Repérage automatique de citations dans des documents journalistiques

Mémoire de Master 2 Recherche SAD

Fabien POULARD

encadré par Nicolas Hernandez et Annie Tartier

Laboratoire d'Informatique de Nantes Atlantique
2, rue de la Houssinière
B.P. 92208
F-44322 NANTES CEDEX 3

RAPPORT DE STAGE DE MASTER RECHERCHE

Octobre 2007

Fabien POULARD (encadré par Nicolas Hernandez et Annie Tartier)

Repérage automatique de citations dans des documents journalistiques

Mémoire de Master 2 Recherche SAD

Repérage automatique de citations dans des documents journalistiques

Mémoire de Master 2 Recherche SAD

Fabien POULARD (encadré par Nicolas Hernandez et Annie Tartier)

fabien.poulard@etu.univ-nantes.fr

Remerciements

Il y a plein de personnes à remercier ... je pourrais les passer une par une en faisant une blague trop cool sur elles, mais à dix minutes de rendre le rappport ça ne va pas être facile.

Je remercie tous les gens du bureau 212 car ils sont vraiement géniaux ! Je les remercie notamment pour avoir sû m'extraire plusieurs fois de plus de 20 minutes de travail d'affilé.

Je remercie Guillaume, aka Bob, pour m'avoir fait ouvrir les yeux sur les subtilités de LaTeXet ainsi permis de réaliser un rapport qui ressemble un peu mieux à ce que je voulais.

Je remercie mes encadrants pour m'avoir accompagner lors de ce voyage initiatique.

Je remercie mon frère d'être toujours en vie et je lui dédie ce mémoire, juste comme ça, parce que je l'aime.

Introduction

Toute nouvelle découverte se fonde sur les briques d'une connaissance auparavant acquise. Ptolémée Ier l'avait bien compris lorsqu'il fit construire la bibliothèque d'Alexandrie et y accumula des manuscrits de tout le monde antique. Cette bibliothèque — compilant environ 700 000 ouvrages à son apogée — attira les plus grands savants de l'époque : Zénodote, Euclide, Archimède... Les ouvrages au sein de cette bibliothèque s'empilèrent sans trop s'attacher à leurs origines et notamment à leurs auteurs. Ainsi, des scribes recopiaient les manuscrits présents à bord des bateaux entrant au port d'Alexandrie.

Les textes représentent aujourd'hui encore la manière la plus aisée de diffuser de l'information. Les scribes ont disparus, laissant place aux photocopieurs et à la numérisation, permettant ainsi une copie et une diffusion à une échelle bien plus importante. Le contrôle de la paternité des écrits devient de plus en plus difficile dans notre société dite "de l'information" et il est nécessaire de mettre au point des outils permettant de profiter de l'apparition de la numérisation.

Le but de ce stage de Master 2 Systèmes d'Aide à la Décision (SAD) est de mettre en place des techniques permettant de détecter de manière automatisée les citations au sein d'écrits journalistiques. Ce travail s'inscrit dans le projet ANR Piithie (Plagiats et Impacts de l'Information Textuelle recHerchée dans un contexte InterlinguE) dont l'objectif est de permettre la mise en place d'outils facilitant la prise de décisions sur le statut de plagiat des textes.

Sans prendre de position sur ce que l'on considère comme reprise textuelle respectueuse ou non de l'auteur original, nous tentons de caractériser le concept de citations afin de faciliter la prise de décision quant à cette reprise. Le concept même de citation est compris différemment selon qu'il s'agisse d'une citation d'un article scientifique, ou bien une citation au sens littéraire du terme. Dans le premier cas, il s'agit de se fonder sur des travaux antérieurs sans citer précisemment un passage des dits travaux, alors que dans le deuxième cas, un extrait du texte original est repris et intégré. Nous nous attacherons uniquement à la citation littéraire, cette dernière reflétant plus fidèlement le domaine sur lequel porte le stage, à savoir les textes journalistiques. Toute référence au terme citation dans la suite du rapport fera donc référence à la citation littéraire.

