU TYPE INFIDELE – RESOLUTION AVEC L'IDENTIFICATION DES VERBES PREDICATIFS.

Nous souhaitons, dans ce chapitre, parler de notre système de résolution des anaphores du type infidèle avec la deuxième méthode : l'utilisation des patrons syntaxiques pour le repérage des antécédents. Nous allons aborder de divers aspects : la description du corpus choisi pour ce type d'anaphore ; les règles d'appariement (ou les paramètres utilisés pour traiter des anaphores infidèles) ; et l'implémentation de ses modules de traitement.

1. DESCRIPTION DU CORPUS

Nous avons choisi un corpus de test qui se compose d'articles aspirés automatiquement sur le site en ligne du journal http://www.lemonde.fr/ à la rubrique faits-divers.

La plupart de ces articles dans ce journal parlent des accidents, des rixes entre individus, ou des enquêtes policières. Nous essayons de limiter la recherche des mots désignant la classe <humain> (équivalent de la classe <personne> dans la résolution des anaphores infidèles sans l'identification des verbes) dans ce corpus aux victimes et aux auteurs de l'acte. Précisément, nous traitons uniquement les reprises de la classe <Victime> et nous essayons de trouver leurs antécédents.

Prenons un exemple :

« ***Un homme*** *a été grièvement blessé à la gorge lors d'une rixe entre* ***passagers*** *éméchés, mercredi 2 octobre au soir dans un TGV près de la gare du Creusot (Saône-et-Loire).* ***La victime****, dont le pronostic vital n'est pas engagé, selon la SNCF, a été transportée consciente à l'hôpital du Creusot.* »

La victime est la reprise

un homme est l'antécédent à trouver.

Après avoir étés aspirés, les articles seront mis dans un fichier texte avec leur titre, leur sous-titre éventuel, et leur contenu qui est le corps de l'article. Nous pouvons également obtenir les informations concernant l'auteur et la date de création de l'article, mais ces informations ne seront pas exploitées dans notre travail. Les articles sont séparés par un trait de séparation.

Les articles aspirés sont correctement écrits, il n'y pas ou très peu de fautes d'orthographes

2. RÈGLES D'APPARIEMENT

2.1. LE CALCUL DE SAILLANCE

Comme entrée du module de calcul de saillance, on prend la sortie du module d'extraction des GN. Son but est d'apparier deux syntagmes nominaux dont l'un étant le GN analysé comme expression anaphorique et l'autre comme son antécédent potentiel puis d'afficher leur degré d'association avec le score de saillance dépassant un seuil fixé.

Nous combinons cette fois deux types de méthodes : des patrons syntaxiques et des méthodes basées sur le calcul des points de ressemblance entre deux GN. En sortie, ce module fournit une nouvelle structure de données : pour chaque syntagme nominal analysé, nous associons les GN antécédents qui lui sont associés, ces GN seront retenus en raison de leur meilleur score.

Nous partons sur le principe : trouver le GN antécédent relevant de la classe <Humain> pour un GN relevant de la classe <Victime>. Autrement dit, nous essayons de trouver tout d'abord les expressions qui désignent la ou les victimes dans les articles. Nous essayons, par la suite, d'identifier les expressions qui désignent soit la victime elle-même mais exprimée par une autre expression), soit l'agresseur qui provoque l'accident ou l'acte criminel. Techniquement, nous souhaitons affiner nos règles d'association en nous appuyant non seulement sur les caractéristiques de l'expérience réalisée auparavant (anaphores nominales du type infidèle - sans analyse de verbes prédicatifs) mais aussi d'autres restrictions concernant la présence de certaines formes verbales et non verbales.

Notre travail se divise en deux étapes :

- L'identification des antécédents pour un GN relevant de la classe <Victime> (l'antécédent peut être la victime ou l'auteur de l'agression), en nous basant sur le mode de calcul de la saillance.
- L'attribution du rôle à chaque antécédent trouvé (victime ou agresseur ?).

Pour la première étape, chacun des antécédents potentiels va se voir attribuer une pondération de saillance correspondant à sa probabilité d'être remarqué en premier, comme dans l'expérience 1. Le calcul de la saillance dépend aussi des facteurs de saillance et le poids final de la saillance est compté avec la moyenne pondérée de tous les poids de chaque facteur. Les facteurs de saillance peuvent être :

- La proximité des antécédents candidats de la classe <Humain> par rapport à la classe <Victime> (Les antécédents candidats plus proches sont plus saillants)
- La distribution des antécédents candidats (les candidats qui se trouvent dans le titre de l'article sont plus saillants)
- L'accord en nombre entre l'antécédent candidat et les termes de la classe <Victime>, s'il s'agit d'identifier la victime. Par exemple :

  Un couple […]. Les victimes […]

  Un couple […]. La victime […]

  Le premier exemple, *les victimes* est au pluriel et *un couple* est aussi au pluriel, ils sont en accord en nombre, alors on peut attribuer un poids de saillance au GN *un couple*. Mais dans le deuxième exemple, *la victime* et *un couple* ne s'accordent pas, on ne peut pas donner un poids de saillance au GN *un couple.*
- Aucun accord en genre et en nombre ne sera imposé s'il s'agit d'identifier l'agresseur d'une victime, car ce peut être deux personnes différentes.
- Le traitement du pluriel est intégré avec l'identification des GN du type : *un couple de, deux hommes, le quinquagénaire et son épouse, les deux victimes, ces victimes...*

- L'identification des fonctions syntaxiques des GN est intégrée grâce à l'identification des verbes et des prépositions. Par exemple :

  ***Joe**, **la victime**, souffre beaucoup après avoir été hospitalisé.*

  Lorsqu'un antécédent candidat garde la fonction « sujet », il a plus de poids de saillance que lorsqu'il garde la fonction objet.
- L'identification de l'état apposition d'un GN grâce à la présence de la virgule. Si l'antécédent est en apposition avec <victime>, nous accordons un poids de saillance.

Pour la deuxième étape, l'analyse des verbes prédicatifs est introduite afin d'identifier les auteurs et les victimes. Le type des verbes prédicatifs et leur distribution jouent un rôle important dans l'identification des causateurs et des victimes. Nous avons analysé deux types de verbes : les verbes associés aux auteurs et les verbes associés aux victimes.

Par exemple pour les verbes à la forme active :
[Causateur] + *[tuer | assassiner | agresser …]*
[Victime] + *[retrouver (mort) |mourir|subir (les coups)]*
Pour les verbes à la forme passive, la distribution des verbes peut être changée, par exemple :
[Victime] + *[être tué(es) | être assassiné(es) | agressé(es) …]*

Par exemple :

Un homme a poignardé une femme […] La victime a été grièvement blessée […]

Le verbe *[poignarder]* est à la forme active. Le GN de la classe <Humain> avant le verbe est identifié comme auteur probable et le GN de la classe <Humain> après le verbe est identifié comme victime probable.

Le verbe *[blesser]* est à la forme passive. Le GN de la classe <Humain> avant le verbe est identifié comme victime probable.

Si le verbe indiquant l'acte du crime n'est pas repéré, le GN de la classe <Humain> peut être considéré comme appartenant à la classe <Victime>

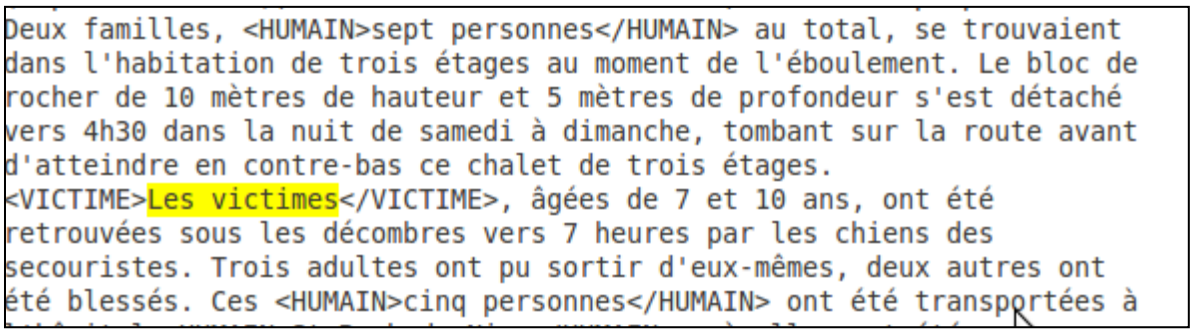


Figure 35 : Le verbe indiquant l'acte du crime n'est pas repéré

2.2. PROCÉDURE DE DÉCISION

Nous obtenons le meilleur résultat lorsque le poids de saillance final est fixé à 0.5, nous éliminons les candidats dont le poids de saillance final est en dessous de 0.5. Si plus d'un candidat est retenu avec un poids de saillance au dessus du seuil, nous allons choisir le candidat ayant le poids de saillance le plus élevé. Si deux éléments de la liste ont le même poids, nous choisissons l'élément le plus proche du GN analysé.

Si aucun des éléments de la liste n'obtient ce seuil, il peut y avoir deux cas :

- Soit le GN n'a pas d'antécédent
- Soit la reconnaissance est manquante (le cas du silence).

Le tableau de poids de saillance que nous proposons ci-dessous nous a donné les meilleurs résultats :

Indices	Poids de saillance proposé
Proximité = 0 (Candidat se trouve dans la phrase	0.5

où se trouve le nom anaphorique (S))	
Proximité = 1 (Candidat se trouve dans la phrase qui précède le nom anaphorique (S-1))	0.4
Proximité = 2 (Candidat se trouve dans la phrase avant (S-2))	0.3
Antécédent se trouvant dans le Titre	0.8
Accord en nombre entre « Victime » et Antécédent candidat	0.3
Antécédent est en apposition avec « Victime »	0.7
Antécédent candidat est sujet	0.8
Antécédent candidat est attribut	0.7
Antécédent candidat est COD	0.6
Antécédent candidat est dans la proposition subordonnée	0.3
Le déterminant de l'antécédent candidat est possessif	-0.6

Tableau 7 : Poids de saillance pour la résolution des anaphores infidèles - avec analyse de SV

Nous avons ajouté un critère pour éliminer les faux candidats. Le critère concerne le déterminant de l'antécédent candidat. Les travaux de Buvet (Buvet, 2015) nous ont guidée dans ce choix. Par exemple :

*Le seul défaut qui dessert **le pc**, c'est **son disque dur***

Dans cet exemple, le GN *pc* a un déterminant défini et le GN *disque dur* a un déterminant possessif. La seule interprétation possible : *le disque dur* appartient au *pc* comme une partie de son ensemble.

Si l'on change le déterminant, l'exemple sera interprété différemment :

*Le seul défaut qui dessert **son pc**, c'est **le disque dur** (*)*

le disque dur dans ce cas peut être compris comme un objet indépendant.

3. MODE D'IMPLEMENTATION

Notre étape de prétraitement est réalisée à l'aide d'Unitex.

- Nous identifions d'abord les termes de la classe <Victime> : les termes « Victime » sont identifiés à formes définies (la/les/ces/cette + « victime(s) » par exemple), simplement avec les graphes d'Unitex.

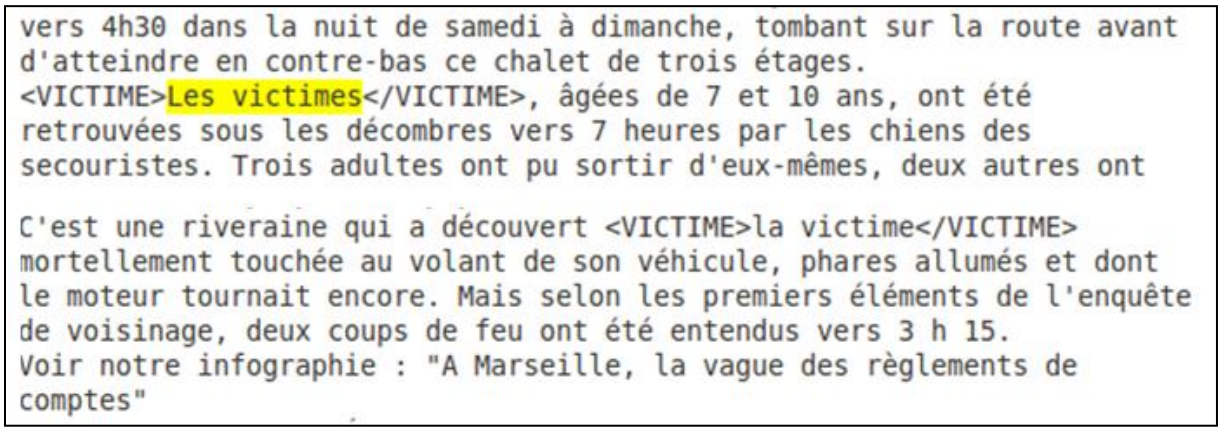


Figure 36 : Identification de la classe <Victime>

- Nous identifions ensuite tous les groupes nominaux évoquant les personnes grâces aux graphes d'Unitex et à nos dictionnaires spécifiques. Et à la sortie nous obtenons des noms communs et noms propres de personnes :

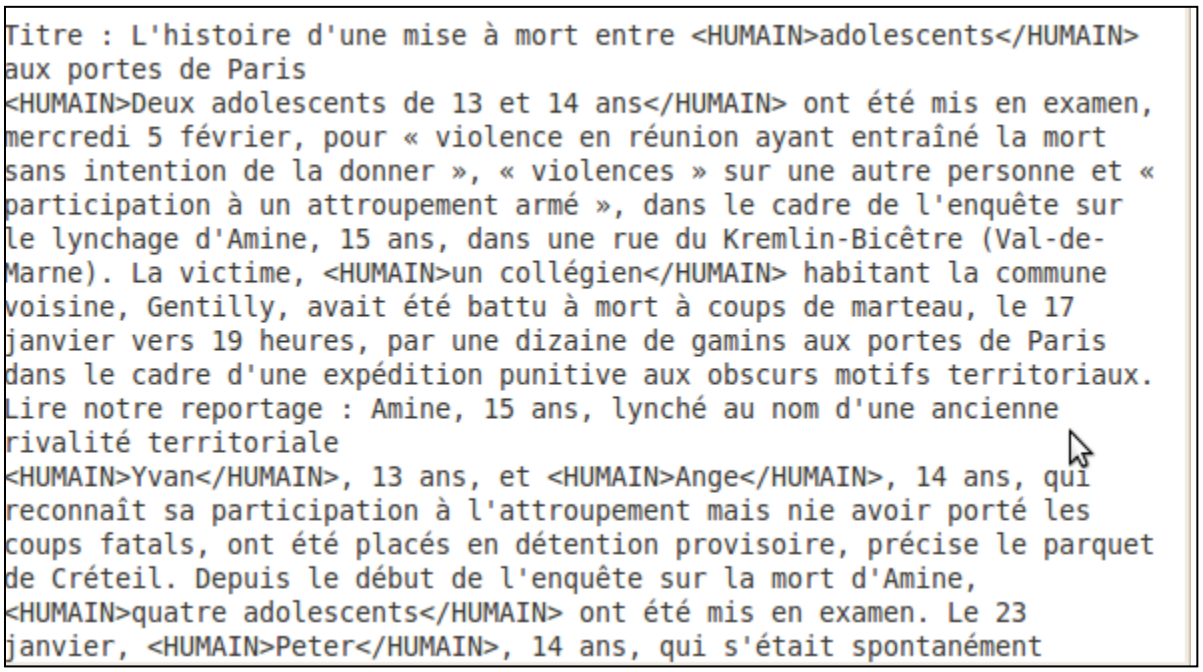


Figure 37 : Annotation des noms communs et noms propres

Nous avons implémenté un dictionnaire de noms propres dans l'extraction des GN. Notre dictionnaire contient uniquement des prénoms. Nous limitons notre extraction de noms propres aux prénoms et rejetons les structures composées (du type *Prénom* + *Nom* ou *Nom* + *Prénom)* car les structures composées ne nous ont pas donné de bons résultats : nous avons obtenu beaucoup de bruits pour un nombre assez limité d'occurrences, le style du corpus (faits-divers) impose.

- Notre travail consiste ensuite à identifier différentes formes (verbales et non verbales) marquant l'existence de l'acte du crime. Pour cela, nous avons élaboré différentes listes, chaque liste est marquée avec un code différent.

- Des formes verbales :

  Formes actives

  <Agresseur> + verbe : *prendre la fuite, commettre un meurtre…*

  <Agresseur> + verbe (+ <Victime>) : *tuer, agresser, menacer, brûler, poignarder, massacré …*

  Verbe + <Agresseur> : *menotter, attraper, poursuivre*

  <Victime> + verbe : *se suicider, retrouver mort*

(<Agresseur> +) ? verbe + <Victime> : *tuer, agresser, menacer, brûler, poignarder, massacré, piéger, filmer …*

Formes passives

<Victime> + verbes (+ <Agresseur>) ? : *être tuer par, être agresser par, être menacer par …*

<Victime> + verbes : *être retrouver mort ; (le corps) être découvert*

- Des formes non-verbales :

  Forme non-verbale + <Victime> : *La mort de, le corps de…*

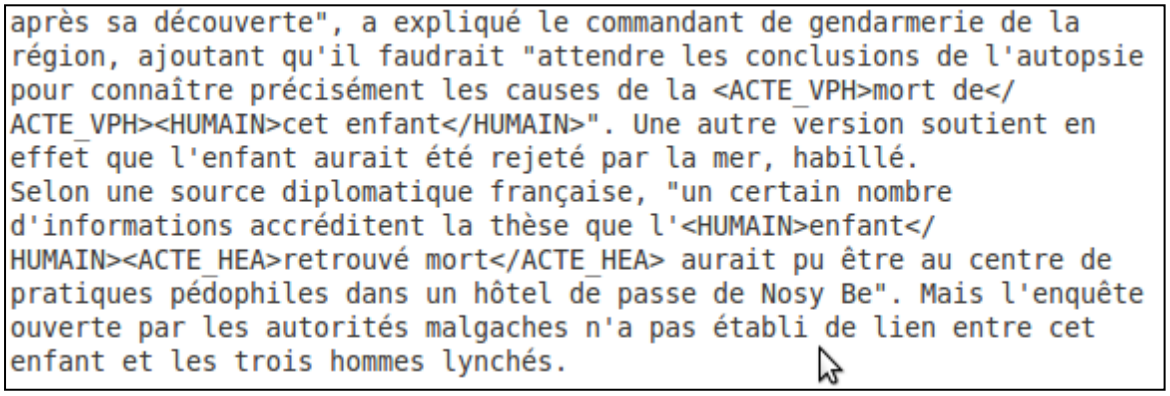


Figure 38 : Extraction des formes non-verbales

Après avoir obtenu un corpus annoté lors de l'étape de prétraitement, nous passons à l'étape suivante : le calcul du poids de saillance avec l'application de l'algorithme de recherche d'antécédents (intégré en Python). Le corpus est mis au format XML qui facilite la structuration des données, on numérote les articles et des lignes dans chaque article.

L'algorithme implique les choix suivants :

- Identification de certains antécédents grâce uniquement aux graphes Unitex, c'est le cas des règles grammaticales sûres et visibles qui permettent le choix. Par exemple : <Victime> *a été tué(es) par* <Agresseur>
- Pour les antécédents qu'on ne peut pas identifier avec les graphes, une autre étape de sélection aura lieu : le calcul de la saillance pour tous les

antécédents qui sont relativement proche de la classe <Victime> (nous avons choisi la distance de 2 et 3 phrases.

- La procédure de décision à l'aide de la table du poids de saillance pour sélectionner l'élément préféré d'une liste des candidats antécédents.

Résultats :

```
Titre 1 - victimes : Trois hommes âgés d'une vingtaine d'années#1.0 |
Brice Robin#0.3
Titre 2 - la victime :  passagers#1.0 | Un homme#1.0
Titre 3 - la victime : Un homme de 34 anss#1.0 | un homme#0.5
Titre 4 - les victimes : deux enfants#1.0 | sept personnes#0.3 |
pompiers#0.3
Titre 5 - la victime : une fillette de 12 ans#1.0 | Un adolescent#0.5 |
Un adolescent de 16 ans#0.3 | enfants#0.17
Titre 6 - la victime : Un homme#1.0 | Un homme d'une trentaine
d'années#1.0
Titre 7 - la victime : Bernard Mazières#1.0 | Le fils#0.5 | Lucas#0.3 |
Lucas Mazières#0.3 | Dany Manfoumbi#0.3 | femme de ménage#0.17|
mineurs#0.17 | Dany Manfoumbi#0.17
Titre 8 - victime : une septuagénaire#0.3 | pompiers#0.3 | les
pompiers#0.17
```

Figure 39 : Affichage du résultat - format tableau

Quelques remarques après la réalisation de l'expérience :

1. La reconnaissance des actes du crime est un bon appui pour le choix de l'antécédent.

2. Les erreurs sont dues principalement à la méconnaissance ou la mauvaise reconnaissance des termes désignant les humains.

3. Le traitement du pluriel n'est pas tout à fait satisfaisant, il demande une analyse plus profonde.

Pour améliorer le résultat, nous pourrions :

- Améliorer le traitement du pluriel

- Réaliser une stop-list pour éliminer certains cas d'erreur comme : victimes d'inondations, victimes de la discrimination etc.

- Appliquer le même travail pour traiter d'autres types d'anaphore ou pour traiter d'autres sujets

CHAPITRE 6. ANAPHORES DU TYPE ASSOCIATIF

Nous envisageons, dans ce chapitre, d'exploiter les occurrences des anaphores associatives du type méronymique relevées dans le corpus et de procéder une résolution en cherchant leur antécédent probable. Nous nous fondons sur la théorie des classes d'objets afin de décrire le type de relation anaphorique établie entre l'expression anaphorique et son antécédent.

1. DESCRIPTION DU CORPUS

Le trait définitoire des anaphores associatives nominales méronymiques réside dans le statut du nom anaphorique qui doit être marqué sémantiquement comme étant une partie d'un tout. Notre étude, s'inscrivant dans la perspective du traitement automatique des anaphores associatives, insiste sur la description du fonctionnement syntaxique, sémantique et référentiel des relations anaphoriques qui entretiennent un rapport méronymique (ou la relation du type parti – tout) entre les expressions anaphorique et les antécédents.

L'algorithme de résolution des anaphores nominales de type associatif est un algorithme qui, après l'analyse d'un corpus non structuré aspiré sur les pages web, vise à repérer automatiquement les antécédents et les associer aux reprises dans les structures d'anaphores associatives méronymiques (relation du type partie-tout entre l'antécédent et l'expression anaphorique).