La détection de citation ne se limite pas au repérage, mais pose également le problème de définir les bornes d'une citation qui ne sont pas forcément explicitées par la typographie. De plus, la citation ne se limite pas à l'extrait de texte recopié verbatim, mais également aux variations de ce texte imposées par son intégration au texte englobant. Enfin, la citation s'accompagne souvent d'une référence à l'auteur original, le repérage de cette référence pouvant faire appel à un analyseur d'entité nommées.

Le premier chapitre du rapport introduit le concept de discours rapporté, puis de citation. Nous mettons en évidence dans cette partie, les spécificités de la citation par rapport au discours rapporté,

après avoir présenté les différentes tentatives de définition du concept de citation. À la fin de cette partie, nous affûtons notre approche en nous intéressant au cas des citations présentes au sein des textes journalistiques.

Le deuxième chapitre du rapport informe sur la manière dont a été constitué notre corpus ainsi que les premières informations concernant les formes que nous avons pu en retirer. Nous abordons également les problématiques qui nous ont mené à choisir une annotation sous forme de “segments citationnels”.

Le troisième chapitre présente notre approche méthodologique dans la mise en place de l’algorithme de détection de citations. Nous y présentons notamment notre vision de la structure des segments citationnels journalistiques, et nous discutons les marques que nous avons considérées d’intérêt pour l’apprentissage automatique dédié à chacune des composantes de cette structure.

Le quatrième chapitre traite dans un premier temps de l’architecture et de l’implémentation de la chaîne de traitement réalisant l’algorithme discuté dans la partie précédente. Nous analysons ensuite les résultats de l’apprentissage supervisé pour nos différents composants, puis d’une manière plus globale, les résultats concernant notre chaîne de traitement.

Finalement, la conclusion offre des ouvertures sur l’amélioration du système de détection proposé.

Chapitre 1

La notion de citation dans la littérature

Qu'elle soit littéraire ou scientifique, de texte ou d'auteur, la citation semble prendre bien des formes. Afin de mieux cadrer le sujet de notre travail, nous tentons dans ce chapitre de définir au mieux cet objet linguistique qu'est la citation. À l'aide de définitions proposées dans des travaux antérieurs tout d'abord, nous tentons d'isoler les éléments communs à toutes les définitions de citations, que celles-ci se veuillent littéraires, linguistiques ou bien opérationnelles. Nous tâchons ensuite d'identifier les différentes styles qui permettent aux utilisateurs de citations de les intégrer à leurs écrits. Nous abordons alors les styles littéraires direct, indirect, indirect libre ou narrativisés, mais également des styles plus spécifiques à la citation journalistique. Finalement, nous essayons de faire le point sur les différentes techniques de détection automatique qui ont été proposées afin de les confronter à notre besoin.

1.1 D'une définition littéraire à une définition opérationnelle

On peut considérer la citation du point de vue de sa réalisation linguistique, cependant elle peut être tellement diffuse au sein du discours qu'elle n'en est "plus palpable". Nous préférons considérer la citation comme une entité avant tout conceptuelle.

Afin d'apréhender au mieux ce concept, il est nécessaire de mettre en place une certaine terminologie fixant au préalable les caractéristiques des autres concepts et réalisations linguistiques auxquelles nous faisons appel.

1.1.1 Définitions proposées dans la littérature

La citation est, selon notre intuition, une notion mieux définie que le discours rapporté. Toutefois, toute citation implique une mise en oeuvre du discours rapporté. Nous tâchons dans un premier temps de définir et encadrer la notion de citation pour pouvoir ensuite définir les différences avec le discours rapporté.

Terminologie

Afin de mener notre discussion, nous posons les termes suivants. Cette terminologie est partiellement tirée de plusieurs travaux antérieurs, notamment [Mourad & Desclès, 2002] et [Giguet & Lucas, 2004].

Une **citation** prend place au sein d'un **texte englobant**, i.e. tout ce qui n'est pas lié à la citation au sein du texte et qui encadre donc cette dernière. La combinaison du texte englobant et de la citation forme le **texte complet** : article, oeuvre littéraire, ... Il est dans un premier temps important de remarquer que la citation est partie prenante du texte, elle y est attachée et sans elle le texte serait différent. Ainsi, la citation prend part au message à destination du lecteur.