Nous avons choisi de travailler sur un corpus provenant des avis de consommateurs sur des produits électroménagers ou des produits hi-tech, aspiré sur le site :

http://www.60millions-mag.com/forum/

Le corpus des textes étudiés qui viennent exclusivement de forums concerne des avis de consommateurs sur différents types de produit qu'on appelle les « artefacts ». Nous avons remarqué que la présence des paires : *artefacts-parties* est assez régulière dans ce type de corpus, la classe <artefact> est devenu ainsi un outil intéressant.

Sur le site, les avis sont regroupés par famille de produit concerné, par site de vente (vendeur). Nous pouvons aussi obtenir un corpus spécialisé en procédant une aspiration par mot-clé.

Pour détecter les thématiques principales de chaque texte (de quels produits parlent les consommateurs), il faut identifier des noms d'artefact (par exemple, *casserole*) ou

des noms de marque (par exemple, *Tefal*), et les catégoriser en leur ajoutant une étiquette sémantique.

L'étiquetage des noms d'artefact et des noms de marque n'est cependant pas suffisant pour la détection d'une thématique principale dans un texte, il faut aussi dégager des chaînes de référence. Pour ce faire, une des pistes envisagées est de détecter toutes sortes de relations transphrastiques. Parmi celles-ci, il y a l'anaphore associative méronymique. Ce qui implique de :

- détecter des groupes nominaux définis et non définis;
- établir comment ces groupes nominaux sont impliqués dans des relations du type anaphore associative;
- s'appuyer sur les groupes nominaux formés de noms d'artefact en précisant lesquels fonctionnent comme des holonymes et lesquels fonctionnent comme des méronymes.

2. RÈGLES D'APPARIEMENT

Ce travail, qui repose principalement sur la théorie des classes d'objets, demande des ressources lexicales très riches. Pour atteindre nos objectifs, nous devons construire de nombreux dictionnaires qui servent à attribuer des informations sémantiques aux GN. Nous avons deux groupes d'étiquettes : le premier groupe concerne la classe de l'artefact, c'est-à-dire, de quel type d'artefact il s'agit. Le deuxième groupe concerne le domaine de l'artefact, c'est-à-dire, le domaine d'usage de l'artefact.

Pour obtenir le plus de détails possible sur le type d'artefact, nous avons insisté sur trois niveaux de précisions de ces étiquettes. En effet, concernant la classe de l'artefact, nous pouvons avoir : l'hyperclasse (*appareil, organe*..), la classe (*appareil_signalisation, appareil_cuisson, organe_prehension*…) et la sous-classe (*appareil_cuisson_rechaud*…).

Par exemple, un *four micro-onde* doit avoir :

L'hyperclasse = Appareil

La classe = Appareil_cuisson

La sous_classe = Appareil_cuisson_rechaud

Comme un artefact peut appartenir à différents domaines, nous nous sommes limités à trois domaines.

Par exemple, une *machine à café* peut appartenir au :

Domaine 1 = Alimentation

Domaine 2 = Equipement_menager

La résolution de ce type d'anaphore se fonde sur la recherche des antécédents candidats pour une expression anaphorique repérée. La recherche d'antécédent repose non seulement sur la distribution des GN mais aussi sur la comparaison des informations sémantiques obtenues. Pour deux GN, plus ils ont de points communs, plus ils ont de chance d'être en relation.

Notre méthode pour traiter ce type d'anaphore a été modifiée par rapport aux méthodes précédentes : nous avons combiné la comparaison des informations d'ordre sémantique avec la méthode de recherche d'antécédent (le calcul du poids de saillance), avec le tableau de l'attribution du poids de saillance suivant :

Indice	Poids de saillance proposé
Proximité = 0	0.8
Proximité = 1	0.7

Proximité = 2	0.5
Reprise et Antécédent candidat ont la même Hyperclasse	0.7
Reprise et Antécédent candidat ont la même Classe	0.8
Reprise et Antécédent candidat ont la même Sous-classe	0.9
Reprise et Antécédent candidat ont le même Domaine 1	0.7
Reprise et Antécédent candidat ont le même Domaine 2	0.6
Reprise et Antécédent candidat ont le même Domaine 3	0.5

Tableau 8 : Poids de saillance pour la résolution des anaphores associatives

Par rapport aux tableaux utilisés dans les méthodes précédentes, nous n'avons pas gardé ici les informations concernant la comparaison du genre et du nombre des GN. En effet, le GN désignant un artefact et le GN désignant sa partie partagent un lien sémantique mais par forcément de lien syntaxique.

De plus, avec cette méthode, nous devons préciser quel est le type de relation qui existe entre eux : ils sont *méronymiques* ou *synonymiques* ? Nous avons certaines règles pour cela. En effet, la relation entre une reprise et un antécédent est du type méronymique si :

- le déterminant de l'antécédent candidat est du type Un N et le déterminant de la reprise est du type possessif. Par exemple :
  *J'ai acheté **une montre** […]. **Sa trotteuse** ne bouge pas.*
- l'antécédent candidat et la reprise n'appartiennent pas à la même hyperclasse et un des deux appartient à l'hyperclasse <Appareil>.

En revanche, la relation entre une reprise et un antécédent est du type synonymique si :

- l'antécédent candidat et la reprise appartiennent à la même hyperclasse.
- le déterminant de l'antécédent candidat est du type Un N et le déterminant de la reprise est du type Ce N. Par exemple :
  *On vient de me livrer **une machine à expresso** […].Je souhaite bien me rétracter et avoir un remboursement car **cette machine** tombe déjà en panne après 2 jours.*

3. MODE D'IMPLEMENTATION

3.1. PRÉTRAITEMENT

Le corpus brut va d'abord passer à l'étape d'annotation grâce aux graphes d'Unitex et à différents dictionnaires d'artefact. Il s'agit de classifier tous les syntagmes nominaux selon leur type d'artefact (appareil ou accessoires) et leurs domaines, grâce aux étiquettes :

- *l'hyperclasse (HC) :* organe, dispositif, appareil…
- *la classe (CC)* : partie_vetement, partie_generique…

- *la sous-classe (SC)*
- *le domaine1 (D1)*
- *le domaine2 (D2)*
- *le domaine3 (D3)*

Généralement, chaque GN sera annoté avec au moins une hyperclasse et un domaine. Plus les informations sont nombreuses, plus la comparaison est précise.

Les informations concernant le déterminant dans le GN, comme le type du GN (déterminant défini ou indéfini), le sous-type du GN (ce N, le N, un N…), sont aussi annotées.

Après l'annotation, le corpus brut sera transformé au format XML qui affiche aussi les informations comme : le numéro du commentaire, le numéro de la phrase.

```
1  <?xml version='1.0' encoding='UTF-8' ?>
2  <document>
3          <message id='1'>
4                  <phrase id='1'>
5  Je viens de recevoir <groupe_nominal type='DEF' sous_type='CEN'><determinant type='CE'>cette</determinant><nom
   hyperclasse='APPAREIL' classe='APPAREIL_SIGNALISATION' sous_classe=''  domaine1='TECHNIQUES' domaine2='VIE_QUOTIDIENNE'
   domaine3=''>montre</nom></groupe_nominal> mais celle-ci ne fonctionne pas, j'ai beau essayer de la régler mais rien ne se
   passe, <groupe_nominal type='DEF' sous_type='LEN'><determinant type='LE'>la</determinant><nom hyperclasse='ORGANE'
   classe='ORGANE_SIGNALISATION' sous_classe=''  domaine1='TECHNIQUES' domaine2='VIE_QUOTIDIENNE' domaine3=''>trotteuse </
   nom></groupe_nominal> ne bouge pas, et <groupe_nominal type='DEF' sous_type='LEN'><determinant type='LE'>l'</
   determinant>heure</groupe_nominal> reste figée.
6                  </phrase>
7          </message>
8          <message id='2'>
9                  <phrase id='1'>
10 J'ai acheté <groupe_nominal type='NON_DEF' sous_type='DNUMN'><determinant type='DNUM'>une</determinant><nom
   hyperclasse='APPAREIL' classe='APPAREIL_CUISSON' sous_classe='CAFETIERE'  domaine1='ALIMENTATION'
   domaine2='TRAVAUX_ET_EQUIPEMENTS_MENAGERS' domaine3=''>machine à expresso </nom></groupe_nominal> mais <groupe_nominal
   type='DEF' sous_type='LEN'><determinant type='LE'>le</determinant><nom hyperclasse='DISPOSITIF'
   classe='DISPOSITIF_MAINTIEN' sous_classe=''  domaine1='ALIMENTATION' domaine2=' TRAVAUX_ET_EQUIPEMENTS_MENAGERS'
   domaine3=''>support filtre à café </nom></groupe_nominal> se bouche systématiquement lorsque l'on met <groupe_nominal
   type='NON_DEF' sous_type='DUN'><determinant type='DU'>du</determinant><nom hyperclasse='ALIMENT' classe=''
   sous_classe=''  domaine1='ALIMENTATION' domaine2='' domaine3=''>café moulu</nom> </groupe_nominal>.
11                 </phrase>
12         </message>
```

Figure 40 : Transformation du corpus au format xml

3.2. RESOLUTION

La procédure de décision pour sélectionner l'élément à choisir dans une liste des candidats antécédents pour un GN est faite après l'attribution du poids de saillance à chaque GN.

L'algorithme de la résolution des anaphores associatives est implémenté de la même façon que celui de la résolution des anaphores infidèles. Ainsi, pour chaque article, le premier GN ne sera pas traité comme expression anaphorique car il n'a pas d'antécédent. Les autres GN de l'article doivent passer à la procédure de recherche d'antécédents suivante :

- Nous mettons tous les GN de l'article dans des listes différentes, chaque liste contient les GN de la *phrase n* à la *phrase n+2* (si la proximité décidée est 2) ou la *phrase n+3* (si la proximité décidé est 3).
- Nous attribuons un poids de saillance à chaque élément GN de la liste. Le poids de saillance est décidé arbitrairement en fonction de leur possibilité d'être repéré comme antécédent. Le tableau des poids de saillance proposé contient les facteurs de saillance élémentaire. Le poids de saillance final d'un GN correspond à la moyenne pondéré de tous ses poids élémentaires.
- Nous procédons à la recherche d'antécédent en nous basant sur le calcul du poids de saillance de tous les autres GN de la liste. Chaque élément d'une liste sera analysé, l'un après l'autre, comme expression anaphorique possible. Nous devons comparer le poids de saillance total de chaque élément GN de la liste avec le GN en question. Plus ils partagent de points communs, plus le score (le poids final de saillance) est élevé. Les meilleurs candidats qui ont un score plus élevé que le seuil fixé seront sélectionnés comme antécédents probables du GN en question.

  La recherche d'antécédent est réalisée dans l'ordre suivant :

  Du premier article au dernier article

  De la première liste à la dernière liste de GN de chaque article

Du dernier GN au premier GN de chaque liste

```
 1 trotteuse : montre = 1.4
 2 support filtre à café : machine à expresso = 1.2
 3 café moulu : machine à expresso = 0.7 | support filtre à café = 0.7
 4 partie haute : théière = 0.8 | Bodum = 0.8
 5 veste : veste = 0.9 | celio = 0.3
 6 roues : valises = 0.8
 7 fermeture : roues = 1.0 | valises = 0.6
 8 bords : produit = 0.8
 9 semelles : bottines = 1.0
10 chaussures : bottines = 0.9 | semelles = 0.7
11 semelles : semelles = 1.2 | chaussures = 1.2 | bottines = 1.0
12 pile : montre = 1.2
13 coté : toaster = 1.0
14 bouton : toaster = 1.0 | coté = 0.8
15 couvercle : presse-légumes = 0.8
16 morceaux : presse-légumes = 0.8 | couvercle = 0.8
17 culotte : soutien-gorge = 0.9 | étiquette = 0.3 | CHANTELLE = 0.1
```

Figure 41 : Affichage du résultat au format tableau

PARTIE 3. ANALYSE DES DONNEES

Chapitre 7 : Evaluation

Chapitre 8 : Analyse des erreurs

Chapitre 9 : Réponses aux questionnements

Dans notre travail, nous avons plusieurs choses à évaluer, mais nous mettons l'importance surtout sur l'extraction des expressions référentielles et leur attribution au bon référent.

Avant de donner une évaluation globale sur l'évaluation des expressions référentielles extraites, nous avons décidé d'évaluer notre travail module par module, dans la mesure où nous pouvons voir l'apport de certains paramètres dans l'évolution générale des résultats, tels que : *Apport de chaque méthode d'extraction ; l'apport de l'extraction des noms propres et des syntagmes nominaux et verbaux ; l'apport des indices contextuels : compatibilité du genre, du nombre, la distance; l'apport du parallélisme des fonctions syntaxiques ; l'apport du choix de la taille du corpus, le type de corpus…*

Nous nous fondons sur le calcul de la Précision, du Rappel et du F-mesure pour évaluer la performance de notre système de résolution des anaphores nominales.

CHAPITRE 7. ÉVALUATION

Nous avons testé différentes méthodes de résolution d'anaphores, chaque méthode utilise des paramètres différents. Dans ce chapitre, nous détaillons tout d'abord les résultats obtenus. Nous présentons, par la suite, une comparaison de ces résultats avant d'expliquer la différence entre chaque méthode en montrant l'apport de chaque paramètre utilisé.

1. EXTRACTION DES GROUPES NOMINAUX

1.1. LA RÉSOLUTION DES ANAPHORES DE TYPE INFIDÈLE – SANS EXTRACTION DES SYNTAGMES VERBAUX

1.1.1. L'EXTRACTION DES GN

Nous n'avons pas eu beaucoup de difficultés dans l'extraction des GN. En effet, pour un corpus ayant un thème spécifique comme les faits-divers, nous avons obtenu un résultat souhaité.

Taux de précision	98,4%
Taux de rappel	96,2%
F-mesure	97,3%

Tableau 9 : Résultat de l'extraction des GN (méthode 1)

Ce résultat a montré le succès de notre méthode de construction des dictionnaires de GN.

En effet, les dictionnaires sont construits en récursivité : nous proposons une liste minimum des GN de base de la classe <Personne>, puis en observant les contextes gauches et droits des GN, nous avons construit des patrons syntaxiques qui nous permettent, à leur tour, de retrouver d'autres GN existants dans le corpus.

1.1.2. EXTRACTION DES GN RÉCURSIFS

Nous privilégions la longueur maximale des GN dans l'extraction. Ainsi, la récursivité des GN est une de nos premières règles. Si un GN récursif est identifié, les GN composants ne seront pas pris en compte. Par exemple, si l'on n'affiche pas le mode GN récursif, le repérage du GN dans les expressions *les deux filles du couple*, ou *un voisin du couple* sera :

*Les deux <person style='color: red;' class='generique' g='f' n='p'>**filles**</person> du <person style='color: red;' class='generique' g='m' n='s'>**couple**</person> Al-Hilli avaient survécu, et l'aînée, âgée de 7 ans, avait évoqué la présence d'un seul "méchant" : le tireur.*

*Il s'agit d'un <person style='color: red;' class='generique' g='m' n='s'>**voisin**</person> du <person style='color: red;' class='generique' g='m' n='s'>**couple**</person>*

En nous basant sur la règle de la longueur maximale, nous avons pu améliorer nos résultats. Ainsi, l'extraction des GN nous a permis d'augmenter sensiblement le taux de f-mesure. Si l'on n'applique pas la règle en privilégiant uniquement les GN unitaires, les résultats obtenus n'étaient que :

Taux de précision	89,7%
Taux de rappel	96,2
F-mesure	93,6%

Tableau 10 : Extraction des GN récursif (méthode 1)

On constate que le repérage des GN récursifs nous a permis un apport dans le taux de précision. Le taux de rappel ne change pourtant pas.

Taux de précision	+8,7%
Taux de rappel	0%
F-mesure	+4,5%

1.1.3. EXTRACTION DES GN COMPOSES

Comme nous l'avons expliqué auparavant, nos dictionnaires prennent en entrée des GN composés de base tels que *petite-fille, grand-mère, ex-femme…* La structure *GN+modifiants* n'est pas encore traitée car elle est plus compliquée à mettre en place. Pourtant, nous avons remarqué que comme la structure *GN+adjectif de nationalité* ou des GN composés provenant de deux GN simples existants dans nos dictionnaires *(GN + GN)* existent assez régulièrement dans notre corpus, nous avons décidé ainsi de traiter spécialement ces cas.

GN + GN comme *: jeunes filles, jeune femme*

GN + adjectif de nationalité, comme : *adolescent turc, pédophile belge, serveuse canadienne…*

En effet, les GN composés occupent une bonne place dans notre corpus. Sans le repérage des GN composés, nous avons ce résultat :

Taux de précision	91,1%
Taux de rappel	94,8%
F-mesure	92,9%

Tableau 11 : Extraction des GN composés (méthode1)

Avec le repérage des GN composés, les résultats sont améliorés :

Taux de précision	+7,3%
Taux de rappel	+1,4%
F-mesure	+4,4%

1.2. LA RESOLUTION DES ANAPHORES DU TYPE INFIDELE – AVEC L'EXTRACTION DES SYNTAGMES VERBAUX

1.2.1. L'EXTRACTION DES GN

L'extraction des GN dans ce travail nous a permis d'avoir le résultat suivant :

Taux de précision	92,9%
Taux de rappel	82,5%
F-mesure	87,4%

Tableau 12 : Extraction des GN (méthode 2)

1.2.2. L'EXTRACTION DES NOMS PROPRES

Bien que l'extraction des noms propres nous ait permis d'améliorer légèrement le résultat au niveau de la précision et au niveau du rappel, le traitement des noms propres reste encore une tâche difficile.

Taux de précision	64,7%
Taux de rappel	73,3%
F-mesure	68,8%

Tableau 13 : Extraction des noms propres (méthode 2)

Ainsi, l'identification des noms propres nous donne :

Taux de précision	+1,5%
Taux de rappel	+0,8%
F-mesure	+1,1%

1.3. LA RESOLUTION DES ANAPHORES ASSOCIATIVES

1.3.1. L'ANNOTATION LEXICO-SYNTAXIQUES DES NOMS D'ARTEFACT

L'annotation des classes et des domaines relève de l'annotation des étiquettes lexico-syntaxiques des noms d'artefact dans le corpus. Les noms d'artefact dénotent des objets dont la création est conditionnée par une fonctionnalité précise. Les noms d'artefact peuvent être des noms simples (*marteau*) ou des noms construits du type dérivé (*tondeuse*). Ils peuvent se conçevoir comme des holonymes (*voiture*) ou des méronymes (*roue*).

Nous avons divisé les classes en trois niveaux : les hyperclasses, les classes et les sous-classes. Les hyperclasses servent surtout à définir si l'artefact est un appareil, un organe, ou un accessoire… Les classes et les sous-classes servent à préciser le type de l'artefact. De même, nous avons accordé trois domaines maximum à chaque artefact.

poignée de porte/,.N/H_ORGANE/C_ORGANE_PREHENSION/D1_HABITATION/D2_TRAVAUX_ET_EQUIPEMENTS_MENAGERS/D3_TECHNIQUES

montre/,.N/H_APPAREIL/C_APPAREIL_SIGNALISATION/SC_/D1_TECHNIQUES/D2_VIE_QUOTIDIENNE/D3

L'annotation lexico-syntaxiques des noms d'artefact nous a donné ce résultat :

Taux de précision	96,7%
Taux de rappel	96,7%
F-mesure	96,7%

Tableau 14 : Extraction des noms d'artefact (méthode 3)

La difficulté de l'annotation des étiquettes lexico-syntaxiques des artefacts s'explique dans la mesure où certains substantifs se comportent comme un holonyme ou comme un méronyme selon les contextes.

2. LA RÉSOLUTION

2.1. LA RÉSOLUTION DES ANAPHORES DE TYPE INFIDÈLE – SANS L'EXTRACTION DES SYNTAGMES VERBAUX

2.1.1. AVEC NOS CHOIX DE PARAMÈTRES

Nous avons obtenu des résultats très différents selon la combinaison (ou les jeux d'association) des paramètres qui servent de base à nos calculs du poids de saillance. Si nous mettons tous les paramètres en mode actif, et que les scores élémentaires sont proposés ainsi :

	Indices	Scores proposés
1	Proximité = 0	**0.8**
2	Proximité = 1	**0.6**
3	Antécédent se trouvant dans le titre de l'article	**0.5**
4	Accord en nombre entre <GN> et l'antécédent candidat	**0.8**
5	Accord en genre entre <GN> et l'antécédent candidat	**0.7**
6	Antécédent est en apposition avec <GN>	**0.6**
7	Antécédent et <GN> partagent les mêmes traits sémantiques	**0.9**

Figure 42 : Paramètres pour la résolution des anaphores infidèles – méthode 1

Nous obtenons les résultats suivants :

Taux de précision	92,5%
Taux de rappel	88,7%
F-mesure	90,6%

Tableau 15 : Résultat de la résolution (méthode 1)

En effet, sur 53 anaphores relevées lors d'une annotation manuelle, 49 ont été correctement détectées. Le manque d'identification a été découvert dans 6 cas.

2.1.2. APPORT DE LA COMPATIBILITE DU GENRE, DU NOMBRE

Le taux de précision a augmenté sensiblement grâce à nos critères de compatibilité du genre et du nombre. Ainsi, sans l'activation de cette option, nous avons obtenu seulement :

Taux de précision	52,8%
Taux de rappel	41,5%
F-mesure	46,5%

Ainsi, la compatibilité en genre et en nombre est un critère important dans ce travail, car elle nous a fait gagner un taux important :

Taux de précision	+39,7%
Taux de rappel	+42,7%
F-mesure	+44,1%

Tableau 16 : Apport de la compatibilité genre/nombre

2.1.3. LA DISTANCE

La distance constitue une fenêtre de recherche des candidats antécédents, ou un filtre de sélection des antécédents. Dans la mesure où il s'agit d'un filtre important (la fenêtre devient plus petite), on s'attend à ce qu'il diminue le rappel.

En effet, nous obtenons les meilleurs résultats lorsque la distance est de 2 phrases. Il s'agit de notre résultat avec les paramètres par défaut.