La personne qui rédige le texte englobant et y intègre la citation est l'**auteur**. Comme l'énonce [Mourad & Desclès, 2001], le simple fait d'inclure une citation démontre un engagement de l'auteur sur ce qu'il rapporte. Il n'y a donc pas de citation neutre, la sélection même du segment de discours à rapporter dans son texte est un acte d'engagement. La personne — ou l'entité (e.g. "le congrès", la vox popula, ...) — qui avant l'auteur avait transmis le segment de discours que l'auteur rapporte est la **source**. L'auteur fait souvent référence à la source à l'aide d'une expression textuelle avoisinant le texte englobé. Nous appellerons cette expression le **locuteur**. Dans le domaine journalistique, le narrateur et l'auteur réfèrent à la même personne. Cette bijection entre auteur et narrateur n'est pas systématique. Ainsi, on trouve parfois dans la littérature le terme **locuteur primaire** pour traiter du narrateur et **locuteur secondaire** pour ce que nous considérons comme le locuteur.

La source peut référer à la personne à l'origine même du segment de discours rapporté, mais également à une entité intermédiaire. On considère comme **chaîne de reprise** la plus petite liste ordonnée des sources tel que :

- le premier élément de la liste soit la personne, ou l'entité à l'origine même du segment de discours rapporté ;
- le dernier élément de la liste soit l'auteur ;
- chaque source de la liste est telle qu'elle a repris le segment de discours de la source située juste avant elle.

Nous introduisons le concept de **degré** caractérisant une source. La source à l'origine même du segment du discours est de degré 0, et les autres sources ont pour degré le nombre de sources qui les sépare de la source de degré 0 au sein de la chaîne de reprise, elles comprises.

Finalement, on appelle **discours source** le discours — écrit ou oral — de la source dans lequel l'auteur a prélevé le passage qu'il a par la suite intégré à son texte englobant. L'**extrait à intégrer** est le passage prélevé au sein du discours source par l'auteur, le **texte englobé** correspond à l'extrait à intégrer transformé pour s'incorporer au texte englobant. La **citation** dénomme le processus global ainsi que la finalité du dit processus. À ce stade, il n'est pas possible de donner une dimension linguistique à la citation.

Définition littéraire

L'Académie Française propose, au sein de son dictionnaire [Académie Française, 1992], une définition au terme citation. Nous nous intéressons ici aux deux dernières versions du dictionnaire.

La définition proposée dans la huitième édition du dictionnaire de l'Académie Française, publié en 1920, est la suivante :

> **8e édition du dictionnaire**
> *Allégation d'un passage, d'une autorité, etc., soit que l'on rapporte le passage, etc., soit que l'on se contente d'indiquer où il se trouve.*

Lors de la publication de la neuvième édition du dit dictionnaire en 1992, la définition du terme citation avait quelque peu évoluée :

9e édition du dictionnaire
Paroles, ou phrase, passage, texte empruntés à un auteur et que l'on reproduit textuellement, de vive voix ou par écrit, pour illustrer, éclairer ou appuyer ce que l'on veut dire.

La différence majeure entre la huitième édition et la neuvième est sans aucun doute la disparition du passage "se contente d'indiquer où il se trouve". Cet extrait faisait très certainement référence au type de citation que l'on peut trouver au sein des articles scientifiques. L'auteur énonce une idée, le plus souvent avec ses propres mots, et indique sous la forme d'une incise entre crochets une référence à la source dans laquelle il a puisé cette idée. Dans le cadre de ce travail, notre définition rencontre celle de 1992, dans la mesure où nous ne traitons pas du cadre particulier de la "citation" scientifique qui est plus proche de la référence bibliographique.

L'énumération des unités qui peuvent être citées au début de la deuxième définition montre que tout ce qui relève du discours peut être repris sous la forme d'une citation. La reprise s'inscrit non seulement au sein d'un acte de communication ("ce qui est dit"), mais elle est partie prenante de cet acte puisqu'elle y joue un rôle : "illustrer, éclairer ou appuyer".

En résumé, la citation consiste à la reprise d'un élément du discours au sein d'un autre discours dans un but défini. Bien que la citation puisse être également orale, nous nous concentrons uniquement sur le médium écrite.