Pourtant, avec la distance de 1 phrase, nous n'avons rien gagné en précision et nous avons beaucoup plus de silences (le rappel diminue de 18,9%) et le f-mesure diminue ainsi de 11,0%

Taux de précision	92,5%
Taux de rappel	69,8%
F-mesure	79,6%

Tableau 17 : Résultat avec la distance = 1 (méthode 1)

Avec la distance de 3 phrases, nous avons plus de bruits (la précision diminue de 21,1%). Le F-mesure ne change pas beaucoup par rapport à la distance de 1 phrase (-10,73% contre 10,99%) même si nous avons gagné de 1,8% en rappel :

Taux de précision	71,4%
Taux de rappel	90,5%
F-mesure	79,8%

Tableau 18 : Résultat avec la distance = 3 (méthode 1)

2.1.4. L'APPORT DU CHOIX DE LA TAILLE ET DU GENRE DU CORPUS

Quelque soit la taille du corpus utilisée, les résultats varient très peu, avec une f-mesure allant de 90,6% à 90,9%. En étendant le corpus à 80% du total, contre 40% prévu, on obtient les résultats suivants :

Taux de précision	91,9%
Taux de rappel	89,9%
F-mesure	90,9%

Tableau 19 : Résultat lorsque la taille du corpus change (méthode 1)

La différence avec les résultats précédents est de :

Taux de précision	+0,6%
Taux de rappel	-1,2%
F-mesure	-0,3%

En effet, les résultats finaux ne diffèrent pas beaucoup car le rappel augmente, tandis que la précision baisse, expliquée par un plus grand taux d'erreurs.

Pourtant, lorsque le genre du corpus change, les résultats changent sensiblement.

Le corpus sur les faits-divers obtient les meilleurs résultats et le corpus sur les avis de consommateurs obtient les moins bons résultats :

Taux de précision	67,1%
Taux de rappel	47,8%
F-mesure	55,8%

Tableau 20 : Résultat lorsque le thème du corpus change (méthode 1)

2.2. RÉSOLUTION DES ANAPHORES DE TYPE INFIDÈLE – AVEC L'EXTRACTION DES SYNTAGMES VERBAUX

2.2.1. AVEC NOS CHOIX DE PARAMÈTRES

Avec nos paramètres choisis par défaut, qui sont :

	Indices	Scores proposés
1	Proximité = 0	**0.5**
2	Proximité = 1	**0.4**
3	Proximité = 2	**0.3**
4	Antécédent se trouvant dans le titre de l'article	**0.8**
5	Accord en nombre entre <GN> et antécédent candidat	**0.3**
6	Antécédent est en apposition avec <GN>	**0.7**
7	Antécédent candidat est sujet	**0.8**
8	Antécédent candidat est attribut	**0.7**
9	Antécédent candidat est COD	**0.6**
10	Antécédent candidat est dans la proposition subordonnée	**0.3**
11	Le déterminant de l'antécédent candidat est possessif	**-0.6**

Figure 43 : Paramètres pour la résolution des anaphores infidèles – méthode 2

Nous obtenons le meilleur résultat pour la résolution des anaphores infidèles avec l'extraction des syntagmes verbaux :

Taux de précision	85,1%
Taux de rappel	68,1%
F-mesure	75,7%

Tableau 21 : Résultat de la résolution (méthode 2)

Par rapport à la méthode utilisée dans la résolution des anaphores infidèles sans l'extraction des syntagmes verbaux, cette méthode donne un résultat plus bas, expliqué par le manque de la compatibilité du genre et du nombre entre le GN analysé et son antécédent candidat. La compatibilité est un paramètre important, mais nous ne pouvons pas l'appliquer dans ce travail car les objectifs des deux méthodes sont différents.

2.2.2. APPORT DE L'IDENTIFICATION DES FONCTIONS SYNTAXIQUES

L'identification des groupes verbaux nous a permis de repérer les fonctions syntaxiques de certains GN. Et nous avons utilisé cette étiquette dans la résolution des anaphores.

C'est en accordant une priorité aux antécédents selon qu'ils sont sujets ou compléments d'objet que nous avons obtenu le meilleur résultat. Sans ces étiquettes, les résultats sont légèrement modifiés :

Taux de précision	80,1%
Taux de rappel	63,8%
F-mesure	71,0%

Tableau 22 : Sans l'identification des fonctions syntaxiques (méthode 2)

Nous avons également choisi d'accorder plus de poids de saillance aux antécédents ayant la même fonction syntaxique que le SN à résoudre. Cependant comme le repérage des verbes et des fonctions syntaxiques des GN est assez limité, les résultats ne s'en ressentent pas beaucoup.

2.2.3. APPORT DES PATRONS SYNTAXIQUES

Grâce à l'identification des structures spécifiques, nous avons identifié en partie des candidats antécédents au GN.

Sans les patrons, le calcul des poids de saillance nous donne les résultats assez faibles :

Taux de précision	63,8%
Taux de rappel	59,6%
F-mesure	61,6%

Tableau 23 : Résultat sans implémentation des patrons syntaxiques (méthode 2)

Les paramètres choisis par défaut nous ont ainsi permis de gagner 14,02% en f-mesure

Pour certains types de corpus, les GN se trouvant dans le titre sont privilégiés, et ainsi obtiennent plus de poids de saillance. Si l'on désactive cette option, le résultat se modifie légèrement en taux de rappel (+2,1%), et en f-mesure (+1,31%) :

Taux de précision	85,1%
Taux de rappel	66,0%
F-mesure	74,3%

2.3. RESOLUTION DES ANAPHORES ASSOCIATIVES

2.3.1. AVEC NOS CHOIX DE PARAMETRES

Nous avons choisi 0.5 pour le seuil de sélection pour les anaphores associatives, et nos scores élémentaires sont choisis selon le tableau suivant :

	Indices	Scores proposés
1	Proximité = 0	0.8
2	Proximité = 1	0.7
3	Proximité = 2	0.5
4	<GN> et Antécédent candidat ont la même Hyperclasse	0.7
5	<GN> et Antécédent candidat ont la même Classe	0.8
6	<GN> et Antécédent candidat ont la même Sous-classe	0.9
7	<GN> et Antécédent candidat ont le même Domaine 1	0.7
8	<GN> et Antécédent candidat ont le même Domaine 2	0.6
9	<GN> et Antécédent candidat ont le même Domaine 3	0.5

Figure 44 : Paramètres pour la résolution des anaphores associatives

Nous obtenons le résultat :

Taux de précision	64,7%
Taux de rappel	73,3%
F-mesure	68,7%

Tableau 24 : Résultat de la résolution des anaphores associatives

Par rapport aux autres types d'anaphore, la résolution des anaphores associatives obtient les résultats moins élevés, ce qui explique la complexité de la tâche.

2.3.2. APPORT DU FILTRE PAR DETERMINANT

Nous avons proposé un filtre de faux antécédents candidats : nous imposons la déduction du poids de saillance si l'antécédent candidat possède un déterminant possessif. Ce filtre devrait nous permettre d'enlever certains cas d'erreurs. Pourtant, le résultat que nous obtenons ne change pas :

Taux de précision	64,7%
Taux de rappel	73,3%
F-mesure	68,8%

Tableau 25 : Résultat sans l'identification des déterminants

Ce résultat peut être expliqué par la taille du corpus. Pour un corpus plus grand, nous pourrions trouver des cas où la structure *[son GN] […] [le GN]* existe. Et si les autres paramètres permettent d'établir une relation anaphorique entre ces deux GN, le filtre va baisser le poids de saillance et empêcher que l'antécédent candidat soit retenu.

3. APPORT DES AUTRES PARAMÈTRES

3.1. EXTRACTION DES SYNTAGMES VERBAUX

Les syntagmes verbaux sont étiquetés dans la résolution des anaphores infidèles (deuxième méthode). L'étiquetage des syntagmes verbaux nous a permis d'obtenir une bonne précision, mais le silence reste encore très important, expliqué par un taux de rappel très faible :

Taux de précision	76,8%
Taux de rappel	48,7%
F-mesure	59,6%

Tableau 26 : Résolution des anaphores infidèles sans l'extraction des syntagmes verbaux (méthode 2)

3.2. EXTRACTION DES DÉTERMINANTS

Les déterminants annotés ne sont pas uniquement de deux types : déterminant défini (*le N*) et déterminant indéfini (*un N*), nous traitons également les déterminants

spéciaux comme *ledit* + *N; le même* + *N* ; *un autre* + *N* ; *le second* + *N* ; *le premier* + *N* ; *de nombreux* + *N* ; *un/deux/trois…/dix* + *N*

Dans le premier corpus, nous avons obtenu une très bonne précision dans l'extraction des déterminants. Le taux de rappel que nous obtenons n'est pourtant pas très élevé, car nous avons rencontré quelques difficultés dans :

- L'annotation des déterminants lorsqu'ils sont des chiffres (par exemple : *Les 58 passagers*)
- Des formes qui nous empêchent d'afficher certaines informations concernant le genre ou le nombre du GN comme : *l', des, d'* + N

Le résultat que nous obtenons est :

Taux de précision	92,7%
Taux de rappel	84,8%
F-mesure	88,6%

Tableau 27 : Extraction des déterminants (méthode 1)

Ainsi, nous avons le récapitulatif des scores suivant :

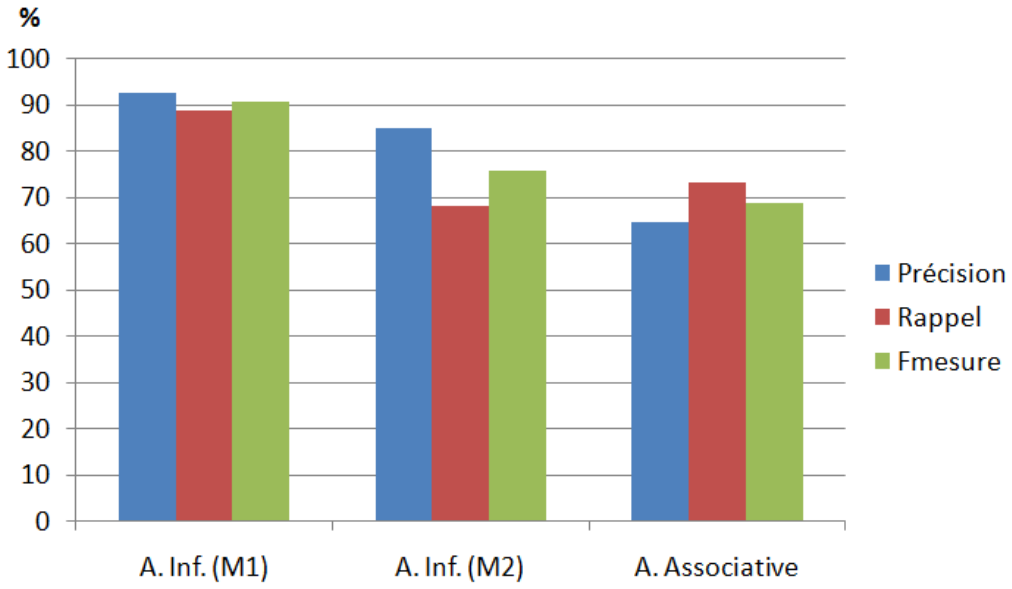


Figure 45 : Récapitulatif des scores

CHAPITRE 8. ANALYSE DES ERREURS

Si la comparaison des résultats obtenus avec l'annotation manuelle donne l'évaluation quantitative, l'analyse des erreurs et du silence relève de l'évaluation qualitative – sujet de ce chapitre.

Nous abordons d'abord des erreurs au niveau du prétraitement, comme l'extraction des syntagmes nominaux, des syntagmes verbaux. Nous allons détailler aussi quelques points qui montrent les limites de notre programme puis nous parlons également de nos contraintes pendant la résolution.

1. LES ERREURS

Comme nous l'avons déjà dit, les ressources lexicales jouent un rôle très important dans la résolution des anaphores. Pourtant, l'exploitation des ressources lexicales n'est pas une tâche facile dans l'extraction des syntagmes nominaux.

Nous avons observé plusieurs types d'erreurs, qu'on peut classer selon des caractéristiques communes.

1.1. LES NOMS PROPRES

La détection des noms propres pose certains problèmes dans notre travail : même avec un dictionnaire des noms propres bien couvrants statistiquement, il n'est pas possible de réaliser un dictionnaire recensant à chaque instant tous les noms propres dans nos corpus. Il faut donc trouver et coder des règles pour permettre au système de détection d'identifier les expressions correspondant à des noms propres et de trouver à quoi de telles expressions font référence sans avoir forcément besoin d'un dictionnaire. Les critères sont très difficiles à trouver, le critère de la majuscule par exemple n'est pas toujours valide, tous les noms propres n'ont pas forcément de majuscule au début, et un logiciel de détection doit être capable de détecter un nom propre même si sa majuscule a été oubliée. Et même dans un dictionnaire, des ambiguïtés persistent : *France* = prénom ou pays ou nom de société ?

On peut faire appel à des règles du type : « après la préposition *en*, le mot *France* correspond au pays ». Il faut établir des grammaires locales pour désambiguïser les expressions à partir du contexte.

Néanmoins, si on dispose de dictionnaires d'analyse morphologique pour les mots, on ne dispose pas d'un tel outil pour les noms propres. Par exemple, si on cherche (Léo) Ferré, on trouvera le participe passé du verbe ferrer : le dictionnaire morphologique analysera le nom propre comme une forme d'un mot.

Par exemple, notre système ne peut pas distinguer les noms communs et les noms propres dans le cas suivant :

Liam <person style='color: red;' class='famille' g='m' n='p'>Cousins</person> a dû répondre de ses actes devant le tribunal.

Le mot *Cousins* est un nom propre, mais il a été repéré comme un GN commun. Normalement, notre système commence par l'extraction des noms propres. *Cousins* dans ce contexte est un nom de famille, il n'appartient pas à la famille des prénoms courants et il n'a donc pas été extrait. Pourtant, lors de la deuxième extraction concernant la classe <Personne>, le mot *cousins* a été considéré comme appartenant de la classe <famille> au pluriel.

1.2. LES NOMS COMPOSES

Une autre erreur assez courante concerne les noms composés. En effet, l'apparition des ADJ dans un GN pose parfois problèmes, comme dans cet exemple :

[…]une très <person style='color: red;' class='famille' g='f' n='s'>belle fille</person>

Le GN *belle fille* a été repéré comme un mot de la classe <famille>. Pourtant, la présence du déterminant *une* avec l'adverbe *très* montre qu'il y a une erreur dans le repérage du GN.

1.3. ERREURS PROVENANT DU CORPUS D'ORIGINE

Certaines erreurs d'annotation ne proviennent pas de notre système d'extraction, mais proviennent de l'incohérence dans le texte d'origine, par exemple :

Des adolescentes de 13 et 14 ans soupçonnées d'avoir kidnappé une enfant de 3 ans dans un magasin Primark. {S}Un enfant de trois ans avait disparu ce mercredi après-midi d'un

magasin Primark à Newcastle, en Angleterre.{S} Ses ravisseurs seraient deux jeunes filles de 13 ans et 14 ans.{S} La fillette de trois ans a finalement été retrouvée saine et sauve par la police à 18 kilomètres de Newcastle.

Dans le premier paragraphe, le mot *enfant* est identifié comme féminin (***une*** *<person style='color: red;' class='famille' g='f' n='s'>enfant</person> de 3 ans)* mais dans le deuxième paragraphe, il est identifié comme masculin (*Un <person style='color: red;' class='famille' g='m' n='s'>enfant</person> de trois ans)*

Cette erreur, provenant du corpus d'origine, nous a empêchée de mener une bonne recherche d'antécédent.

En outre, nous avons aussi des erreurs causées par des signes de ponctuation, comme dans cet exemple :

{S}Le <person style='color: red;' class='metier' g='m' n='s'>conducteur</person>, <person style='color: red;' class='origine' g='m' n='s'>français</person> d'origine, venait d'Avignon et roulait avec une Audi Quatro

Dans cette phrase, le mot « français » est un adjectif et non pas un GN. Il a été quand-même extrait par Unitex à cause de la présence d'une virgule qui se place juste devant. En effet, les marquages du GN que nous considérons dans notre travail est : la présence des déterminants ; des ponctuations ; ou le marquage du début de phrase.

1.4. LA CLASSIFICATION DES TYPES D'ANAPHORES

Pour la résolution des anaphores associatives, nous avons enregistré 6 cas d'erreurs qui appartiennent à deux types d'erreur suivants :

a, Étiquetage des anaphores fidèles

Je vous ai déjà écrit pour vous dire que ma veste Celio présentait des taches blanches. Au vu du prix de la veste je souhaitais la conserver et non la renvoyer.

veste : veste = 0.9 | celio = 0.3

Je viens de recevoir les bottines que j'avais commandées. Les semelles ne peuvent pas coller dans les chaussures et il y a beaucoup de colle sur les semelles

semelles : semelles = 1.2 | chaussures = 1.2 | bottines = 1.0

Au point de vue de la résolution des anaphores générales, ce sont des anaphores du type fidèle, qui nécessitent d'être distinguées avec les anaphores du type infidèle – le sujet de notre thèse.

Ce n'est donc pas une erreur de la résolution mais une erreur au niveau de la classification des types d'anaphores.

Dans ces exemples, notre programme a choisi le candidat *veste* pour la reprise *veste* et le candidat *semelles* pour la reprise *semelles* car ayant le même forme lexicale, ces paires de GN partagent au maximum des ressemblances au niveau de leurs attributs (hyperclasse, classe, domaines etc.), ce à quoi s'ajoute une distance assez proche. Ainsi, le poids de saillance est très élevé.

b, Etiquetage des anaphores infidèles

Un autre type d'erreur implique l'étiquetage des GN synonymiques et pas seulement les types partie-tout. Par exemple, les paires de GN suivants sont quand-même associées : *chaussures* et *bottines ; fermeture* et *roues ; culotte* et *soutien-gorge*.

Je viens de recevoir les bottines que j'avais commandées. Les semelles ne peuvent pas coller dans les chaussures et il y a beaucoup de colle sur les semelles

chaussures : bottines = 0.9 | semelles = 0.7

J'ai acheté plusieurs CHANTELLE. La taille du soutien-gorge n'est pas du 90B même si c'est ce qui est marqué sur l'étiquette. La culotte est déchirée.

culotte : soutien-gorge = 0.9 | étiquette = 0.3 | CHANTELLE = 0.1

J'ai commandé trois valises Platinium. Après seulement deux utilisations les roues sont cassées et la fermeture aussi.

fermeture : roues = 1.0 | valises = 0.6

Avec ces exemples, nous pouvons trouver les limites de notre système, c'est le manque de la classification des types d'anaphores. Après la résolution, nous pourrions ajouter un module de classification des types d'anaphore pour préciser de quel type d'anaphore s'agit-il : associative méronymique, associative synonymique, ou simplement anaphores infidèles.

2. NOS DIFFICULTES ET LES CONTRAINTES DE NOTRE SYSTEME

2.1. LE CHOIX DES CLASSES

Nous avons aussi des difficultés dans le choix des classes. En effet, certains mots peuvent appartenir à plusieurs classes, et seul le contexte nous permet de les distinguer. Par exemple le mot *fille* appartient à la classe <famille > dans :

Un homme a été placé en garde à vue pour homicide volontaire de sa fille.

Il a deux filles, une fille de 15 ans habite chez lui et une autre, étudiante, à Lyon.

Mais il appartient également à la classe <générique> dans :

Le corps d'une jeune fille a été découvert à Vincennes

2.2. TRAITEMENT DES NOMBRES EN LETTRES

Nous avons rencontré certaines difficultés liées à la convention d'annotation des nombres en lettres. Pour la plupart des cas, les nombres de 1 à 10 sont écrits en

lettres. Au-delà de 10, les nombres sont souvent écrits en chiffres. Cette remarque nous a permis une convention de traitement : les déterminants en nombre (de deux à dix) ou en chiffre sont marqués au pluriel.

Pourtant, il existe des cas d'exception comme :

[…] ont repêché douze corps dans la Seine l'an dernier.

Le mot *corps* n'a pas été annoté au pluriel car le mot *douze* n'a pas été traité comme un déterminant au pluriel.

2.3. ANNOTATION DU GENRE ET DU NOMBRE

Après l'application des dictionnaires d'Unitex, chaque entrée dans le texte est associée à une entrée unique dans le dictionnaire. Nous avons appliqué la règle du genre masculin privilégié pour les mots dont on ignore le genre (ils peuvent être à la fois masculins ou féminins), et la règle du singulier privilégié pour les mots dont on ignore le nombre (ils peuvent être à la fois singuliers ou pluriels). La vérification et la correction du genre et du nombre des GN annotés en contexte est donc nécessaire. Pourtant, il existe certains cas plus difficiles à traiter car le déterminant du GN n'est pas suffisant pour définir le genre et le nombre des GN.

Une première difficulté relevée est comment déterminer le bon genre du GN. Dans cette phrase, le déterminant « *une* » permet de déterminer le genre du GN *enfant* :

Des <person style='color: red;' class='generique' g='f' n='p'>adolescentes</person> de 13 et 14 ans soupçonnées d'avoir kidnappé une <person style='color: red;' class='famille' g='m' n='s'>enfant</person> de 3 ans dans un magasin Primark.

Après la correction, nous avons :

Des <person style='color: red;' class='generique' g='f' n='p'>adolescentes</person> de 13 et 14 ans soupçonnées d'avoir kidnappé une <person style='color: red;' class='famille' g='f' n='s'>enfant</person> de 3 ans dans un magasin Primark.

Pourtant, l'analyse du déterminant ne nous permet pas suffisamment de connaître le genre du mot *enfant* dans d'autres cas.

Pour pouvoir bien annoter les informations morphosyntaxiques d'un GN, nous devons parfois passer par une analyse grammaticale complète dans le contexte général, comme dans les exemples suivants :

*Deux heures après avoir été prise en charge, l' <person det='def' **g='m'** nb='s' style='color: red;' class='famille'>**enfant**</person>, originaire du Texas a poussé son dernier souffle.{S} Daisy, une <person det='ind' **g='f'** nb='s' style='color: red;' class='generique'>**fillette**</person> de 14 mois, décède à la suite d'une anesthésie administrée par son <person det='poss' g='m' nb='s' style='color: red;' class='metier'>dentiste</person>*

Si nous ne connaissons pas les informations morphosyntaxiques du mot *enfant* dans la première phrase, les informations morphosyntaxiques du mot *fillette* dans la deuxième phrase nous permettent de trouver son genre.

Mais les informations morphosyntaxiques ne sont pas toujours existantes, au cas où le mot *fillette* est imbriqué dans un autre mot (la règle de la longueur maximale du GN impose le repérage du GN *mère de la fillette*), où ses informations morphosyntaxiques ne sont pas annotées, comme le cas suivant :

*{S}Le <person det='def' g='m' nb='s' style='color: red;' class='generique'>suspect</person>, prénommé Wang et âgé de 33 ans, aurait maîtrisé la <person det='def' g='f' nb='s' style='color: red;' class='famille'>mère</person> et décapité l' <person det='def' g='m' nb='s' style='color: red;' class='famille'>**enfant**</person> avec un couteau de cuisine. «la <person det='def' g='f' nb='s' style='color: red;'*

*class='famille'>mère de la **fillette** </person> s'est ruée sur Wang pour le repousser, mais il était plus fort et l'a poussée sur le côté», précise la police.*

La mise en place d'une telle analyse est encore plus difficile au cas où on doit se baser sur les autres mots et non pas uniquement les GN repérés. Par exemple :

*{S}Les versions de Cécile Bourgeon et Berkane Maklouf divergent sur les causes du décès, mais pas sur l'enterrement.{S} Découvrant l' <person det='def' g='m' nb='s' style='color: red;' class='famille'>**enfant**</person> **morte** le dimanche matin, ils l'ont **mise**, **nue**, dans un sac et **placée** dans le coffre de la voiture.*

*L' <person det='def' **g='m'** nb='s' style='color: red;' class='famille'>**enfant**</person> serait **née** le 9 août 2012*

*l' <person det='def' **g='m'** nb='s' style='color: red;' class='famille'>**enfant**</person> avait été **retrouvée** saine et sauve et que l'alerte enlèvement avait été levée, sans plus de précision.*

{S}Après avoir été relâchée par ses ravisseurs, l'<person style='color: red;' class='famille' g='m' n='s'>enfant</person> était parvenue à s'enfuir et avait trouvé de l'aide auprès d'un automobiliste à proximité du lieu de l'agression.

*L' <person det='def' g='m' nb='s' style='color: red;' class='famille'>enfant</person> **âgée** de 9 mois était **enfermée** dans une voiture*

Dans ces exemples, les signes qui montrent le genre du mot *enfant* peuvent être repérés grâce aux ADJ *mise, nue, placée* au participe passé *née, parvenue, enfermée* et à l'expression figée *retrouvée saine et sauve*

Nous avons aussi des cas plus difficiles, où l'on ne peut pas définir le genre du GN comme dans cet exemple :

> *C'est des enfants sauvages, qui ont poussé tout seul à la Courneuve.*

Aucun autre signe contextuel ne nous permet de préciser le genre du GN *enfants* dans cet exemple. Le mot *enfants* est donc annoté par défaut comme un groupe nominal masculin, appartient à la classe <générique>, au pluriel.

Nous n'avons pas d'information concernant le genre pour les déterminants suivants :

Des, autres, quelques, ces, les, l', notre, votre, leur, mes, tes, ses, nos, vos, leurs, deux, trois, quatre, cinq, six, sept, huit, neuf, dix, plusieurs, aux.

Nous n'avons pas d'information sur le genre, ni sur le nombre des GN pour les déterminants suivants : *de, d'*.

Les erreurs de l'annotation du nombre proviennent souvent des mots ayant « s » en terminaison, comme « français » ou « fils ».

2.4. TRAITEMENT DU LANGAGE WEB

Une autre difficulté provient du traitement du langage web, lorsque les auteurs des textes utilisent le langage SMS, des images créées avec les caractères spéciaux font usage sans contrainte des espaces ou des sauts à la ligne pour attirer l'attention des lecteurs, ou lorsqu'ils font des erreurs orthographiques ou grammaticales.

Par exemple, les fautes d'orthographe sont assez régulières dans les commentaires ou les forums en ligne :

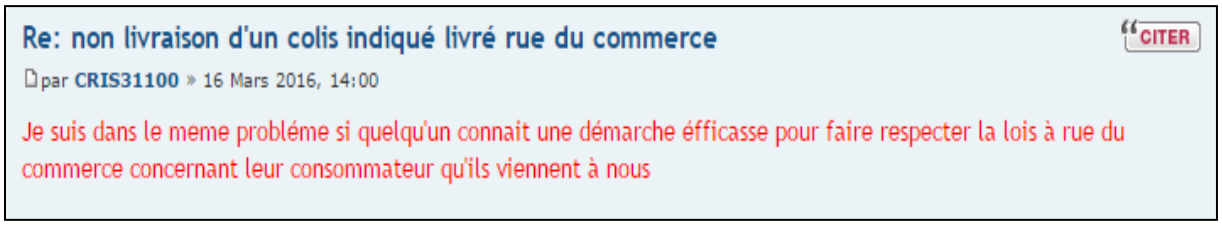


Figure 46 : Fautes d'orthographe typiques dans les commentaires

Ce phénomène est assez répandu dans les commentaires, les blogs ou les forums. Pour utiliser un corpus aspiré sur le web, l'identification des blocs de textes pertinents est indispensable.

Ainsi, il serait nécessaire d'insérer un module permettant de filtrer de façon automatique et séparer la partie texte considérée comme pertinente des textes considérés comme « bruits ». Pourtant, il s'agit d'une autre problématique dans le TALN, et la plupart des aspirateurs de site web actuels intègrent le module de nettoyage, mais ils ne sont pas aussi performants face à la variété des structures des pages web et la complexité de la tâche.

CHAPITRE 9. REPONSES AU QUESTIONNEMENT

Nous souhaitons réserver ce chapitre pour présenter nos réflexions sur les méthodes que nous avons utilisées, visant à apporter les réponses au questionnement que nous avons posé au début de cette thèse : 1) Comment exploiter l'étiquetage anaphorique pour l'interprétation des textes ? 2) Quel est l'apport des ressources linguistiques pour le traitement automatique ? 3) Quelle analyse rétroactive sur la théorie permet la mise en place de l'outil informatique ?

Pour la dernière partie de ce chapitre, nous souhaitons présenter nos réflexions sur nos méthodes de travail.

1. REPONSES AUX QUESTIONS

S'inscrivant dans la perspective du TALN, notre étude a ainsi insisté sur la description du fonctionnement syntaxique, sémantique et référentiel des relations anaphoriques en revenant à la littérature qui porte sur trois questions.

1.1. PREMIERE QUESTION

La première question que nous avons posée était : *Comment exploiter l'étiquetage anaphorique pour l'interprétation des textes ?* A notre connaissance, l'annotation des anaphores nominales reste toujours une tâche difficile à l'heure actuelle, car il n'existe aucun outil d'annotation adapté pour la résolution de ce type d'anaphore.

La plupart des travaux existants sur la résolution des relations anaphoriques est réalisée dans des limites : soit le nombre d'expressions anaphoriques nominales est limité, soit le corpus d'évaluation est suffisamment court pour une évaluation manuelle. Pour le moment, personne n'a proposé un outil complet équivalent pour l'annotation anaphorique nominale pour le français, et ce genre de problème est encore largement non résolu.

Face à la difficulté de la tâche, nous avons choisi d'utiliser la méthode manuelle pour l'évaluation. Pour faciliter la tâche, après le traitement, nous avons affiché les résultats en deux formats, un tableau qui favorise l'évaluation statistique et une sortie au format de texte, avec l'affichage en couleur des GN identifiés et leurs scores équivalents.

L'évaluation manuelle est relativement simple, mais elle a beaucoup de limites :

- Le résultat peut changer d'un échantillon à un autre.
- A chaque fois que nous changeons un paramètre, le résultat change, et nous devons compter à nouveau le nombre de relations anaphoriques établies. Il s'agit d'une étape coûteuse en temps.

1.2. DEUXIEME QUESTION

La deuxième question que nous avons posée était : *Quel est l'apport des ressources linguistiques pour le traitement automatique ?* Partant de l'idée que la formation des anaphores ne peut être séparée de l'interprétation des schémas syntaxiques, de la morphologie et de la sémantique de ce type de phénomène, nous nous sommes fixés comme objectif une résolution automatique des anaphores nominales en nous basant sur les études lexicales, sémantiques et syntaxiques de ce phénomène. En effet, pour la résolution des anaphores nominales, nous avons besoin non seulement des connaissances sémantiques (Exp : On parle depuis des années d'un nouveau **bâtiment** dans ce quartier, mais la construction de l'**immeuble** vient d'être commencée), mais aussi des connaissances morphosyntaxiques (Par exemple : A chaque fois que j'utilise ma nouvelle **cafetière**, je sens une odeur bizarre dans le **café)** ou même des connaissances encyclopédiques (Je n'aime pas **les chiens**, à part les **labradors**).

L'idée des dictionnaires électroniques nous a permis d'atteindre notre objectif. En effet, certaines notions extralinguistiques sont notamment utiles dans la désambigüisation des unités lexicales. Il faut compter les indications d'hyperclasses, de classes, de sous-classes et de domaines. Le rattachement variable d'un mot à un domaine en fonction du degré de précision que l'on attribue à ce type d'information explique la distinction entre les champs de domaines (Buvet and Mathieu-Colas, 1999). Par exemple, le mot « trotteuse » peut être considéré par les lexicographes comme relevant du domaine de l'horlogerie, ou plus précisément, une partie de la montre, ou d'un autre domaine.

La notion de classe sémantique sur laquelle s'appuient les dictionnaires électroniques dans la description des unités lexicales est primordiale pour réduire la polysémie, mais les notions de domaines demeurent indispensables. Par exemple, les mots

« trotteuse », « aiguille » ont en commun d'appartenir à la classe d'objets (matériel d'une montre) et diffèrent seulement par leur appartenance respective aux domaines.

1.3. TROISIEME QUESTION

Quelle analyse rétroactive de la théorie permet la mise en place de l'outil informatique ?

Pour répondre à cette question, nous revenons sur les propriétés de la langue qui permettent la résolution automatique des anaphores nominales.

Nous nous trouvons devant les différentes informations qui concernent la nature des éléments qui peuvent entretenir une relation anaphorique. Cela nous a mis devant la nécessité de rappeler les différentes théories qui permettent la réalisation de nos études, comme les contraintes syntaxiques (l'accord en genre et en nombre, l'accord en fonction syntaxique du GN), contraintes sémantiques (la classe, la sous-classe et les domaines des GN), la distribution des GN (position d'emphase du GN, la proximité des GN dans la relation référentiel avec un autre GN), l'extraction des patrons syntaxiques (la présence des verbes indicateur, les structures), l'extraction des déterminants (y compris les déterminants spéciaux comme *le dit* + *N; le même* + *N ; un autre* + *N ; le second* + *N ; le premier* + *N ; les lettres)*, l'extraction des GN (y compris les GN spéciaux comme les GN composés, les sigles, les noms propres, les récursifs,…) .

Les résultats obtenus nous ont montré l'intérêt de théories linguistiques vis-à-vis les différents points en rapport avec le phénomène des anaphores nominales : le substantif de l'expression anaphorique, la détermination de l'expression anaphorique et la distribution de l'expression anaphorique.

2. RÉFLEXIONS SUR NOS MÉTHODES

Nos deux premières méthodes relèvent des méthodes pauvres en connaissances. La première méthode se fonde uniquement sur le calcul des poids de saillance et la deuxième méthode se fonde aussi sur les patrons syntaxiques. Même si ces méthodes dépendent moins des ressources lexicales, elles nous permettent d'obtenir de bons résultats. Notre troisième méthode dépend majoritairement des ressources lexicales, les critères de sélection des antécédents se basent sur les dictionnaires.

Après avoir observé et analysé les résultats obtenus, nous avons fait certaines remarques concernant nos difficultés rencontrées.

La première difficulté provient du manque d'informations concernant le contexte linguistique dans nos corpus. Le contexte est nécessaire à la compréhension et l'interprétation des corpus, mais les termes dans les ressources lexicales sont hors contexte, et la classification des entrées dans nos dictionnaires reste encore très arbitraire. Un <artefact> peut être un appareil dans un contexte, mais il devient une partie de l'ensemble dans un autre contexte.

La deuxième difficulté concerne la désambigüisation au niveau des unités lexicales et au niveau des énoncés. Les langues naturelles ont des caractéristiques fondamentales, notamment concernant le rapport signifiant-dénotation. En effet, on peut avoir un signifiant et plusieurs dénotations, il s'agit alors d'ambiguïté inhérente à l'énoncé. Les connaissances encyclopédiques ou linguistiques permettent de lever certaines ambiguïtés qu'une machine ne peut gérer puisqu'elle ne possède pas de telles connaissances encyclopédiques ou grammaticales (syntaxiques, lexicales …), connaissances qu'on ne sait pas encore coder.

La langue naturelle laisse une liberté d'expression très grande, trop difficile à coder et ingérable pour un ordinateur qui ne peut pas percevoir par exemple les liens de synonymie, d'hyperonymie… L'interprétation d'un texte n'est pas une opération simple, elle met en œuvre chez les individus une multitude de mécanismes plus ou

moins inconscients. En revanche, le langage informatique est quelque chose de très rigide, qui ne souffre aucune erreur orthographique ou syntaxique du fait de l'absence de capacités de compréhension et d'interprétation. La constitution et l'utilisation d'un langage informatique demandent une très grande rigueur, une parfaite exhaustivité et une formalisation exemplaire des règles régissant le langage. Néanmoins, il est impossible d'acquérir des dictionnaires exhaustifs.

Pour atteindre un meilleur résultat, il faut savoir suffisamment comment la langue fonctionne, et traduire les mécanismes syntaxiques, morphologiques, sémantiques et phonétiques en langage formalisé. Pour cela, nous avons quelques propositions :

- Il serait intéressant de pouvoir réunir dans un même corpus les différentes analyses que nous avons faites. Toutefois, nous nous sommes aperçu que : plus le traitement du corpus est spécifique, plus le résultat est satisfaisant.
- Il serait aussi intéressant si nous améliorons les ressources linguistiques existantes en intégrant un système d'acquisition automatique des ressources lexicales à notre étude avant de mettre en œuvre le système de résolution.
- D'un autre côté, l'augmentation de la quantité des patrons syntaxiques aiderait aussi à l'amélioration des résultats. Pour l'instant, nous utilisons seulement les patrons dans la résolution des anaphores nominales du type infidèle avec l'identification des syntagmes verbaux. Comme les patrons syntaxiques nous ont permis d'augmenter la précision et le rappel de notre programme, une augmentation de la quantité des patrons nous permettra d'augmenter le f-mesure, ou d'améliorer nos résultats.
- Une autre proposition qu'on peut réaliser est l'intégration d'un module de reconnaissance des groupes nominaux pluriels et d'un module d'évaluation plus efficace dans le système
- En outre, comme l'outil UNITEX nous permet de travailler avec différentes langues, appliquer ces méthodes pour le traitement d'autres langues serait réalisable.

CONCLUSION

L'objet de notre thèse est l'étude des formes de reprises anaphoriques de type nominal en mettant en avant les techniques de résolution automatique des anaphores afin d'obtenir une compréhension automatique des textes. Dans cet objectif, notre travail de recherche s'est attaché plus particulièrement au fait d'établir, dans la perspective du TAL, les propriétés des anaphores nominales qui doivent être prises en compte afin d'une part d'identifier automatiquement les anaphores et leurs sources et d'autre part d'établir comment les anaphores nominales concourent à la cohésion des textes.

Pour y arriver, nous avons tenté d'analyser, dans la première partie, les aspects linguistiques de l'anaphore nominale en termes de la typologie, la syntaxe et la sémantique. En effet, la résolution automatique des anaphores n'a pas qu'un aspect pratique, son intérêt théorique est plus important qu'on peut le penser ; puisqu'on veut faire faire à des machines un certain nombre d'opérations sur la langue, il faut passer par l'explicitation de toutes les connaissances nécessaires à ces tâches et par leur formalisation. La problématique du traitement automatique nous pousse à avoir un point de vue plus formel sur les connaissances linguistiques qu'une théorie qui se traduirait plus par un discours que par un ensemble de points formalisés.

Ce sont ces analyses en amont de tout traitement informatique qui nous ont offert une vue globale dans la mise en œuvre d'un algorithme pour notre système. En effet, nous avons pris en considération des caractéristiques linguistiques et certaines théories pour réaliser notre travail, telles que : la théorie des classes d'objet, la théorie des trois fonctions primaires.

La partie principale de notre thèse a été consacrée à la description de notre travail, avec l'analyse du sujet de l'anaphore nominale sous l'aspect du traitement automatique : le choix du corpus, la modélisation efficace des anaphores, le principe

de décision grâce au calcul du poids sémantique alloué à chaque antécédent candidat, l'évaluation du système de résolution. Pour la résolution des anaphores infidèles, nous avons utilisé la méthode pauvre en connaissance, c'est à dire utiliser uniquement des connaissances du type morphosyntaxique des GN. Pour la résolution des anaphores associative, nous avons utilisé la méthode riche en connaissance, qui mobilise également les connaissances sémantiques des GN.

Nous avons essayé différentes études pour tester les méthodes utilisées et la fonctionnalité des paramètres utilisés, et les résultats obtenus sont encourageants : nous obtenons les meilleurs résultats pour la résolution des anaphores infidèles avec la première méthode, sans l'analyse des groupes verbaux, à 90,6% de taux de F-mesure. La deuxième méthode, avec l'intégration de l'extraction des groupes verbaux, nous a donné 75,7% de réussite et nous avons obtenu seulement 68,7% de réussite pour la résolution des anaphores associatives.

Pour les questions de recherches que nous avons posées, ce travail a apporté les réponses sur l'importance du phénomène de l'anaphore dans l'interprétation des textes, sur l'importance des ressources linguistiques des théories linguistiques dans ce sujet.

Ainsi, sur le plan théorique, la réalisation de cette étude a permis de mettre en lumière les techniques de résolution des anaphores nominales avec la notion de poids sémantique, qui contribue à la mise en place de notre programme.

Sur le plan applicatif, notre thèse peut donner lieu à une meilleure compréhension des textes.

BIBLIOGRAPHIE

Abeillé, A., 2001. Un corpus français arboré : quelques interrogations. TALN.

Aone, C., Bennett, S.W., 1995. Evaluating automated and manual acquisition of anaphora resolution strategies. St. Cruz N. M., Proceedings of the 33rd annual meeting of the ACL 122–129.

Aone, C., Ramos-Santacruz, M., 2000. RESS: A large-scale relation and event extraction system. Seattle USA, Proceedings of ANLP.

Apothéloz, D., 2002. La construction du lexique français. Ophrys.

Apothéloz, D., 1995a. Rôle et fonctionnement de l'anaphore dans la dynamique textuelle. Librairie Droz, Genève-Paris.

Apothéloz, D., 1995b. Référents clandestins et anaphores atypiques. Trav. Neuchâtel. Linguist. Inst. Sci. Lang. Commun. Neuchâtel Suisse 143–173.

Baldwin, B., 1997. CogNIAC : high precision coreference with limited knowledge and linguistic resources. Madrid, Proceedings of ACL/EACL workshop on Operational factors in pratical, robust anaphora resolution.

Beust, P., Nicolle, A., 1997. L'identité dans un modèle de la dépendance nominale.

Bonhomme, M., 1998. Les figures clés du discours., Seuil. ed.

Boudreau, S., Kittredge, R., 2006. Résolution d'anaphores et identification des chaînes de coréférence : une approche « minimaliste. Univ. Montr.

Buvet, P.-A., 2015. Fonction argumentale et possessivation. Tunisie, Synergies 147–163.

Buvet, P.-A., 2013. La dimension lexicale de la détermination en français, Honoré Champion. ed. Paris.

Buvet, P.-A., 2011. Categorisation semantico-enonciative du lexique a partir d'un dictionnaire electronique. Fontes Métotodos E Novas Technol., Os di.ci.o.na.rios 75–96.

Buvet, P.-A., 2009. Les dictionnaires électroniques du modèle des classes d'objets. Larousse, Langages 63–79.

Buvet, P.-A., 1998. Détermination et classes d'objets. Langages 32, 91–102.

Buvet, P.-A., Grezka, A., 2007. Elaboration d'outils méthodologiques pour décrire les prédicats du français. John Benjamins BV Amst., Lingvisticae Investigationes.

Buvet, P.A., Mathieu-Colas, M., 1999. Les champs 'domaine' et 'sous-domaine' dans les dictionnaires électroniques. Cahiers de Lexicologie, Centre National de la Recherche Scientifique, 173–191.

Cardie, C., Wagstaff, K., 1999. Noun phrase coreference as clustering. In Empirical Methods in Natural Language Processing.

Chaumartin, F.-R., 2007. Résolution d'anaphores dans une encyclopédie en langue anglaise : conception, implémentation et évaluation des performances. La résolution des anaphores en Traitement Automatique des Langues.

Chébouti, K., 2014. Le vocabulaire médical du point de vue des trois fonctions primaires.

Condamines, A., 2005. Anaphore nominale infidèle et hyperonymie : le rôle du genre textuel. Rev. Sémant. Pragmatique.

Condamines, A., 2002. Corpus Analysis and Conceptual Relation Patterns. Terminology 8, 141–162.

Connolly, D., Burger, J.D., Day, D.S., 1994. A machine learning approach to anaphoric reference. ACL, Proceedings of the International Conference on New Methods in Language Processing.

Corblin, F., 1995. Les formes de reprise dans le discours. Anaphores et chaînes de référence. Presses Universitaires de Rennes.

Corblin, F., Laborde, M.-C., 2004. Anaphore nominale et référence mentionnelle : Le premier, le second, l'un et l'autre, in: De Mulder, W. et Al. (2001) Anaphores Pronominales et Nominales. Etudes Pragma-Sémantiques, Rodopi., Pp 99-121.

Cornish, F., 2001. l'anaphore pronominale indirecte : une question de focus, Université de Toulouse-Le Mirail. ed, in Anaphores pronominales et nominales : études pragma-sémantiques. Faux-titre.

Dagan, I., Itai, A., 1990. Automatic acquisition of constraints for the resolution of anaphora references and syntactic ambiguities. Proceedings, International Conference on Computational Linguistic 3, 330–332.

Dessaux, A.-M., 1976. Déterminants nominaux et paraphrases prépositionnelles : problèmes de description syntaxique et sémantique du lexique. La Rousse, Langue française 30, 44–62.

Dispy, M., 2014. Lire, écrire et écouter à l'école primaire. Press. Univ. Namur.

Ducrot, O., Todorov, T., 1972. Dictionnaire encyclopédique des sciences du langage.

Fontanier, P., 1977. Les Figures du Discours, Flammarion. ed. Paris.

Garera, N., Yarowsky, D., 2006. Resolving and generating definite anaphora by modeling hypernymy using unlabeled corpora. Proceedings of the Conference on Natural Language Learning.

Gasperin, C., Vieira, R., 2004. Using word similarity lists for resolving indirect anaphora. Barcelona, Proceedings of ACL Workshop on Reference Resolution and its Applications 40–46.

Grevisse, M., 1986. Grammaire transformationelle du français, Syntaxe de l'adverbe. ed. ASSTRIL.