Définition linguistique

Les principaux travaux linguistiques sur le repérage automatique des citations ont été réalisés par les laboratoires de recherche du LaLIC et du GREYC. Nous présentons les travaux du GREYC, qui sont selon nous plus opérationnels, à la section 1.1.1.

Dans [Mourad & Minel, 2000], les auteurs, G. Mourad et J.L. Minel, définissent une citation comme "toute partie d'un acte de communication, lu ou entendu, prélevé en premier temps et greffé : soit exactement à la lettre et marqué typographiquement, soit par d'autre expression en deuxième temps dans un autre acte de communication". On retrouve dans cette définition la trame de base de toute citation, i.e. le prélèvement d'une partie d'un discours par l'auteur pour l'intégrer dans son propre discours. On peut donc considérer deux processus distincts dans l'acte de citation :

1. la sélection de la partie à citer
2. l'intégration de la sélection

Mourad et Minel proposent deux processus d'intégration de la sélection. Le premier consiste à rapporter "exactement à la lettre" et à "marquer typographiquement" le passage rapporté. Ce premier processus est le plus respectueux de l'extrait à intégrer. En effet, le texte englobé est très proche de l'extrait à intégrer. De plus, dans le texte englobant le texte englobé est clairement délimité par des marques typographiques.

Le deuxième processus consiste à rapporter à l'aide "d'autres expressions". Cela engendre une différence nette en terme de mots utilisés ou bien d'agencement de ces mots entre l'extrait à intégrer et le texte englobé. Le résultat de ce processus est donc moins fidèle que celui du processus précédent.

Cette définition nous force à nous interroger sur la fidelité d'une citation. G. Mourad et J.L. Minel nous proposent ici une typologie simple, consistant à considérer d'un côté les citations fidèles à la lettre près, et de l'autre côté toutes les autres.

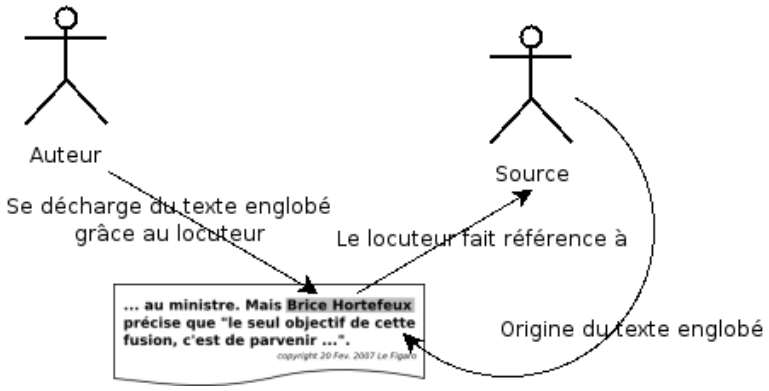


FIG. 1.1 – Relation entre auteur et locuteur

Sans entrer dans le jeu des catégories intermédiaires, nous pensons qu'il existe un niveau de fidelité intermédiaire entre le verbatim et la modification des expressions. Ce niveau est typique du processus d'intégration puisqu'il regroupe toutes les modifications grammaticales nécessaires à l'injection de l'extrait à intégrer au sein du texte englobant : changement de genre, ou bien modification du temps des verbes pour assurer la concordance. Il ne s'agit plus d'une reprise verbatim, mais les expressions ne sont modifiés qu'en surface.

La définition de Mourad et Minel décrit donc les deux étapes du processus de citation et introduit une typologie simple basée sur la fidelité du texte englobé par rapport à l'extrait à intégrer.

Sans contredire la définition précédente, G. Mourad et J.P. Desclès, proposent une nouvelle version plus axée sur le rôle du "locuteur". Ainsi, ils considèrent comme citation "un segment de texte englobé que l'auteur du texte englobant fait prendre en charge par un locuteur (singulier ou collectif) explicite" [Mourad & Desclès, 2002]. S'attachant beaucoup moins au processus et plus à la finalité les auteurs mettent en avant la nécessité de présenter explicitement un locuteur qui va prendre en charge le texte englobé.