Grevisse, M., 1961. Le bon usage, Gembloux (Belgique). ed. Duculot.

Gross, G., 1998. Pour une typologie des prédicats nominaux. Colloq. Upps. En Linguist. Fr., Prédication, assertion, information.

Gross, G., 1996. Les expressions figées en français: noms composés et autres locutions, L'essentiel. Ophrys.

Gross, M., 1997. Construction of Local Grammars. MIT Press, Finite-State Language Processing, Cambridge, Mass. 329–352.

Gross, M., 1989. The Use of Finite Automata in the Lexical Representation of Natural Language. Comput. Sci. 377, Electronic Dictionaries and Automata in Computational Linguistics 34–50.

Grosz, B.J., Joshi, A.K., Weinstein, S., 1983. Providing a unified account of definite noun phrases in discourse pages 44–50. Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics.

Habert, B., 1997. Des corpus représentatif : de quoi, pour quoi, comment ?

Hearst, M.A., 1992. Automatique acquisition of hyponyms from large text corpora. COLING Nantes.

Hirst, G., 1981. Anaphora in Natural Language Understanding: A Survey. Lecture Notes in Computer Science.

Hobbs, J.R., 1993. Summaries from Structure. Schloss Dagstuhl, Proceedings, Workshop on Summarizing Text for Intelligent Communication.

Hobbs, J.R., 1978. Resolving Pronoun References. Lingua 311–338.

Householder, F.-W., 1981. The Syntax of Apollonius Dyscolus, Amsterdam. ed, Studies in the History of Linguistics.

Ingria, R.J.P., Stallard, D., 1989. A computational mechanism for pronominal reference. Vanc. Br. Columbia, Proceedings of the 27th Annual Meeting of the ACL 262–271.

Jean-Louis, L., 2012. Approches supervisées et faiblement supervisées pour l'extraction d'événements et le peuplement de bases de connaissances, thèse en informatique. ed.

Kassel, G., 2009. Vers une ontologie formelle des artefacts. 20es Journ. Francoph. En Ingénierie Connaiss.

Kennedy, C., Boguraev, B., 1996. Anaphora for everyone : pronominal anaphora resolution without a parser. Stroudsburg PA USA, Proceedings of the 16th conference on Computational linguistics 1, 113–118.

Kleemann-Rochas, C., Farina, G., Fernandez, M., Michel, M., 2003. Comment rédiger un rapport, un mémoire, un projet de recherche, une activité de recherche en cours ?

Kleiber, G., 2001. anaphore associative, linguistique nouvelle. PUF.

Kleiber, G., 2000. Le possessif via l'anaphore associative. Univ. Marc Bloch Strasbg. Scolia.

Kleiber, G., 1999a. Anaphore associative et relation partie-tout : condition d'aliénation et principe de congruence ontologique. Lang. Fr. 70–100.

Kleiber, G., 1999b. un puzzle rérérentiel en anaphore associative. Lingua Port. Estrut. Usos E Contrastes.

Kleiber, G., 1992. Anaphore associative et inférences. Actes VIIème Colloq. Int. Linguis-Tique Paris, Lexique et inférence(s) 175–201.

Kleiber, G., 1981. Problèmes de référence. Descriptions Définies et Noms Propres. Klincksieck.

Kleiber, G., Charolles, M., David, J., Schnedecker, C., 1994. L'anaphore associative. Aspects linguistiques, psycholiguistiques et automatiques. Paris Klincksieck.

Lappin, S., Leass, H.J., 1994. An algorithm for pronominal anaphora resolution, Computational linguistics.

Lavalley, R., 2012. Extraction automatique de segments textuels, détection de rôles, de sujets et de polarités. IUP GMI Avignon.

Le Moigne, J.-L., 1984. La théorie du système général. Théorie de la modélisation, Presses Universitaires de France. ed.

Le Pesant, D., 2002a. La détermination dans les anaphores. Langages.

Le Pesant, D., 2002b. la détermination dans les anaphores fidèles et infidèles. Langages, La détermination au regard de la diversité lexicale.

Lerat, P., 1981. Les Noms de relation. Cahiers de lexicologie 55–65.

Liang, T., Lin, Y.-H., 2005. Anaphora Resolution for Biomedical Literature by Exploiting Multiple Resources. Springer-Verl. Berl. Heidelb. 742–753.

Magri-Mourgues, V., 2015. L'anaphore rhétorique dans le discours politique. L'exemple de N. Sarkozy. Rev. Sémio-Linguist. Textes Discours.

Maillard, M., 1974. Essai de typologie des substituts diaphoriques, Langue française.

Markert, K., Nissim, M., 2005. Comparing Knowledge Sources for Nominal Anaphora Resolution. Comput. Linguist. 31.

Mathieu-Colas, M., 1998. Illustration d'une classe d'objets : les voies de communication. Langages 32, 77–90.

Mathilde Salles, 2010. Anaphore associative et relations de cohérence : une expression particulière de la relation Assertion-Indice. Discours En Ligne.

Mejri, S., 2009. Le mot : problématique théorique. Le Français Moderne 68–82.

Mel'cuk, I.A., Clas, A., Polguère, A., 1995. Introduction à la lexicographie explicative et combinatoire. Duculot.

Milner, J.-C., 1982. Ordres et raisons de langue, Seuil. ed. Paris.

Milner, J.-C., 1976. Réflexions sur la référence. Larousse, Lexique et Grammaire.

Mitkov, R., 2003. The Oxford Handbook of Computational linguistics, Ruslan Mitkov. ed, Oxford Handbooks in Linguistics.

Mitkov, R., 1999. Anaphora resolution: the state of the art. Sch. Lang. Eur. Stud. Univ. Wolverh.

Mitkov, R., 1998a. Evaluating anaphora resolution approaches.

Mitkov, R., 1998b. Robust pronoun resolution with limited knowledge.

Mitkov, R., 1996. Anaphora resolution: a combination of linguistic and statistical approaches. Lanc. UK, Proceedings of the Discourse Anaphora and Anaphor Resolution (DAARC'96).

Mouton, C., 2007. Résolution des anaphores : Implémentation et Evaluation de l'algorithme de Lappin&Leass.

Nasukawa, T., 1994. Robust method of pronoun resolution using full-text information. Proceedings of the 15th International Conference on Computational Linguistics 1157–1163.

Nazarenko, A., 2004. Donner accès au contenu des documents textuels : acquisition de connaissances et analyse de corpus spécialisés.

Neveu, F., 2004. Dictionnaire des sciences du langage. Armand Colin.

Ng, V., 2010. Supervised Noun Phrase Coreference Research: The First Fifteen Years. Proc. 48th Annu. Meet. Assoc. Comput. Linguist. 1396–1411.

Ng, V., Cadie, C., 2002. Improving Machine Learning Approcahes to Corefrence Resolution. In Proceedings of ACL'02 104–111.

Nouioua, F., 2007. Heuristique pour la résolution d'anaphores dans les textes d'accidents de la route. La résolution des anaphores en Traitement Automatique des Langues.

Paumier, S., 2015. Unitex 3.1beta - Manuel d'utilisation.

Pepin, L., 2009. La coréférence dans la narration, première partie. Univ. Qué. À Rimouski Format PDF, 16 pages.

Perdicoyanni-Paléologou, H., 2001. Le concept d'anaphore, de cataphore et de déixis en linguistique française. Rev. Québécoise Linguist., UQAM, Montréal.

Poesio, M., Paolo-Ponzetto, S., Versley, S., 2010. Computational models of anaphora resolution : a survey.

Poibeau, T., 2005. Parcours interprétatifs et terminologie. Rouen, Actes TIA.

Polguère, A., 2003. Lexicologie et sémantique lexicale: notions fondamentales. Les presses de l'Université de Montréal.

Posturzynska-Bosko, M., 2009. La détermination du nom anaphorique associatif en moyen français dans Le Livre des fais et bonnes meurs du sage Roy Charles V de Christine de Pizan.

Rey, A., 2010. Dictionnaire historique de la langue Française, Le Rober SEJER.

Riegel, M., Pellat, J.C., Rioul, R., 1994. Grammaire méthodique du français. Paris : Presses universitaires de France.

Sagot, B., Danlos, L., 2008. Méthodologie lexicographique de constitution d'un lexique syntaxique de référence pour le français. Nancy Fr., Actes du colloque Lexicographie et informatique : bilan et perspectives.

Sagot, B., Fišer, D., 2008. Construction d'un wordnet libre du français à partir de ressources multilingues. TALN, Traitement Automatique des Langues Naturelles.

Sahiri, L., 2013. Le bon usage de la répétition dans l'expression écrite et orale. Mon petit éditeur.

Saint-Dizier, P., 2006. SÉMANTICLOPÉDIE.

Salmon-Alt, S., 2001. Du corpus à la théorie : l'annotation (co-)référentielle. Hermès Paris, Traitement Automatique des Langues (T.A.L.).

Saussure, F., 1967. Cours de la linguistique générale.

Sidner, C., 1979. Towards a Computational Theory of Definite Anaphora Comprehension in English Discourse, Massachusetts Institute of Technology, USA. ed, Ph.D. thesis.

Soon, W.M., Ng, H.T., Lim, D.C.Y., 2001. A machine learning approach to coreference resolution of Noun phrases. Computational Linguistique 521–544.

Tamba, I., 1991. Organisation hiérarchique et relations de dépendance dans le lexique. Inf. Gramm. 50, 43–47.

Theissen, A., 2001. La concurrence entre un SN défini fidèle et un SN défini totalement fidèle, Faux Titre. ed.

Van Peteghem, M., 2001. « Autre » et « même » sans nom : anaphore nominale ou pronominale, Anaphores pronominales et nominales - etude pragma sémantique. Faux Titre.

Vieira, R., Poesio, M., 2000. An Empirically-Based System for Processing Definite Descriptions. Computational Linguistics 539–593.

Weissenbacher, D., Nazarenko, A., 2007. Identifier les pronoms anaphoriques et trouver leurs antécédents : l'intérêt de la classification bayésienne. ATALA Fr., Traitement Automatique des Langues Naturelles 145–155.

Whittake, S., Handelshøyskole, N., 2002. Portrait de l'expression anaphorique Ledit N. Rom. Forum.

ANNEXE

1. LE TRAITEMENT DES ANAPHORES INFIDELES

Le script pour aspirer le corpus de fait-divers (sur la page sudinfo.be):

```
# -*- coding: utf-8 -*-

import re
import urllib2
from bs4 import BeautifulSoup

# ce programme donnera en sortie un fichier url.txt qui contient les url obtenus lorsqu'on
# aspire le site : http://www.sudinfo.be/12/actualite/faits-divers

# générer la liste des noms de pages de 0 à 3
def generation_url() :
    liste_page = []
    for i in range (0, 15) :
        pages = "http://www.sudinfo.be/12/actualite/faits-divers?page=%i" % i
        liste_page.append(pages)
    return liste_page

# input : nom d'une page, output : le contenu html
def print_content(page_url) :
    hdr = {'User-Agent':'Mozilla/5.0'}
    req = urllib2.Request(page_url,headers=hdr)
    response = urllib2.urlopen(req)
    html = response.read()
    return html

# input : 1 url, output : liste url
def out_url_all(url) :
    text_input = print_content(url)
    url_out = []
    for match in re.finditer (r"\<h2\> \<a href=\"\/((.)*)\"\>", text_input):
        url_list = "http://www.sudinfo.be/%s" % match.group(1)
        url_out.append(url_list)
    return url_out

# input : rien, output : liste all url
def all_url_list() :
    list_url_in = generation_url()
    list_url_out = []
    for url in list_url_in :
        list_url_out.append(out_url_all(url))
```

```
def all_url_list() :
    list_url_in = generation_url()
    list_url_out = []
    for url in list_url_in :
        list_url_out.append(out_url_all(url))
    return list_url_out

fileout = open("corpus_FD.txt", "w")

def parse_url():
    liste_url = all_url_list()
    i = 0
    for l_url in liste_url :
        for url in l_url :
            print "%d - %s" % (i, url)
            i = i + 1
            hdr = {'User-Agent':'Mozilla/5.0'}
            req = urllib2.Request(url,headers=hdr)
            try:
                resp = urllib2.urlopen(req)
                html = resp.read()
            except urllib2.HTTPError, error:
                html = error.read()
            soup = BeautifulSoup(html)
            for t in soup.find_all("h1") :
                data = t.get_text().encode("utf-8")
                fileout.write("-------------------------\nTitre : %s\n" % data)
            for st in soup.find_all("div", {"class": "chapeau"}) :
                data = st.get_text().encode("utf-8")
                fileout.write("Soustitre : %s" % data)
            for p in soup.find_all("div", {"id": "body_text"}) :
                data = p.get_text().encode("utf-8")
                fileout.write("%s\n" % data)


parse_url()
fileout.close()
```

Le corpus obtenu après l'apiration :

Titre : Nord: un couple trouvé mort dans sa maison, le mari a tué sa femme avant de se suicider
Soustitre :
Un couple a été trouvé mort jeudi soir dans sa maison de Petite-Forêt (Nord), le mari ayant probablement tué sa femme avec un fusil de chasse avant de
sources concordantes.

« Nous avons découvert deux corps dans le domicile familial. Il semblerait que le mari ait tiré sur sa femme avec un fusil de chasse avant de retourner
quelques journalistes sur place une substitut du procureur de Valenciennes.
Le drame serait survenu peu avant 20h, a confirmé la préfecture du Nord.
Il s'agirait d'un couple d'une cinquantaine d'années, parents de trois jeunes filles, « fort discret » et en instance de divorce, selon le voisinage, i
l'AFP.
« J'ai vu l'une de leurs filles sortir de leur maison en criant +à l'aide !+. Je me suis rendu sur place et ai vu les deux corps », a raconté un autre
La police de Valenciennes ainsi que la police scientifique s'est rendue sur les lieux.
De nombreux voisins s'étaient regroupés dans la soirée autour de cette maison individuelle, avec un jardin devant, située dans un quartier résidentiel
de cette petite commune d'environ 5.000 habitants, selon un photographe de l'AFP sur place.

Titre : Six mois avec sursis pour un enseignant juif qui avait inventé une agression
Soustitre :
Un enseignant juif qui était accusé d'avoir inventé une agression antisémite quelques jours après les attentats de Paris, suscitant alors un vif émoi,
six mois de prison avec sursis.

Cette condamnation ne sera pas inscrite au casier judiciaire de cet homme de 57 ans, qui a maintenu devant le tribunal sa version des faits.
En novembre, quelques jours après les attentats qui ont fait 130 morts à Paris, il avait affirmé avoir été agressé au couteau par trois hommes se rever
(EI).
«La vérité, c'est qu'il n'a pas été agressé comme il le dit», a asséné le procureur André Ribes, soulignant les doutes émis par toutes les personnes im
policiers, médecins, experts et insistant sur le sérieux de l'enquête menée par le parquet dans un contexte tendu après les attentats parisiens.
«Je n'ai jamais vu des blessures réelles à l'arme blanche comme celles-là», a encore lancé le représentant du ministère public, évoquant des problèmes

Le corpus après le prétraitement avec Unitex :

{S}Titre : Nord: un <person style='color: red;' class='generique' g='m' n='s'>couple</person> trouvé mort dans sa maison, le <person st
color: red;' class='famille' g='m' n='s'>mari</person> a tué sa <person style='color: red;' class='generique' g='f' n='s'>femme</pe
avant de se suicider
{S}Soustitre :
{S}Un <person style='color: red;' class='generique' g='m' n='s'>couple</person> a été trouvé mort jeudi soir dans sa maison de Petite-F
Nord), le <person style='color: red;' class='famille' g='m' n='s'>mari</person> ayant probablement tué sa <person style='color: red;' c
generique' g='f' n='s'>femme</person> avec un fusil de chasse avant de se suicider, a-t-on appris de sources concordantes.
« Nous avons découvert deux corps dans le domicile familial.{S} Il semblerait que le <person style='color: red;' class='famille' g='m'
mari</person> ait tiré sur sa <person style='color: red;' class='generique' g='f' n='s'>femme</person> avec un fusil de chasse avant de
retourner l'arme contre lui », a affirmé à quelques <person style='color: red;' class='metier' g='m' n='p'>journalistes</person> sur pl
person style='color: red;' class='metier' g='m' n='s'>substitut du procureur</person> de Valenciennes.
{S}Le drame serait survenu peu avant 20h, a confirmé la préfecture du Nord.
{S}Il s'agirait d'un <person style='color: red;' class='generique' g='m' n='s'>couple</person> d'une cinquantaine d'années, <person sty
: red;' class='generique' g='m' n='p'>parents</person> de trois <person style='color: red;' class='generique' g='f' n='p'>jeunes fi
person>, « fort discret » et en instance de divorce, selon le voisinage, interrogé par un correspondant de l'AFP.
« J'ai vu l'une de leurs <person style='color: red;' class='generique' g='f' n='p'>filles</person> sortir de leur maison en criant +à l
!+.{S} Je me suis rendu sur place et ai vu les deux corps », a raconté un autre de leurs <person style='color: red;' class='generique'
p'>voisins</person>.
{S}La <person style='color: red;' class='metier' g='m' n='s'>police</person> de Valenciennes ainsi que la <person style='color: red;' c
metier' g='m' n='s'>police</person> scientifique s'est rendue sur les lieux.
{S}De nombreux <person style='color: red;' class='generique' g='m' n='p'>voisins</person> s'étaient regroupés dans la soirée autour de
maison individuelle, avec un jardin devant, située dans un quartier résidentiel et populaire, habituellement calme, de cette petite com
d'environ 5.000 <person style='color: red;' class='generique' g='m' n='p'>habitants</person>, selon un <person style='color: red;' clas

Le script pour transformer le corpus unitex au format xml :

```
#!/usr/bin/python
# -*- coding: utf-8 -*-

from xml.etree.ElementTree import ElementTree
from xml.etree.ElementTree import Element
import xml.etree.ElementTree as etree
import codecs

# lire fichier d'entrée, vérifier à supprimer le {S} avant Titre
# file_in = codecs.open("FD_1.html", "r", "utf-8")
file_in = codecs.open("FD_cleaned.html", "r", "utf-8")
text = file_in.read()
messages = text.split("_____")

# préparer le contenu XML
root=Element('document') # créer tag document

i = 1
for message in messages :
    article=Element('article') # tao tag article
    root.append(article) # tag article dans tag document
    article.set('id', str(i)) # creer attributs aux articles
    i = i + 1
    phrases = message.split("{S}")
    j = 0
    for phrase in phrases :
        title_mark = "Titre : "
        subtitle_mark = "Soustitre : "
        if title_mark in phrase :
            title=Element('titre') # créer tag titre
            article.append(title) # tag titre dans tag article
            title.text = phrase
        elif subtitle_mark in phrase :
            subtitle=Element('soustitre') # créer tag titre
            article.append(subtitle) # tag titre dans tag article
            subtitle.text = phrase
        else :
            line=Element('phrase') # créer tag phrase
            article.append(line) # tag phrase dans tag article
            line.text = phrase
            line.set('id', str(j)) # creer attributs à la phrase
            j = j + 1

# creation d'un fichier XML de sortie
tree=ElementTree(root)
file_out = codecs.open("FD_cleaned.xml","w", "utf-8")
file_out.write("<?xml version='1.0' encoding='UTF-8' ?>")
tree.write(file_out)
```

Le corpus transformé au format xml :

```
<?xml version="1.0" encoding="UTF-8"?>
<document>
  <article id="1">
    <phrase id="0"/>
    <titre>Titre : Nord: un <person det='ind' g='m' nb='s' style='color: red;' class='generique'>couple</person> trouvé mort dans sa maison,
      g='m' nb='s' style='color: red;' class='famille'>mari</person> a tué sa <person det='poss' g='f' nb='s' style='color: red;' class='gene
      avant de se suicider </titre>
    <phrase id="1">Soustitre : </phrase>
    <phrase id="2">un <person det='ind' g='m' nb='s' style='color: red;' class='generique'>couple</person> a été trouvé mort jeudi soir dans
      Forêt (Nord), le <person det='def' g='m' nb='s' style='color: red;' class='famille'>mari</person> ayant probablement tué sa <person
      style='color: red;' class='generique'>femme</person> avec un fusil de chasse avant de se suicider, a-t-on appris de sources concorda
      découvert deux corps dans le domicile familial.</phrase>
    <phrase id="3"> Il semblerait que le <person det='def' g='m' nb='s' style='color: red;' class='famille'>mari</person> ait tiré sur sa <pers
      style='color: red;' class='generique'>femme</person> avec un fusil de chasse avant de retourner l'arme contre lui », a affirmé à quelc
      g='m' nb='p' style='color: red;' class='metier'>journalistes</person> sur place une <person det='ind' g='f' nb='s' style='color: red;' c
      du procureur</person> de Valenciennes. </phrase>
    <phrase id="4">Le drame serait survenu peu avant 20h, a confirmé la préfecture du Nord. </phrase>
    <phrase id="5">Il s'agirait d'un <person det='ind' g='m' nb='s' style='color: red;' class='generique'>couple</person> d'une cinquantaine
      <person det='ind' g='f' nb='p' style='color: red;' class='generique'>jeunes filles</person>, « fort discret » et en instance de divorce,
      interrogé par un correspondant de l'AFP. « J'ai vu l'une de leurs <person det='poss' g='f' nb='p' style='color: red;' class='generique'>f
      leur maison en criant +à l'aide !+.</phrase>
    <phrase id="6"> Je me suis rendu sur place et ai vu les deux corps », a raconté un autre de leurs <person det='poss' g='m' nb='p' style='co
      class='generique'>voisins</person>. </phrase>
    <phrase id="7">la <person det='def' g='f' nb='s' style='color: red;' class='metier'>police</person> de Valenciennes ainsi que la <person
      style='color: red;' class='metier'>police</person> scientifique s'est rendue sur les lieux. </phrase>
    <phrase id="8">De nombreux <person det='ind' g='m' nb='p' style='color: red;' class='generique'>voisins</person> s'étaient regroupés d
      cette maison individuelle, avec un jardin devant, située dans un quartier résidentiel et populaire, habituellement calme, de cette petite
      <person det='ind' g='m' nb='p' style='color: red;' class='generique'>habitants</person>, selon un <person det='ind' g='m' nb='s' sty
      class='metier'>photographe</person> de l'AFP sur place. </phrase>
  </article>
  <article id="2">
    <phrase id="0"> </phrase>
    <titre>Titre : Six mois avec sursis pour un <person det='ind' g='m' nb='s' style='color: red;' class='metier'>enseignant</person> juif qui
      agression </titre>
    <phrase id="1">Soustitre : </phrase>
    <phrase id="2">un <person det='ind' g='m' nb='s' style='color: red;' class='metier'>enseignant</person> juif qui était accusé d'avoir inve
```

Le script de traitement des anaphores infidèles (résultat affiché en tableau) :

```python
#!/usr/bin/python2
# -*- coding: utf-8 -*-

#./searchAntecedent.py -n 2 test_input.xml
import os,sys,re,getopt

try:
    import xml.etree.cElementTree as ET
except ImportError:
    import xml.etree.ElementTree as ET

usage = "usage: cherchePerson.py -n antecedentCutOff xmlFile.xml\n"
xmlFile = ""
cutOff = 100 #par défault il n'y a pas de limite pour chercher l'antécédent

#parsing de la ligne de paramètre
try:
    opts, args = getopt.getopt(sys.argv[1:],"n:",["-n"])
except getopt.GetoptError, err:
    # print help information and exit:
    print str(err) # will print something like "option -a not recognized"
    sys.stderr.write(usage) #afficher le message d'erreur
    sys.exit() #quitter le programme

for opt, arg in opts:
    if opt in ("-n"):
        cutOff = int(arg)
    else:
        sys.stderr.write("Error: unrecognized option '%s'\n"%opt)
        exit()
for arg in args:
    if xmlFile=="":
        xmlFile=arg
    else:
        sys.stderr.write(usage) #afficher le message d'erreur
        sys.exit() #quitter le programme
        exit()

if xmlFile == "":
    sys.stderr.write(usage) #afficher le message d'erreur
    sys.exit() #quitter le programme
```

```
#==========


#classe pour stocker les informations sur les antécédents
class antecedent():
    #initialisation de la class à partir des paramètres qui sont passés
    def __init__(self, text, clas, gen, nb, sentId, articleID):
        self.label = text
        self.CC = clas # CC contient la class
        self.G = gen # G contient le genre
        self.N = nb # N contient le nombre
        self.aId = int(articleID) # contient le numéro de l'article
        self.sId = int(sentId) # contient le numéro de phrase
        # dictionnaire qui contient les scores de saillence pour tous les antécédents. La clé est l'antécédent, la valeur est le score
        self.scoreDict = {}
    def computeScore(self, antecedentToCompare):
        total = 0 #initialisation du score à 0
        score_CC, score_G, score_N = 0,0,0
        # distance en nombre de phrase entre les persons
        sentDist = abs(self.sId - antecedentToCompare.sId)
        if (sentDist == 0):
            score_proxi = 1
        elif sentDist > 1 and sentDist <= 2:
            score_proxi = 0.8
        elif sentDist > 2 and sentDist <= 4:
            score_proxi = 0.7
        else:
            score_proxi = 0
        #facon simple de calculer le score :
        #si classA = classB alors score_CC = 1, sinon score_CC = 0
        # on peut changer pour que cela soit plus souple que 0 ou 1
        if (self.CC == antecedentToCompare.CC):
            score_CC = 1
        if (self.G == antecedentToCompare.G):
            score_G = 1
        if (self.N == antecedentToCompare.N):
            score_N = 1
        total = float(score_CC + score_G + score_N + score_proxi)/ 4
        total = round(total, 2)
```

```
        total = round(total, 2)
        # ajouter le score de saillance au dictionnaire -- note vérifier les doublons dans les antécédents
        self.scoreDict[antecedentToCompare] = total
        return total

# fonction qui lit le fichier xml et renvoie une liste des persons dans l'ordre du document
def readXml(inFile):
    outLst = [] # liste qui contient les persons d'un document
    tree = ET.ElementTree(file=inFile)
    #expression xpath pour parcourir chaque article
    articleLst = tree.findall('//article')
    for article in articleLst:
        #identifiant de l'article
        articleID = article.get("id")
        for phrase in list(article):
            #identifiant de la phrase dans l'article
            phraseId = phrase.get("id")
            #parcourir chaque balise person dans la phrase courante
            personLst = phrase.findall(".//person")
            # parcours toutes les balises <person> du fichier et construit un object antecedent pour chaque person
            for person in personLst:
                person_attrib = person.attrib # person_attrib est un dictionnaire qui contient toutes les valeurs
                #construire un object antecendent à partir du tag
                obj_antecendent = antecedent(person.text, person_attrib['class'], person_attrib['g'], person_attrib['n'], phraseId, articleID)
                #ajoute l'antecedent à la liste
                outLst.append(obj_antecendent)
    return outLst

# classe principale
def main():
    #lire le fichier xml et stocker les persons dans une liste
    #la liste les ajoute dans leurs ordres d'apparition dans le document
    person_lst = readXml(xmlFile) # contient la liste des persons du document. Précisément la liste des instances de la classe antecedent.

    # pour chaque person de la liste on cherche son antécédent. On commence par la fin de la liste: on prends le dernier élément de la liste et on calcul
    inv_personLst = list(reversed(person_lst)) # inverser la liste des persons pour itérer à partir du dernier élément (le dernier élément devient le prem
    for i in range(0,len(person_lst)): #itération à partir du dernier élément
        current_person = inv_personLst[i]
        for j in range(i+1, len(person_lst)):
            prev_person = inv_personLst[j]
```

```python
# classe principale
def main():
    #lire le fichier xml et stocker les persons dans une liste
    #la liste les ajoute dans leurs ordres d'apparition dans le document
    person_lst = readXml(xmlFile) # contient la liste des persons du document. Précisément la liste des instances de la classe antecedent.

    # pour chaque person de la liste on cherche son antécédent. On commence par la fin de la liste:
    #on prends le dernier élément de la liste et on calcul le score de saillence avec tous les éléments précédents
    inv_personLst = list(reversed(person_lst)) # inverser la liste des persons pour itérer à partir du dernier élément (le dernier élément devient le p
    for i in range(0,len(person_lst)): #itération à partir du dernier élément
        current_person = inv_personLst[i]
        for j in range(i+1, len(person_lst)):
            prev_person = inv_personLst[j]
            #ne pas calculer le score lorsque les person dans la limite
            if current_person.aId != prev_person.aId:
                continue
            #on vérifie que la distance entre l'antécédent et le person est inférieure à la limite cutOff
            if abs(current_person.sId - prev_person.sId) > cutOff :
                continue
            #calculer le score entre le person courant et l'antécédent
            current_person.computeScore(prev_person)
    #pour chaque person affiche le texte du person et les antécédents avec leurs scores
    fichier = open("resultat_FD.txt", "w")
    for person in person_lst:
        outLine = person.label.strip() + " : " + " | ".join([("%s = %s")%(k.label.strip(),v) for (k,v) in sorted(person.scoreDict.iteritems(),
            key=lambda x:(-1)*x[1])])
        fichier.write(outLine.encode("utf-8"))
        fichier.write("\n")
    fichier.close()

main()
```

Résultat de la résolution (affiché en tableau) :

```
mari :
femme : mari = 0.5
mari : mari = 1.0 | femme = 0.5
femme : femme = 1.0 | mari = 0.5 | mari = 0.5
mari : mari = 1.0 | mari = 1.0 | femme = 0.5 | femme = 0.5
femme : femme = 1.0 | femme = 1.0 | mari = 0.5 | mari = 0.5 | mari = 0.5
journalistes : mari = 0.5 | mari = 0.5 | mari = 0.5 | femme = 0.25 | femme = 0.25 | femme = 0.25
substitut du procureur : journalistes = 0.75 | mari = 0.75 | mari = 0.75 | mari = 0.75 | femme = 0.5 | femme = 0.5 | femme = 0.5
parents : journalistes = 0.7 | substitut du procureur = 0.45 | femme = 0.45 | mari = 0.45 | femme = 0.45 | mari = 0.45 | femme = 0.45 | mari = 0.45
jeunes filles : parents = 0.75 | femme = 0.7 | femme = 0.7 | femme = 0.7 | journalistes = 0.45 | substitut du procureur = 0.2 | mari = 0.2 | mari = 0.2 | ma
filles : jeunes filles = 1.0 | parents = 0.75 | femme = 0.7 | femme = 0.7 | femme = 0.7 | journalistes = 0.45 | substitut du procureur = 0.2 | mari = 0.2 |
voisins : parents = 0.75 | jeunes filles = 0.5 | filles = 0.5
voisins : voisins = 0.95
photographe : voisins = 0.5 | voisins = 0.45
homme :
hommes : homme = 0.5
procureur : homme = 0.5 | hommes = 0.5
pompiers : procureur = 0.75 | hommes = 0.75 | homme = 0.25
policiers : pompiers = 1.0 | procureur = 0.75 | hommes = 0.75 | homme = 0.25
experts : pompiers = 1.0 | policiers = 1.0 | procureur = 0.75 | hommes = 0.75 | homme = 0.25
quinquagénaire : homme = 0.95 | procureur = 0.5 | hommes = 0.5 | pompiers = 0.25 | policiers = 0.25 | experts = 0.25
responsables : pompiers = 0.75 | policiers = 0.75 | experts = 0.75 | procureur = 0.5 | hommes = 0.5 | quinquagénaire = 0.5 | homme = 0.45
président : procureur = 0.95 | pompiers = 0.7 | policiers = 0.7 | experts = 0.7 | responsables = 0.5 | quinquagénaire = 0.5 | hommes = 0.45
professeur : président = 0.75 | responsables = 0.7 | quinquagénaire = 0.7
adolescent : quinquagénaire = 0.95 | professeur = 0.75 | président = 0.5 | responsables = 0.45
turc : adolescent = 0.75 | professeur = 0.75 | quinquagénaire = 0.7 | président = 0.5 | responsables = 0.45
homme :
jeune femme :
garçon : jeune femme = 0.75
Britannique : garçon = 0.75 | jeune femme = 0.5
jeune fille : jeune femme = 1.0 | garçon = 0.75 | Britannique = 0.5
jeune femme : jeune fille = 1.0 | jeune femme = 1.0 | garçon = 0.75 | Britannique = 0.5
Britannique : Britannique = 0.75 | garçon = 0.5 | jeune fille = 0.25 | jeune femme = 0.25 | jeune femme = 0.25
copains : garçon = 0.5 | Britannique = 0.5 | jeune femme = 0.25 | jeune fille = 0.25 | jeune femme = 0.25 | Britannique = 0.25
amies : copains = 0.75 | jeune femme = 0.5 | jeune fille = 0.5 | jeune femme = 0.5 | Britannique = 0.25 | garçon = 0.25 | Britannique = 0.0
ami : garçon = 0.95 | jeune femme = 0.7 | jeune fille = 0.7 | Britannique = 0.7 | jeune femme = 0.7 | Britannique = 0.5 | copains = 0.5 | amies = 0.25
jeune femme : jeune femme = 0.95 | jeune fille = 0.95 | jeune femme = 0.95 | ami = 0.75 | garçon = 0.7 | amies = 0.5 | Britannique = 0.45 | Britannique = 0.
jeune fille : jeune femme = 0.75 | amies = 0.7 | ami = 0.5 | copains = 0.45 | Britannique = 0.45
```

Le script de traitement des anaphores infidèles (résultat au format html) :

```
#!/usr/bin/python2
# -*- coding: utf-8 -*-


import os,sys,re,getopt

try:
    import xml.etree.cElementTree as ET
except ImportError:
    import xml.etree.ElementTree as ET

#classe pour stocker les informations sur les antécédents
class antecedent():
    #initialisation de la class à partir des paramètres qui sont passés
    def __init__(self, personTail, personId, text, clas, gen, nb, sentId, articleID):
        self.label = text
        self.suite = personTail
        self.id = personId
        self.CC = clas # CC contient la class
        self.G = gen # G contient le genre
        self.N = nb # N contient le nombre
        self.aId = int(articleID) # contient le numéro de l'article
        self.sId = int(sentId) # contient le numéro de phrase
        self.scoreDict = {} # dictionnaire qui contient les scores de saillence pour tous les antécédents
        # La clé est l'antécédent, la valeur est le score

    def computeScore(self, antecedentToCompare):
        total = 0 #initialisation du score à 0
        score_CC, score_G, score_N = 0,0,0
        # distance en nombre de phrase entre les persons
        sentDist = abs(self.sId - antecedentToCompare.sId)
        if (sentDist == 0):
            score_proxi = 1
        elif sentDist > 1 and sentDist <= 2:
            score_proxi = 0.8
```

```
        elif sentDist > 1 and sentDist <= 2:
            score_proxi = 0.8
        elif sentDist > 2 and sentDist <= 4:
            score_proxi = 0.7
        else:
            score_proxi = 0
        # calculer le score : si classA = classB alors score_CC = 1, sinon score_CC = 0
        if (self.CC == antecedentToCompare.CC):
            score_CC = 1
        if (self.G == antecedentToCompare.G):
            score_G = 1
        if (self.N == antecedentToCompare.N):
            score_N = 1
        total = float(score_CC + score_G + score_N + score_proxi)/ 4
        total = round(total, 2)

        # ajouter le score de saillance au dictionnaire -- note vérifier les doublons dans les antécédents
        self.scoreDict[antecedentToCompare] = total
        return total

# fonction qui lit le fichier xml et renvoie une liste des persons dans l'ordre du document
def readXml(inFile):
    outLst = [] # liste qui contient les persons d'un document
    tree = ET.ElementTree(file=inFile)
    # expression xpath pour parcourir chaque article
    articleLst = tree.findall('//article')
    for article in articleLst:
        i = 1
        #identifiant de l'article
        articleID = article.get("id")
        for phrase in list(article):
            Lst = [] # liste qui contient les phrase.text + les persons
            #identifiant de la phrase dans l'article
            phraseId = phrase.get("id")
            if phrase.text :
```

```
            phraseId = phrase.get("id")
            if phrase.text :
                phraseText = phrase.text # vérifier si le text n'est pas vide, prendre le texte de <phrase>
            else :
                phraseText = "" # si le texte est vide, il aura le NoneType et non pas du type str
            #parcourir chaque balise <person> dans la phrase courante
            personLst = phrase.findall(".//person")
            # parcours toutes les balises <person> du fichier et construit un object antecedent pour chaque person
            for person in personLst:
                person_id = i
                i = i + 1
                personLst[0].text = phraseText + "<person style=\"color: red;\">" + person.text + "</person>"
                if person.tail :
                    person_Tail = person.tail # vérifier si le text après <person> n'est pas vide
                else :
                    person_Tail = ""
                person_attrib = person.attrib # person_attrib est un dictionnaire qui contient toutes les valeurs : class, g, n
                #construire un object antecendent à partir du tag
                obj_antecendent = antecedent(person_Tail, person_id, person.text, person_attrib['class'], \
                                  person_attrib['g'], person_attrib['n'], phraseId, articleID)
                outLst.append(obj_antecendent) #ajoute l'antecedent à la liste
    return outLst

# classe principale
def main():
    #lire le fichier xml et stocker les persons dans une liste
    #la liste les ajoute dans leurs ordres d'apparition dans le document
    person_lst = readXml("FD_ok.xml") # contient la liste des persons du document. Précisément la liste des instances de la class
    # pour chaque person de la liste on cherche son antécédent.
    # On commence par la fin de la liste pour comparer l'élément suivant avec l'élément précédent
    # on prends le dernier élément de la liste et calcul le score de saillance avec tous les éléments précédents
    inv_personLst = list(reversed(person_lst)) # inverser la liste des persons pour itérer à partir du dernier élément
    # (le dernier élément devient le premier).
    for i in range(0,len(person_lst)): #itération à partir du dernier élément
```

```
    # (le dernier élément devient le premier).
    for i in range(0,len(person_lst)): #itération à partir du dernier élément
        current_person = inv_personLst[i]
        for j in range(i+1, len(person_lst)):
            prev_person = inv_personLst[j]
            # ne pas calculer le score lorsque les person dans la limite
            if current_person.aId != prev_person.aId:
                continue
            # on vérifie que la distance entre l'antécédent et le person est inférieure à la limite cutOff(2)
            if abs(current_person.sId - prev_person.sId) > 2 :
                continue
            # calculer le score entre le person courant et l'antécédent
            current_person.computeScore(prev_person)
    #pour chaque person affiche le texte du person et les antécédents avec leurs scores
    fichier = open("resultat_ok.html", "w")
    for person in person_lst:
        outLine = "<p><person>" + person.label.strip() + "</person><sub style=\"color: red;\"> " + \
                    str(person.id) + "</sub><sup style=\"color: blue;\"> " + " | ".join([("%i(%s)")%(k.id,v) \
                    for (k,v) in sorted(person.scoreDict.iteritems(), key=lambda x:(-1)*x[1]) if v > 0.75 ]) + "</sup>" + person.suite
        fichier.write("<?xml version=\"1.0\" encoding=\"UTF-8\"?>")
        fichier.write(outLine.encode("utf-8"))
        fichier.write("\n")
    fichier.close()

main()

```

Résultat de la résolution (affiché au format html) :

Titre : Nord: un couple$_1$ trouvé mort dans sa maison, le

mari$_2$ a tué sa

femme$_3$ avant de se suicider

un couple$_4$ $^{1(0.95)}$ a été trouvé mort jeudi soir dans sa maison de Petite-Forêt (Nord), le

mari$_5$ $^{2(0.95)}$ ayant probablement tué sa

femme$_6$ $^{3(0.95)}$ avec un fusil de chasse avant de se suicider, a-t-on appris de sources concordantes. « Nous avons découvert deux corps dans le domicile familial.

Il semblerait que le mari$_7$ ait tiré sur sa

femme$_8$ avec un fusil de chasse avant de retourner l'arme contre lui », a affirmé à quelques

journalistes$_9$ sur place une

substitut du procureur$_{10}$ de Valenciennes.

Il s'agirait d'un couple$_{11}$ d'une cinquantaine d'années, parents de trois

jeunes filles$_{12}$, « fort discret » et en instance de divorce, selon le voisinage, interrogé par un correspondant de l'AFP. « J'ai vu l'une de leurs

filles$_{13}$ $^{12(1.0)}$ sortir de leur maison en criant +à l'aide !+.

Je me suis rendu sur place et ai vu les deux corps », a raconté un autre de leurs voisins$_{14}$.

la police$_{15}$ de Valenciennes ainsi que la

police$_{16}$ $^{15(1.0)}$ scientifique s'est rendue sur les lieux.

De nombreux voisins$_{17}$ $^{14(0.95)}$ s'étaient regroupés dans la soirée autour de cette maison individuelle, avec un jardin devant, située dans un quartier résidentiel et populaire, hab
petite commune d'environ 5.000

habitants$_{18}$ $^{17(1.0)|14(0.95)}$, selon un

photographe$_{19}$ de l'AFP sur place.

Titre : Six mois avec sursis pour un enseignant$_1$ juif qui avait inventé une agression

un enseignant$_2$ $^{1(0.95)}$ juif qui était accusé d'avoir inventé une agression antisémite quelques jours après les attentats de Paris, suscitant alors un vif émoi, a été condamné jeudi
avec sursis.

Cette condamnation ne sera pas inscrite au casier judiciaire de cet homme$_3$ de 57 ans, qui a maintenu devant le tribunal sa version des faits.

En novembre, quelques jours après les attentats qui ont fait 130 morts à Paris, il avait affirmé avoir été agressé au couteau par trois hommes$_4$ se revendiquant du groupe État isl
qu'il n'a pas été agressé comme il le dit», a asséné le

procureur$_5$ $^{2(0.95)}$ André Ribes, soulignant les doutes émis par toutes les

personnes$_6$ $^{4(1.0)}$ impliquées dans le dossier, pompiers, policiers, médecins, experts et insistant sur le sérieux de l'enquête menée par le parquet dans un contexte tendu après les
jamais vu des blessures réelles à l'arme blanche comme celles-là», a encore lancé le représentant du ministère public, évoquant des problèmes conjugaux comme possible mob

enseignant$_7$ $^{5(1.0)|2(0.95)}$.