Le locuteur auquel les auteurs font allusion est une expression du texte englobant qui fait référence à ce que l'auteur exhibe au lecteur comme la source du texte englobé. La relation entre l'auteur et le locuteur est illustrée par la figure 1.1.

L'utilisation du terme "explicite" n'est pas plus détaillé dans l'article, les auteurs ne nous fournissent donc pas de critère permettant de différencier les locuteurs explicites de ceux qui ne le sont pas. Nous considérons pour notre part qu'un locuteur est considéré explicite lorsqu'il respecte les critères suivants :

- l'expression linguistique utilisée pour désigner le locuteur fait référence à une personne (physique ou morale) ou bien à un groupe de personnes ;
- l'expression introduit le texte englobé dans le co-texte[1] à l'aide d'un verbe d'énonciation ou bien appartient à un syntagme prépositionnel précédant ou suivant le texte englobé ;
- le texte englobé n'est pas séparé de l'expression par d'autres expressions faisant référence à des locuteurs potentiels.

En résumé, les auteurs utilisent indifféremment le terme *citation* pour l'acte de prélèvement d'informations rapportées par la suite, ainsi que pour la représentation linguistique produite par cet acte. La présence

[1] Le co-texte correspond à la partie du contexte accessible uniquement à partir du texte

d'un locuteur, autre que le locuteur principal, prenant en charge les dits rapportés par l'acte de citation est nécessaire selon les auteurs. La présence d'une expression linguistique faisant référence à ce dernier permet d'expliciter sa présence.

Définition opérationnelle

Les subtilités linguistiques augmentant la complexité du repérage automatique, certains chercheurs se sont penchés sur une formalisation plus épurée de la citation afin de permettre un repérage automatique robuste.

Observant la nécessité de constituer des listes importantes de marques linguistiques pour la détection automatique des citations, E. Giguet et N. Lucas [Giguet & Lucas, 2004] (travaux réalisés au laboratoire GREYC) proposent une formalisation épurée des citations en introduisant le concept d'objet citationnel et ses concepts associés.

L'approche de Giguet et Lucas est opérationnelle car elle se base sur le constat de la grande variabilité des formes introductrices des citations induisant le besoin d'une méthode alternative aux listes. Leur proposition est d'abstraire la grande variation des introducteurs en se concentrant sur ce qu'ils appellent les "invariants". En se concentrant sur les aspects communs à toutes les citations, Giguet et Lucas proposent une forme citationnelle abstraite constituée des éléments : source, relateur et discours rapporté. Ils nomment ce triplet la "séquence canonique de la citation". Le discours rapporté correspond à ce que l'on dénommait précédemment comme "texte englobé" alors que la source correspond à notre expression locuteur. Le relateur quant à lui est "le segment établissant la relation entre la source et le discours rapporté" [Giguet & Lucas, 2004]. On peut considérer ce dernier comme une contrainte permettant de s'assurer du caractère "explicite" de la source.

L'abstraction proposée s'accorde donc plutôt bien avec la définition précédemment proposée par G. Mourad. Ils contraignent cependant un peu plus leur formalisation en proposant un modèle censé représenter les formes de citation que l'on peut rencontrer en français. Il se constitue de deux motifs :

– l'ordre normal : `source + relateur + discours rapporté`
– l'ordre inversé : `discours rapporté + relateur + source`

Ce modèle répond à un besoin opérationnel en permettant d'annoter certains objets citationnels sur des critères positionnels dans le cadre du repérage automatique. Il est mis en place pour répondre à un besoin technique. Par exemple, si l'on identifiant une séquence canonique comme *source + ? + discours rapporté*, alors on déduirait à partir du motif d'ordre normal que le deuxième composant correspond au relateur.

Il existe des citations en français qui ne respectent pas le modèle proposé par les chercheurs du GREYC. C'est ainsi le cas de la citation 1.1.1 qui correspondrait à un motif du type `discours rapporté + relateur + source + discours rapporté`. Une majorité des citations respectent toutefois le modèle.