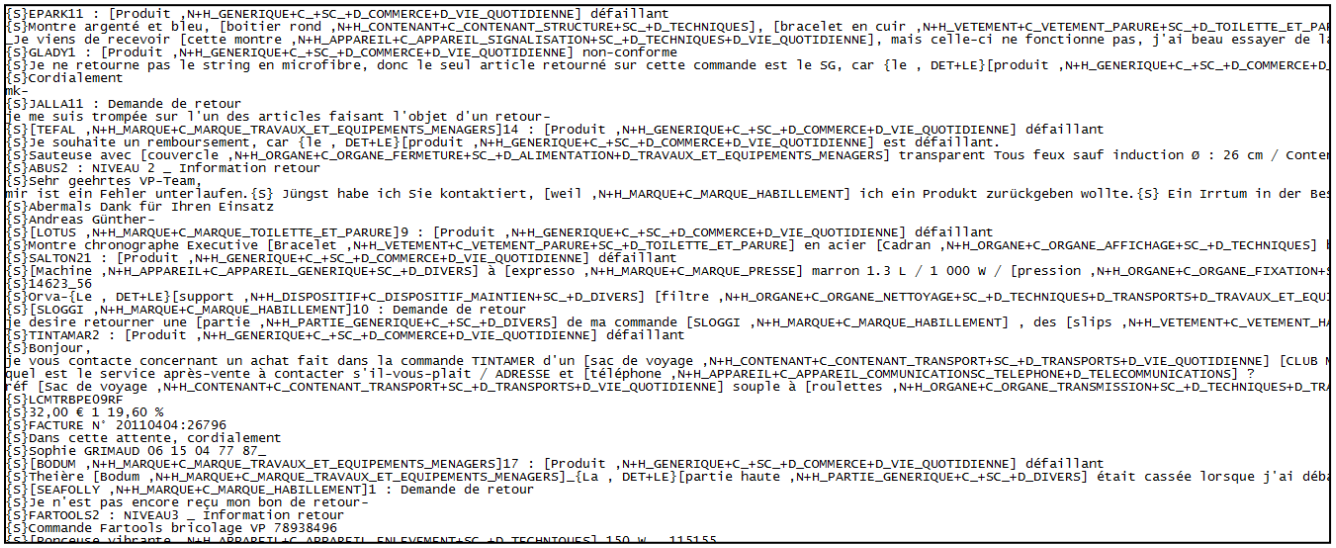
2. LE TRAITEMENT DES ANAPHORES ASSOCIATIVES

La classe de calcul du score dans la résolution des anaphores associative :

```
#classe pour stocker les informations sur les antécédents
class antecedent():
    #initialisation de la class à partir des paramètres qui sont passés
    def __init__(self, text, hyperclass, clas, domain1, domain2, domain3, sentId, messageId):
        self.label = text
        self.HC = hyperclass #HC contient l'hyperclasse
        self.CC = clas # CC contient la classe
        self.D1 = domain1 # D1 contient le domaine1
        self.D2 = domain2 # D2 contient le domaine2
        self.D3 = domain3 # D3 contient le domaine3
        self.mId = int(messageId) # contient le numéro du message
        self.sId = int(sentId) # contient le numéro de phrase du message
        # dictionnaire qui contient les scores de saillence pour tous les antécédents. La clé est l'antécédent, la valeur est le score
        self.scoreDict = {}
    def computeScore(self, antecedentToCompare):
        total = 0 #initialisation du score à 0
        score_HC, score_CC, score_D1, score_D2,score_D3 = 0,0,0,0,0
        # distance en nombre de phrase entre les GNs
        sentDist = abs(self.sId - antecedentToCompare.sId)
        if (sentDist == 0):
            score_proxi = 1
        elif sentDist > 1 and sentDist <= 2:
            score_proxi = 0.8
        elif sentDist > 2 and sentDist <= 4:
            score_proxi = 0.7
        else:
            score_proxi = 0
        #facon simple de calculer le score :
        #si hyperclassA = hyperclassB alors score_HC = 1, sinon score_HC =0
        # on peut changer pour que cela soit plus souple que 0 ou 1
        if (self.HC == antecedentToCompare.HC):
            score_HC = 1
        if (self.CC == antecedentToCompare.CC):
            score_CC = 1
        if (self.D1 == antecedentToCompare.D1):
            score_D1 = 1
        if (self.D2 == antecedentToCompare.D2):
            score_D2 = 1
        if (self.D3 == antecedentToCompare.D3):
            score_D3 = 1
        total = float(score_HC + score_CC + score_D1 + score_D2 + score_D3 + score_proxi)/ 6
        total = round(total, 2)
        #total = score_HC + score_CC + scoreD1 + scoreD2 + scoreD3
        # ajouter le score de saillence au dictionnaire -- note vérifier les doublons dans les antécédents
        self.scoreDict[antecedentToCompare] = total
        return total

```

Le corpus pour la résolution des anaphores associatives (annoté avec Unitex) :



Le corpus pour la résolution des anaphores associatives (nettoyé) :

message
{S}Je viens de recevoir <groupe_nominal type='DEF' sous_type='CEN'><determinant type='CE'>cette</determinant>
{montre,N+H_APPAREIL+C_APPAREIL_SIGNALISATION+SC_+D1_TECHNIQUES+D2_VIE_QUOTIDIENNE+D3_}</groupe_nominal> mais celle-ci ne fonctionne pas, j'ai beau essayer de la régler mais rien ne se passe, <groupe_nominal type='DEF' sous_type='LEN'><determinant type='LE'>la</determinant>{trotteuse ,N+H_ORGANE+C_ORGANE_SIGNALISATION+SC_+D1_TECHNIQUES+D2_VIE_QUOTIDIENNE+D3_}
</groupe_nominal> ne bouge pas, et <groupe_nominal type='DEF' sous_type='LEN'><determinant type='LE'>l'</determinant>heure</groupe_nominal> reste figée.
message
{S}J'ai acheté <groupe_nominal type='NON_DEF' sous_type='DNUMN'><determinant type='DNUM'>une</determinant>{machine à expresso
,N+H_APPAREIL+C_APPAREIL_CUISSON+SC_CAFETIERE+D1_ALIMENTATION+D2_TRAVAUX_ET_EQUIPEMENTS_MENAGERS+_D3} mais <groupe_nominal type='DEF' sous_type='LEN'><determinant type='LE'>le</determinant>{support filtre à café ,N+H_DISPOSITIF+C_DISPOSITIF_MAINTIEN+SC_+D1_ALIMENTATION+D2_ TRAVAUX_ET_EQUIPEMENTS_MENAGERS+D3_} se bouche systématiquement lorsque l'on met <groupe_nominal type='NON_DEF' sous_type='DUN'><determinant type='DU'>du</determinant>{café moulu,N+H_ALIMENT+C_SC_+D1_ALIMENTATION+D2_+D3_} </groupe_nominal>.
message
{S}On m'a livré <groupe_nominal type='NON_DEF' sous_type='DNUMN'><determinant type='DNUM'>une</determinant>{théière
,N+H_CONTENANT+C_CONTENANT_BOISSON+SC_+D1_ALIMENTATION+D2_TRAVAUX_ET_EQUIPEMENTS_MENAGERS+D3_}</groupe_nominal> {Bodum ,N+H_MARQUE+C_+SC_+D_TRAVAUX_ET_EQUIPEMENTS_MENAGERS} il y a <groupe_nominal type='NON_DEF' sous_type='DNUMN'><determinant type='DNUM'>trois</determinant>jours</groupe_nominal> <groupe_nominal type='DEF' sous_type='LEN'><determinant type='LE'>la</determinant>{partie haute ,N+H_PARTIE_GENERIQUE+C_+SC_+D1_DIVERS+D2_+D3_}</groupe_nominal> était cassée lorsque j'ai déballé <groupe_nominal type='DEF' sous_type='LEN'><determinant type='LE'>le</determinant>{produit ,N+H_GENERIQUE+C_+SC_+D1_COMMERCE+D2_VIE_QUOTIDIENNE+D3_}</groupe_nominal>.
message
{S}Je vous ai déjà écrit pour vous dire que <groupe_nominal type='DEF' sous_type='POSSN'><determinant type='POSS'>ma</determinant>{veste
,N+H_VETEMENT+C_VETEMENT_HABILLAGE+SC_+D1_HABILLEMENT+D2_+D3_}</groupe_nominal>{celio ,N+H_MARQUE+C_+SC_+D_HABILLEMENT} présentait <groupe_nominal type='NON_DEF' sous_type='UNN'><determinant type='UN'>des</determinant> taches blanches</groupe_nominal>.{S} Au vu <groupe_nominal type='NON_DEF' sous_type='DUN'><determinant type='DU'>du</determinant>prix</groupe_nominal>de<groupe_nominal type='DEF' sous_type='LEN'><determinant type='LE'> la</determinant>{veste
,N+H_VETEMENT+C_VETEMENT_HABILLAGE+SC_+D1_HABILLEMENT+D2_+D3_}</groupe_nominal> je souhaitais la conserver et non la renvoyer.
message
{S}J'ai commandé <groupe_nominal type='NON_DEF' sous_type='DNUMN'><determinant type='DNUM'>trois</determinant>{valises ,N+H_CONTENANT+C_CONTENANT_TRANSPORT+SC_+D1_TRANSPORTS+D2_+D3_}
</groupe_nominal> Platinium.{S} Après seulement <groupe_nominal type='NON_DEF' sous_type='DNUMN'><determinant type='DNUM'>deux</determinant>utilisations</groupe_nominal> <groupe_nominal type='DEF' sous_type='LEN'><determinant type='LE'>les</determinant>{roues,N+H_DISPOSITIF+C_DISPOSITIF_TRANSPORT+SC+D1_TRANSPORTS+D2_+D3_}</groupe_nominal> sont cassées et <groupe_nominal type='DEF' sous_type='LEN'><determinant type='LE'>la</determinant>{fermeture ,N+H_DISPOSITIF+C_DISPOSITIF_BLOCAGE+SC_+D1_DIVERS+D2_+D3_ }</groupe_nominal> aussi.
message
{S}<groupe_nominal type='DEF' sous_type='LEN'><determinant type='LE'>La</determinant>{cafetière
,N+H_APPAREIL+C_APPAREIL_CUISSON+SC_+D1_ALIMENTATION+D2_TRAVAUX_ET_EQUIPEMENTS_MENAGERS+D3_}</groupe_nominal> fuit énormément lorsque on ajoute <groupe_nominal type='DEF' sous_type='LEN'><determinant type='LE'>l'</determinant>eau</groupe_nominal>.{S} <groupe_nominal type='DEF' sous_type='LEN'><determinant type='LE'>La</determinant>fuite</groupe_nominal> vient de dessous <groupe_nominal type='DEF' sous_type='LEN'><determinant type='LE'>la</determinant>{machine ,N+H_APPAREIL+C_APPAREIL_GENERIQUE+SC_+D1_DIVERS+D2_+D3_}</groupe_nominal>.
message
{S}<groupe_nominal type='DEF' sous_type='LEN'><determinant type='LE'>Le</determinant>{produit ,N+H_GENERIQUE+C_+SC_+D1_COMMERCE+D2_VIE_QUOTIDIENNE+D3_}</groupe_nominal> est arrivé avec <groupe_nominal type='NON_DEF' sous_type='UNN'><determinant type='UNN'>un</determinant>défaut de fabrication</groupe_nominal>: <groupe_nominal type='NON_DEF' sous_type='DNUMN'><determinant type='DNUM'>deux</determinant>{bords ,N+H_PARTIE_GENERIQUE+C_+SC_+D1_DIVERS+D2_+D3_}</groupe_nominal>de <groupe_nominal type='DEF' sous_type='LE'><determinant type=LE'>la</determinant>{poêle ,N+H_CONTENANT+C_CONTENANT_CUISSON+SC_+D1_ALIMENTATION+D2_TRAVAUX_ET_EQUIPEMENTS_MENAGERS+D3_}</groupe_nominal> sont tordus.

Le corpus pour la résolution des anaphores associatives (transformé au format xml) :

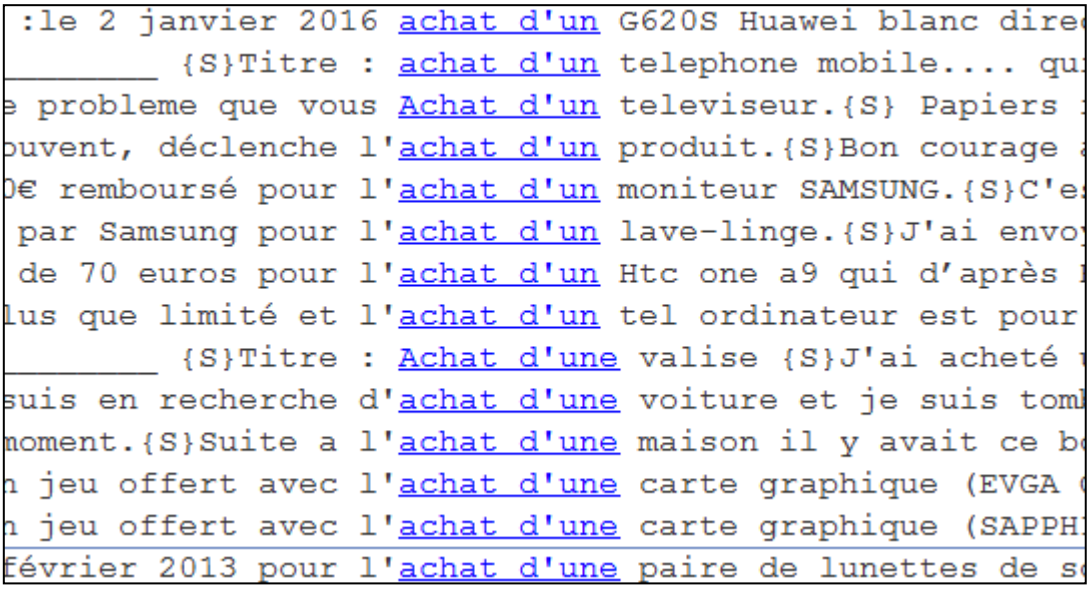
```
<?xml version='1.0' encoding='UTF-8' ?>
<document>
    <message id='1'>
        <phrase id='1'>
Je viens de recevoir <groupe_nominal type='DEF' sous_type='CEN'><determinant type='CE'>cette</determinant><nom hyperclasse='APPAREIL' classe='APPAREIL_SIGNALISATION' sous_classe=''  do
TECHNIQUES' domaine2='VIE_QUOTIDIENNE' domaine3=''>montre</nom></groupe_nominal> mais celle-ci ne fonctionne pas, j'ai beau essayer de la régler mais rien ne se passe, <groupe_nominal
DEF' sous_type='LEN'><determinant type='LE'>la</determinant><nom hyperclasse='ORGANE' classe='ORGANE_SIGNALISATION' sous_classe=''  domaine1='TECHNIQUES' domaine2='VIE_QUOTIDIENNE' dom
>trotteuse </nom></groupe_nominal> ne bouge pas, et <groupe_nominal type='DEF' sous_type='LEN'><determinant type='LE'>l'</determinant>heure</groupe_nominal> reste figée.
        </phrase>
    </message>
    <message id='2'>
        <phrase id='1'>
J'ai acheté <groupe_nominal type='NON_DEF' sous_type='DNUMN'><determinant type='DNUM'>une</determinant><nom hyperclasse='APPAREIL' classe='APPAREIL_CUISSON' sous_classe='CAFETIERE'  do
ALIMENTATION' domaine2='TRAVAUX_ET_EQUIPEMENTS_MENAGERS' domaine3=''>machine à expresso </nom></groupe_nominal> mais <groupe_nominal type='DEF' sous_type='LEN'><determinant type='LE'>l
determinant><nom hyperclasse='DISPOSITIF' classe='DISPOSITIF_MAINTIEN' sous_classe=''  domaine1='ALIMENTATION' domaine2=' TRAVAUX_ET_EQUIPEMENTS_MENAGERS' domaine3=''>support filtre à
nom></groupe_nominal> se bouche systématiquement lorsque l'on met <groupe_nominal type='NON_DEF' sous_type='DUN'><determinant type='DU'>du</determinant><nom hyperclasse='ALIMENT' class
sous_classe=''  domaine1='ALIMENTATION' domaine2='' domaine3=''>café moulu</nom> </groupe_nominal>.
        </phrase>
    </message>
    <message id='3'>
        <phrase id='1'>
On m'a livré <groupe_nominal type='NON_DEF' sous_type='DNUMN'><determinant type='DNUM'>une</determinant><nom hyperclasse='CONTENANT' classe='CONTENANT_BOISSON' sous_classe=''  domaine1
ALIMENTATION' domaine2='TRAVAUX_ET_EQUIPEMENTS_MENAGERS' domaine3=''>théière </nom></groupe_nominal> <nom hyperclasse='MARQUE' classe='' sous_classe=''  domaine1='
TRAVAUX_ET_EQUIPEMENTS_MENAGERS}' domaine2='' domaine3=''>Bodum </nom> il y a <groupe_nominal type='NON_DEF' sous_type='DNUMN'><determinant type='DNUM'>trois</determinant>jours</groupe
> <groupe_nominal type='DEF' sous_type='LEN'><determinant type='LE'>la</determinant><nom hyperclasse='PARTIE_GENERIQUE' classe='' sous_classe=''  domaine1='DIVERS' domaine2='' domaine3
partie haute </nom></groupe_nominal> était cassée lorsque j'ai déballé <groupe_nominal type='DEF' sous_type='LEN'><determinant type='LE'>le</determinant><nom hyperclasse='GENERIQUE' cl
sous_classe=''  domaine1='COMMERCE' domaine2='VIE_QUOTIDIENNE' domaine3=''>produit </nom></groupe_nominal>.
        </phrase>
    </message>
    <message id='4'>
        <phrase id='1'>
Je vous ai déjà écrit pour vous dire que <groupe_nominal type='DEF' sous_type='POSSN'><determinant type='POSS'>ma</determinant><nom hyperclasse='VETEMENT' classe='VETEMENT_HABILLAGE'
sous_classe=''  domaine1='HABILLEMENT' domaine2='' domaine3=''>veste </nom></groupe_nominal><nom hyperclasse='MARQUE' classe='' sous_classe=''  domaine1='HABILLEMENT}' domaine2='' doma
celio </nom> présentait <groupe_nominal type='NON_DEF' sous_type='UNN'><determinant type='UN'>des</determinant> taches blanches</groupe_nominal>.
        </phrase>
        <phrase id='2'>
Au vu <groupe_nominal type='NON_DEF' sous_type='DUN'><determinant type='DU'>du</determinant>prix</groupe_nominal>de<groupe_nominal type='DEF' sous_type='LEN'><determinant type='LE'> la
determinant><nom hyperclasse='VETEMENT' classe='VETEMENT_HABILLAGE' sous_classe=''  domaine1='HABILLEMENT' domaine2='' domaine3=''>veste </nom></groupe_nominal> je souhaitais la conser
non la renvoyer.
        </phrase>
    </message>
    <message id='5'>
        <phrase id='1'>
J'ai commandé <groupe_nominal type='NON_DEF' sous_type='DNUMN'><determinant type='DNUM'>trois</determinant><nom hyperclasse='CONTENANT' classe='CONTENANT_TRANSPORT' sous_classe=''  dom
```

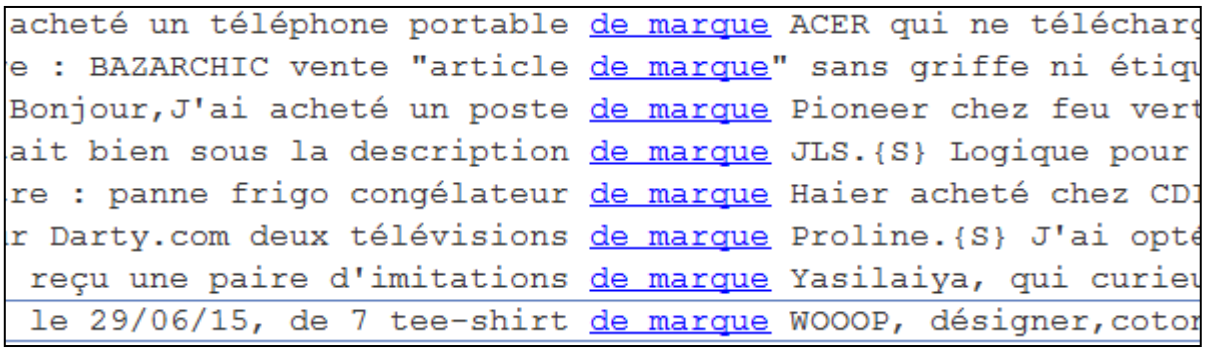
Résultat de la résolution des anaphores associatives

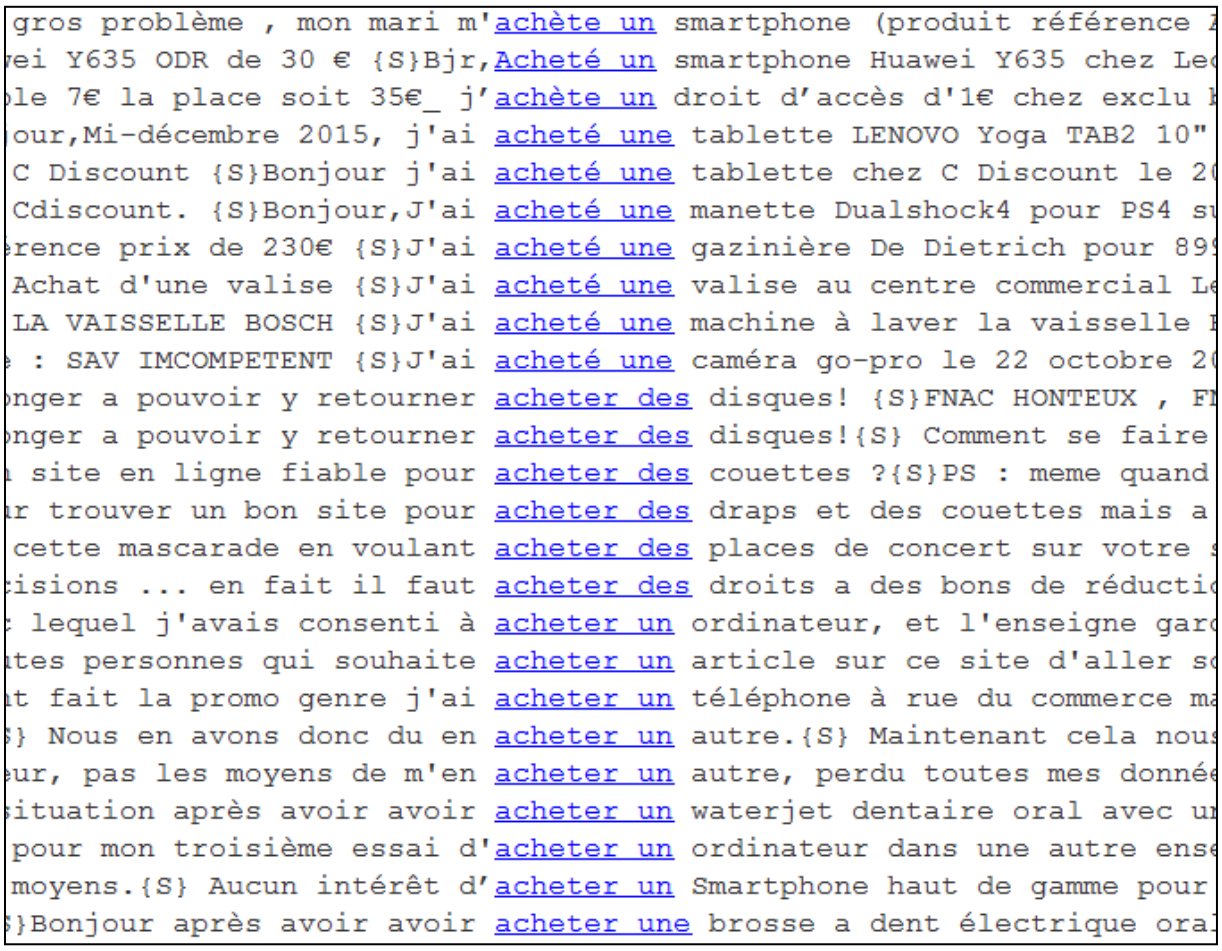
```
trotteuse : montre#0.67
machine ├á expresso :
support filtre ├á caf├® : machine ├á expresso#0.33
caf├® moulu : support filtre ├á caf├®#0.33 | machine ├á expresso#0.33
th├®i├¿re :
Bodum : th├®i├¿re#0.17
partie haute : Bodum#0.67 | th├®i├¿re#0.17
produit : partie haute#0.33 | Bodum#0.33 | th├®i├¿re#0.17
veste :
celio : veste#0.5
veste : veste#0.83 | celio#0.33
valises :
roues : valises#0.5
fermeture : roues#0.67 | valises#0.33
cafeti├¿re :
machine : cafeti├¿re#0.17
produit :
bords : produit#0.33
po├¬le : produit#0.17 | bords#0.17
bottines :
semelles : bottines#0.5
chaussures : bottines#0.83 | semelles#0.67
semelles : semelles#1.0 | chaussures#0.67 | bottines#0.5
montre :
pile : montre#0.5
toaster :
cot├® : toaster#0.17
bouton : cot├®#0.17 | toaster#0.17
presse-l├®gumes :
couvercle : presse-l├®gumes#0.0
morceaux : couvercle#0.17 | presse-l├®gumes#0.0
bo├«te : morceaux#0.17 | couvercle#0.17 | presse-l├®gumes#0.0
CHANTELLE :
soutien-gorge : CHANTELLE#0.33
├®tiquette : soutien-gorge#0.17 | CHANTELLE#0.17
culotte : soutien-gorge#0.83 | CHANTELLE#0.47 | ├®tiquette#0.0
robot multifonctions :
Philips : robot multifonctions#0.17
```