> "Il faut comprendre, a estimé David Morrison astronome au centre Ames de la Nasa en Californie, que si le danger est réel, sa probabilité est faible, mais les conséquences énormes, avec la menace d'une extinction totale de toute vie sur Terre."
>
> © *Le Figaro*

L'approche opérationnelle de [Giguet & Lucas, 2004] nous offre une formalisation épurée : la séquence canonique de la citation où toute citation se compose des trois éléments : source, relateur et discours

rapporté. Le modèle proposé semble cependant trop contraignant, omettant certains “motifs citationnels” courants (placement du locuteur en incise dans le texte englobé notamment).

Synthèse et prise de position

Le processus citationnel est une démarche en deux étapes. La première consiste à prélever un extrait d'un acte d'énonciation, et la seconde à rapporter ces informations. Nous nous intéressons uniquement aux diverses formes d'intégration des informations prélevées et c'est ce que nous appellerons par abus de langage “citation” par la suite.

L'information prélevée dans l'acte de communication puis intégrée au texte englobant est le texte englobé — discours rapporté selon [Giguet & Lucas, 2004] — ce dernier est accompagné d'une expression textuelle faisant référence à la source de l'information : l'expression locuteur. Cette expression locuteur permet d'introduire un locuteur différent qui permet à l'auteur de se désengager des propos qu'il rapporte. Il indique ainsi au lecteur qui a la responsabilité des propos rapportés. Ce locuteur fait référence à une source, ce qui s'accorde avec la définition de citation d'après [Mourad & Desclès, 2002].

Selon [Giguet & Lucas, 2004], une citation s'abstrait en une séquence canonique composée d'une source, d'un relateur et d'un discours rapporté. Le relateur n'a selon nous pas systématiquement d'existence linguistique. La structure de la phrase peut suffire à relier la source au discours rapporté comme l'illustre l'exemple 1.1.1 où la source (soulignée) et le discours rapporté (en italique) sont mis en incises par " :".

une même complainte s'élève de tous les rangs sarkozystes : *le candidat est devenu invisible*.

© *Le Figaro*

La notion de relateur n'en reste pas moins importante puisqu'elle répond d'une certaine façon à l'idée de “locuteur explicite” introduite par [Mourad & Desclès, 2002]. Nous considérons toutefois le relateur comme un concept secondaire par rapport à la source et au discours rapporté.

Nous emploierons donc un formalisme assez proche de celui proposé par [Giguet & Lucas, 2004]. Ainsi, chaque citation devrait se composer :

- d'une expression locuteur introduisant un locuteur faisant lui même référence à la source ;
- du texte englobé ;
- d'une forme linguistique assimilable à un relateur, mais cet objet est optionnel.

En résumé, une citation est un segment textuel contenant un texte englobé — fragment de texte formé à partir d'informations prélevés dans un acte de communication antérieur —, une expression locuteur — expression textuelle introduisant un locuteur — et éventuellement un relateur.

1.2 Style de discours rapporté ou différentes formes d'intégration du discours extrait

Il existe quatre variétés de discours rapporté définis dans la littérature : le discours direct, le discours indirect lié, le discours indirect libre et le discours narrativisé. À ces formes classiques s'ajoutent les formes hybrides de discours indirect (DI) avec îlots textuels et le discours indirect quasi textuel. Nous traiterons dans un premier temps des formes communes de discours rapportés, puis nous traiterons des formes hybrides de discours indirect, spécifiques au style journalistique.

1.2.1 Style direct, indirect, indirect libre et narrativisé

Les formes les plus présentes dans la littérature sont le discours direct, indirect lié, indirect libre et narrativisé. Nous introduisons les spécificités de chacune de ces formes à l'exception du discours indirect libre peu usité. La compréhension des spécificités de ces différentes formes de discours nous permettra par la suite de caractériser les citations y faisant appel.

Style direct

Le discours direct consiste en la transcription exacte d'un discours énoncé par une personne qui n'est pas le narrateur. L'énoncé source est conservé tel quel et introduit au sein du récit à l'aide de marques de ponctuations spécifiques que sont les deux points et les guillemets. La transcription se trouve donc dans un plan syntaxique indépendant du récit dans lequel elle s'inscrit. Le discours direct est ainsi censé être objectif et neutre puisqu'il rapporte les paroles telles quelles.