3. L'ENRICHISSEMENT DES DICTIONNAIRES UNITEX

La recherche des noms d'artefact avec des patrons syntaxiques :







4. LES SCRIPTS

TRAITEMENT DES ANAPHORES INFIDELES – AFFICHAGE AU FORMAT HTML

```
#!/usr/bin/python2

# -*- coding: utf-8 -*-


import os,sys,re,getopt

try:
```

```
    import xml.etree.cElementTree as ET
except ImportError:
    import xml.etree.ElementTree as ET


#classe pour stocker les informations sur les antécédents
class antecedent():
    #initialisation de la class à partir des paramètres qui sont passés
    def __init__(self, personTail, personId, text, clas, gen, nb, sentId, articleID):
        self.label = text
        self.suite = personTail
        self.id = personId
        self.CC = clas # CC contient la class
        self.G = gen # G contient le genre
        self.N = nb # N contient le nombre
        self.aId = int(articleID) # contient le numéro de l'article
        self.sId = int(sentId) # contient le numéro de phrase
        self.scoreDict = {} # dictionnaire qui contient les scores de saillence pour tous les antécédents.
        # La clé est l'antécédent, la valeur est le score


    def computeScore(self, antecedentToCompare):
```

```
total = 0 #initialisation du score à 0

score_CC, score_G, score_N = 0,0,0

# distance en nombre de phrase entre les persons

sentDist = abs(self.sId - antecedentToCompare.sId)

if (sentDist == 0):

   score_proxi = 1

elif sentDist > 1 and sentDist <= 2:

   score_proxi = 0.8

elif sentDist > 2 and sentDist <= 4:

   score_proxi = 0.7

else:

   score_proxi = 0

# calculer le score : si classA = classB alors score_CC = 1, sinon score_CC = 0

if (self.CC == antecedentToCompare.CC):

   score_CC = 1

if (self.G == antecedentToCompare.G):

   score_G = 1

if (self.N == antecedentToCompare.N):

   score_N = 1

total = float(score_CC + score_G + score_N + score_proxi)/ 4

total = round(total, 2)
```

```
        # ajouter le score de saillance au dictionnaire -- note vérifier les doublons dans les antécédents
        self.scoreDict[antecedentToCompare] = total
        return total


# fonction qui lit le fichier xml et renvoie une liste des persons dans l'ordre du document
def readXml(inFile):
    outLst = [] # liste qui contient les persons d'un document
    tree = ET.ElementTree(file=inFile)
    # expression xpath pour parcourir chaque article
    articleLst = tree.findall('//article')
    for article in articleLst:
        i = 1
        #identifiant de l'article
        articleID = article.get("id")
        for phrase in list(article):
            Lst = [] # liste qui contient les phrase.text + les persons
            #identifiant de la phrase dans l'article
            phraseId = phrase.get("id")
```

```
if phrase.text :
    phraseText = phrase.text # vérifier si le text n'est pas vide, prendre le texte de <phrase>
else :
    phraseText = "" # si le texte est vide, il aura le NoneType et non pas du type str
#parcourir chaque balise <person> dans la phrase courante
personLst = phrase.findall(".//person")
# parcours toutes les balises <person> du fichier et construit un object antecedent pour chaque person
for person in personLst:
    person_id = i
    i = i + 1
    personLst[0].text = phraseText + "<person style=\"color: red;\">" + person.text + "</person>"
    if person.tail :
        person_Tail = person.tail # vérifier si le text après <person> n'est pas vide
    else :
        person_Tail = ""
    person_attrib = person.attrib # person_attrib est un dictionnaire qui contient toutes les valeurs : class, g, n
    #construire un object antecendent à partir du tag
```

```
            obj_antecendent = antecedent(person_Tail, person_id, person.text, person_attrib['class'], \
                                person_attrib['g'], person_attrib['n'], phraseId, articleID)
            outLst.append(obj_antecendent) #ajoute l'antecedent à la liste
    return outLst


# classe principale
def main():
    #lire le fichier xml et stocker les persons dans une liste
    #la liste les ajoute dans leurs ordres d'apparition dans le document
    person_lst = readXml("FD_ok.xml") # contient la liste des persons du document. Précisément la liste des instances de la classe antecedent.
    # pour chaque person de la liste on cherche son antécédent.
    # On commence par la fin de la liste pour comparer l'élément suivant avec l'élément précédent
    # on prends le dernier élément de la liste et calcul le score de saillance avec tous les éléments précédents
    inv_personLst = list(reversed(person_lst)) # inverser la liste des persons pour itérer à partir du dernier élément
    # (le dernier élément devient le premier).
    for i in range(0,len(person_lst)): #itération à partir du dernier élément
        current_person = inv_personLst[i]
```

```
    for j in range(i+1, len(person_lst)):
        prev_person = inv_personLst[j]
        # ne pas calculer le score lorsque les person dans la limite
        if current_person.aId != prev_person.aId:
            continue
        # on vérifie que la distance entre l'antécédent et le person est inférieure à la limite cutOff(2)
        if abs(current_person.sId - prev_person.sId) > 2 :
            continue
        # calculer le score entre le person courant et l'antécédent
        current_person.computeScore(prev_person)
    #pour chaque person affiche le texte du person et les antécédents avec leurs scores
    fichier = open("resultat_ok.html", "w")
    for person in person_lst:
        outLine = "<p><person>" + person.label.strip() + "</person><sub style=\"color: red;\"> " + \
                str(person.id) + "</sub><sup style=\"color: blue;\"> " + " | ".join([("%i(%s)")%(k.id,v) \
                for (k,v) in sorted(person.scoreDict.iteritems(), key=lambda x:(-1)*x[1]) if v > 0.75 ]) + "</sup>" + person.suite
        fichier.write("<?xml version=\"1.0\" encoding=\"UTF-8\"?>")
        fichier.write(outLine.encode("utf-8"))
```

```
        fichier.write("\n")
    fichier.close()


main()
```

TRAITEMENT DES ANAPHORES INFIDELES – AFFICHAGE AU FORMAT TABLEAU

```
#!/usr/bin/python2
# -*- coding: utf-8 -*-

#./searchAntecedent.py -n 2 test_input.xml
import os,sys,re,getopt

try:
    import xml.etree.cElementTree as ET
except ImportError:
    import xml.etree.ElementTree as ET

usage = "usage: cherchePerson.py -n antecedentCutOff xmlFile.xml\n"
xmlFile = ""
```

```
cutOff = 100 #par défault il n'y a pas de limite pour chercher l'antécédent


#parsing de la ligne de paramètre
try:
   opts, args = getopt.getopt(sys.argv[1:],"n:",["-n"])
except getopt.GetoptError, err:
   # print help information and exit:
   print str(err) # will print something like "option -a not recognized"
   sys.stderr.write(usage) #afficher le message d'erreur
   sys.exit() #quitter le programme


for opt, arg in opts:
   if opt in ("-n"):
      cutOff = int(arg)
   else:
      sys.stderr.write("Error: unrecognized option '%s'\n"%opt)
      exit()
for arg in args:
   if xmlFile=="":
      xmlFile=arg
   else:
```

```
        sys.stderr.write(usage) #afficher le message d'erreur
        sys.exit() #quitter le programme
        exit()

if xmlFile == "":
    sys.stderr.write(usage) #afficher le message d'erreur
    sys.exit() #quitter le programme
#==========



#classe pour stocker les informations sur les antécédents
class antecedent():
    #initialisation de la class à partir des paramètres qui sont passés
    def __init__(self, text, clas, gen, nb, sentId, articleID):
        self.label = text
        self.CC = clas # CC contient la class
        self.G = gen # G contient le genre
        self.N = nb # N contient le nombre
        self.aId = int(articleID) # contient le numéro de l'article
        self.sId = int(sentId) # contient le numéro de phrase
```

```
    # dictionnaire qui contient les scores de saillence pour tous les antécédents. La clé est l'antécédent, la valeur est le score
    self.scoreDict = {}
  def computeScore(self, antecedentToCompare):
    total = 0 #initialisation du score à 0
    score_CC, score_G, score_N = 0,0,0
    # distance en nombre de phrase entre les persons
    sentDist = abs(self.sId - antecedentToCompare.sId)
    if (sentDist == 0):
      score_proxi = 1
    elif sentDist > 1 and sentDist <= 2:
      score_proxi = 0.8
    elif sentDist > 2 and sentDist <= 4:
      score_proxi = 0.7
    else:
      score_proxi = 0
    #facon simple de calculer le score :
    #si classA = classB alors score_CC = 1, sinon score_CC = 0
    # on peut changer pour que cela soit plus souple que 0 ou 1
    if (self.CC == antecedentToCompare.CC):
      score_CC = 1
```

```
        if (self.G == antecedentToCompare.G):
            score_G = 1
        if (self.N == antecedentToCompare.N):
            score_N = 1
        total = float(score_CC + score_G + score_N + score_proxi)/ 4
        total = round(total, 2)
        # ajouter le score de saillance au dictionnaire -- note vérifier les doublons dans les antécédents
        self.scoreDict[antecedentToCompare] = total
        return total


# fonction qui lit le fichier xml et renvoie une liste des persons dans l'ordre du document
def readXml(inFile):
    outLst = [] # liste qui contient les persons d'un document
    tree = ET.ElementTree(file=inFile)
    #expression xpath pour parcourir chaque article
    articleLst = tree.findall('//article')
    for article in articleLst:
        #identifiant de l'article
        articleID = article.get("id")
```

```
        for phrase in list(article):
            #identifiant de la phrase dans l'article
            phraseId = phrase.get("id")
            #parcourir chaque balise person dans la phrase courante
            personLst = phrase.findall(".//person")
            # parcours toutes les balises <person> du fichier et construit un object antecedent pour chaque person
            for person in personLst:
                person_attrib = person.attrib # person_attrib est un dictionnaire qui contient toutes les valeurs
                #construire un object antecendent à partir du tag
                obj_antecendent = antecedent(person.text, person_attrib['class'], person_attrib['g'], person_attrib['n'], phraseId, articleID)
                #ajoute l'antecedent à la liste
                outLst.append(obj_antecendent)
    return outLst


# classe principale
def main():
    #lire le fichier xml et stocker les persons dans une liste
    #la liste les ajoute dans leurs ordres d'apparition dans le document
```

person_lst = readXml(xmlFile) # contient la liste des persons du document. Précisément la liste des instances de la classe antecedent.

pour chaque person de la liste on cherche son antécédent. On commence par la fin de la liste:

#on prends le dernier élément de la liste et on calcul le score de saillence avec tous les éléments précédents

inv_personLst = list(reversed(person_lst)) # inverser la liste des persons pour itérer à partir du dernier élément (le dernier élément devient le premier).

for i in range(0,len(person_lst)): #itération à partir du dernier élément

current_person = inv_personLst[i]

for j in range(i+1, len(person_lst)):

prev_person = inv_personLst[j]

#ne pas calculer le score lorsque les person dans la limite

if current_person.aId != prev_person.aId:

continue

#on vérifie que la distance entre l'antécédent et le person est inférieure à la limite cutOff

if abs(current_person.sId - prev_person.sId) > cutOff :

continue

#calculer le score entre le person courant et l'antécédent

current_person.computeScore(prev_person)

```
    #pour chaque person affiche le texte du person et les antécédents avec leurs scores
    fichier = open("resultat_FD.txt", "w")
    for person in person_lst:
        outLine = person.label.strip() + " : " + " | ".join([("%s = %s")%(k.label.strip(),v) for
(k,v) in sorted(person.scoreDict.iteritems(),
            key=lambda x:(-1)*x[1])])
        fichier.write(outLine.encode("utf-8"))
        fichier.write("\n")
    fichier.close()


main()
```

LE SCRIPT DE TRAITEMENT DES ANAPHORES ASSOCIATIVES

```
#!/usr/bin/python
# -*- coding: utf-8 -*-
import os,sys,re,getopt
try:
    import xml.etree.cElementTree as ET
except ImportError:
    import xml.etree.ElementTree as ET
```

```
usage = "usage: searchAntecedent.py -n antecedentCutOff xmlFile.xml\n"
xmlFile = ""
cutOff = 100 #par défault il n'y a pas de limite pour chercher l'antécédent

#parsing de la ligne de paramètre
try:
   opts, args = getopt.getopt(sys.argv[1:],"n:",["-n"])
except getopt.GetoptError, err:
   # print help information and exit:
   print str(err) # will print something like "option -a not recognized"
   sys.stderr.write(usage) #afficher le message d'erreur
   sys.exit() #quitter le programme

for opt, arg in opts:
   if opt in ("-n"):
      cutOff = int(arg)
   else:
      sys.stderr.write("Error: unrecognized option '%s'\n"%opt)
      exit()
for arg in args:
   if xmlFile=="":
```

```
        xmlFile=arg
    else:
        sys.stderr.write(usage) #afficher le message d'erreur
        sys.exit() #quitter le programme
        exit()

if xmlFile == "":
    sys.stderr.write(usage) #afficher le message d'erreur
    sys.exit() #quitter le programme
#===========



#classe pour stocker les informations sur les antécédents
class antecedent():
    #initialisation de la class à partir des paramètres qui sont passés
    def __init__(self, text, hyperclass, clas, domain1, domain2, domain3, sentId, messageId):
        self.label = text
        self.HC = hyperclass #HC contient l'hyperclasse
        self.CC = clas # CC contient la classe
        self.D1 = domain1 # D1 contient le domaine1
```

```
        self.D2 = domain2 # D2 contient le domaine2
        self.D3 = domain3 # D3 contient le domaine3
        self.mId = int(messageId) # contient le numéro du message
        self.sId = int(sentId) # contient le numéro de phrase du message
        # dictionnaire qui contient les scores de saillence pour tous les antécédents. La clé est l'antécédent, la valeur est le score
        self.scoreDict = {}
    def computeScore(self, antecedentToCompare):
        total = 0 #initialisation du score à 0
        score_HC, score_CC, score_D1, score_D2,score_D3 = 0,0,0,0,0
        # distance en nombre de phrase entre les GNs
        sentDist = abs(self.sId - antecedentToCompare.sId)
        if (sentDist == 0):
            score_proxi = 1
        elif sentDist > 1 and sentDist <= 2:
            score_proxi = 0.8
        elif sentDist > 2 and sentDist <= 4:
            score_proxi = 0.7
        else:
            score_proxi = 0
        #facon simple de calculer le score :
```

```
#si hyperclassA = hyperclassB alors score_HC = 1, sinon score_HC =0

# on peut changer pour que cela soit plus souple que 0 ou 1

if (self.HC == antecedentToCompare.HC):

   score_HC = 1

if (self.CC == antecedentToCompare.CC):

   score_CC = 1

if (self.D1 == antecedentToCompare.D1):

   score_D1 = 1

if (self.D2 == antecedentToCompare.D2):

   score_D2 = 1

if (self.D3 == antecedentToCompare.D3):

   score_D3 = 1

total = float(score_HC + score_CC + score_D1 + score_D2 + score_D3 + score_proxi)/ 6

total = round(total, 2)

#total = score_HC + score_CC + scoreD1 + scoreD2 + scoreD3

# ajouter le score de saillence au dictionnaire -- note vérifier les doublons dans les antécédents

self.scoreDict[antecedentToCompare] = total

return total
```

```
# fonction qui lit le fichier xml et renvoie une liste des GNs dans l'ordre du document
def readXml(inFile):
    outLst = [] # liste qui contient les GNs d'un document
    tree = ET.ElementTree(file=inFile)
    #expression xpath pour parcourir chaque message
    messageLst = tree.findall('//message')
    for message in messageLst:
        #identifiant du message
        messageId = message.get("id")
        for phrase in list(message):
            #identifiant de la phrase dans le message
            phraseId = phrase.get("id")
            #parcourir chaque balise nom dans la phrase courante
            nameLst = phrase.findall(".//nom")
            # parcours toutes las balises <nom> du fichier et construit un object antecedent pour chaque GN
            for groupe_nom in nameLst:
                nom_attrib = groupe_nom.attrib # nom attrib est un dictionnaire qui contient toutes les valeurs
                #construire un object antecendent à partir du tag
```

```
            obj_antecendent = antecedent(groupe_nom.text, nom_attrib['hyperclasse'], nom_attrib['classe'], nom_attrib['domaine1'], nom_attrib['domaine2'], nom_attrib['domaine2'], phraseId, messageId)

            #ajoute l'antecedent à la liste

            outLst.append(obj_antecendent)

    return outLst


# classe principale

def main():

    #lire le fichier xml et stocker les GNs dans une liste

    #la liste les ajoute dans leurs ordres d'apparition dans le document

    gn_lst = readXml(xmlFile) # contient la liste des GNs du document. Précisément la liste des instances de la classe antecedent.


    # pour chaque GN de la liste on cherche son antécédent. On commence par la fin de la liste: on prends le dernier élément de la liste et on calcul le score de saillence avec tous les éléments précédents

    inv_gnLst = list(reversed(gn_lst)) # inverser la liste des GNs pour itérer à partir du dernier élément (le dernier élément devient le premier).

    for i in range(0,len(gn_lst)): #itération à partir du dernier élément

        current_gn = inv_gnLst[i]

        for j in range(i+1, len(gn_lst)):
```

```
        prev_gn = inv_gnLst[j]
        #ne pas calculer le score lorsque les GNs dont la limite
        if current_gn.mId != prev_gn.mId:
            continue
        #on vérifie que la distance entre l'antécédent et le GN est inférieure à la limite cutOff
        if abs(current_gn.sId - prev_gn.sId) > cutOff :
            continue
        #calculer le score entre le gn courant et l'antécédent
        current_gn.computeScore(prev_gn)
    #pour chaque GN affiche le nom du GN et les antécédents avec leurs scores
    for gn in gn_lst:
        outLine = gn.label.strip() + " : " + " | ".join([("%s#%s")%(k.label.strip(),v) for (k,v) in sorted(gn.scoreDict.iteritems(), key=lambda x:(-1)*x[1]) ])
        print outLine.encode("utf-8")


main()
```

SCRIPT D'ASPIRATION DU CORPUS DE FAIT-DIVERS

```
# -*- coding: utf-8 -*-
```

```
import re

import urllib2

from bs4 import BeautifulSoup


# ce programme donnera en sortie un fichier url.txt qui contient les url obtenus lorsqu'on

# aspire le site : http://www.sudinfo.be/12/actualite/faits-divers


# générer la liste des noms de pages de 0 à 3

def generation_url() :

        liste_page = []

        for i in range (0, 15) :

                pages = "http://www.sudinfo.be/12/actualite/faits-divers?page=%i" % i

                liste_page.append(pages)

        return liste_page


# input : nom d'une page, output : le contenu html

def print_content(page_url) :

        hdr = {'User-Agent':'Mozilla/5.0'}

        req = urllib2.Request(page_url,headers=hdr)
```

```
        response = urllib2.urlopen(req)
        html = response.read()
        return html


# input : 1 url, output : liste url
def out_url_all(url) :
        text_input = print_content(url)
        url_out = []
        for match in re.finditer (r"\<h2\> \<a href=\"\/((.)*)\"\>", text_input):
                url_list = "http://www.sudinfo.be/%s" % match.group(1)
                url_out.append(url_list)
        return url_out


# input : rien, output : liste all url
def all_url_list() :
        list_url_in = generation_url()
        list_url_out = []
        for url in list_url_in :
                list_url_out.append(out_url_all(url))
        return list_url_out
```

```
fileout = open("corpus_FD.txt", "w")


def parse_url():
        liste_url = all_url_list()
        i = 0
        for l_url in liste_url :
                for url in l_url :
                        print "%d - %s" % (i, url)
                        i = i + 1
                        hdr = {'User-Agent':'Mozilla/5.0'}
                        req = urllib2.Request(url,headers=hdr)
                        try:
                                resp = urllib2.urlopen(req)
                                html = resp.read()
                        except urllib2.HTTPError, error:
                                html = error.read()
                        soup = BeautifulSoup(html)
                        for t in soup.find_all("h1") :
                                data = t.get_text().encode("utf-8")
                                fileout.write("-------------------------\nTitre : %s\n" % data)
                        for st in soup.find_all("div", {"class": "chapeau"}) :
```

```
                data = st.get_text().encode("utf-8")
                fileout.write("Soustitre : %s" % data)
            for p in soup.find_all("div", {"id": "body_text"}) :
                data = p.get_text().encode("utf-8")
                fileout.write("%s\n" % data)


parse_url()
fileout.close()
```

SCRIPT POUR TRANSFORMER UN CORPUS BRUT AU FORMAT XML

```
#!/usr/bin/python
# -*- coding: utf-8 -*-

from xml.etree.ElementTree import ElementTree
from xml.etree.ElementTree import Element
import xml.etree.ElementTree as etree
import codecs

# lire fichier d'entrée, supprimer le {S} avant Titre
file_in = codecs.open("FD_cleaned.html", "r", "utf-8")
```

```
text = file_in.read()

messages = text.split("_____")


# préparer le contenu XML

root=Element('document') # créer tag document


i = 1

for message in messages :

        article=Element('article') # tao tag article

        root.append(article) # tag article dans tag document

        article.set('id', str(i)) # creer attributs aux articles

        i = i + 1

        phrases = message.split("{S}")

        j = 0

        for phrase in phrases :

                title_mark = "Titre : "

                subtitle_mark = "Soustitre : "

                if title_mark in phrase :

                        title=Element('titre') # créer tag titre

                        article.append(title) # tag titre dans tag article

                        title.text = phrase
```

```
            elif subtitle_mark in phrase :
                subtitle=Element('soustitre') # créer tag titre
                article.append(subtitle) # tag titre dans tag article
                subtitle.text = phrase
            else :
                line=Element('phrase') # créer tag phrase
                article.append(line) # tag phrase dans tag article
                line.text = phrase
                line.set('id', str(j)) # creer attributs à la phrase
                j = j + 1


# creation d'un fichier XML de sortie
tree=ElementTree(root)
file_out = codecs.open("FD_cleaned.xml","w", "utf-8")
file_out.write("<?xml version='1.0' encoding='UTF-8' ?>")
tree.write(file_out)
```