Les marques ponctuatives que sont les deux-points et les guillemets sont très certainement les éléments les plus caractéristiques du discours direct. Ces marques permettent de clairement délimiter l'énonciation du narrateur des paroles du locuteur, comme dans l'extrait ci-dessous tiré du journal *Le Monde* :

> Le quotidien économique souligne : "Si le rapport ne veut pas associer ces montants à l'idée d'une nouvelle 'cagnotte' budgétaire, ni au débat électoral sur le niveau de prélèvements obligatoires, le montant est équivalent au déficit budgétaire de l'Etat, à savoir 36,5 milliards d'euros l'an dernier."
>
> *Le Monde*

La marque ponctuative " :" suivie de l'ouverture des guillemets marque la rupture entre le plan d'énonciation du narrateur et celui du locuteur. Ces deux plans sont syntaxiquement indépendants puisqu'ils ne correspondent pas à la même énonciation. Le plan du narrateur est cohérent avec l'ensemble du discours au niveau du temps des verbes, de la personne ou encore des repères spatio-temporels. Le plan du locuteur — autre que le narrateur, bien entendu — est caractérisé par le temps des verbes s'articulant autour du présent (*"ne veut pas", "est"*), les pronoms personnels de la première et deuxième personne majoritairement et enfin des repères spatio-temporels relatifs à la situation d'énonciation (*"l'an dernier"*).

Le style direct donne une impression d'objectivité et de neutralité en détachant le plan du narrateur de celui du locuteur cité à l'aide de verbes conjugués sur la base du présent et des repères spatio-temporels relatifs à la situation d'énonciation. Les passages rapportés sont clairement délimités du reste du récit à l'aide des marques ponctuatives du deux points et des guillemets. Ceci en fait un style assez facile à appréhender par le lecteur ou par une méthode automatique.

Style indirect

Le discours indirect est l'intégration dans le flux discursif d'un discours ayant été énoncé dans un cadre temporel antérieur et généralement par une personne différente du narrateur. Cette intégration entraîne l'adoption morphosyntaxique du discours rapporté. L'énoncé original subit des transformations au niveau des pronoms personnels, du temps des verbes ainsi qu'au niveau des indicateurs spatio-temporels. Ces transformations ont pour objectif de placer le récit et le discours rapporté sur un même

plan syntaxique afin que ce dernier soit considéré par le lecteur comme partie du récit. Le discours indirect est moins objectif que le discours direct, permettant à l'auteur une modification de la signification des paroles par une manipulation du co-texte.

Les phrases rapportant du discours indirect sont la plupart du temps composées d'une proposition principale contenant le locuteur ainsi que le verbe amenant le discours rapporté, tandis que ce dernier est placé au sein d'une proposition subordonnée. Ces propositions s'articulent la plupart du temps autour de la particule conjonctive "que", comme dans l'extrait ci-dessous tiré du journal *Le Monde* :

> il a déclaré qu'il pourrait nommer un premier ministre de gauche s'il était élu président de la République.
>
> *Le Monde*

La structuration des phrases contenant du discours indirect en proposition influe directement sur les temps des verbes. En effet, les règles de concordance des temps s'appliquent aux verbes des propositions subordonnées et donc des verbes rapportés. L'application de la concordance oblige donc, afin de conserver les phrases grammaticalement correctes, à modifier le temps des verbes rapportés. Le mode indicatif des verbes rapportés peut ainsi se transformer en conditionnel ou subjonctif selon le temps et mode du verbe de la proposition principale ainsi que la chronologie de l'action rapportée dans la subordonnée par rapport à l'action rapportée dans la principale.

Les pronoms personnels varient selon la personne qui rapporte les paroles et notamment selon la personne utilisée pour le narrateur. L'utilisation d'une première personne pour ce dernier implique une variation assez uniforme des trois personnes, alors que l'emploi d'une troisième personne entraîne une utilisation majoritaire de troisième personne. La détection de la personne utilisée pour le narrateur peut donc avoir son intérêt.

Les indicateurs spatio-temporels du discours indirect sont détachés du cadre de l'énonciation originale afin de s'inscrire logiquement dans la narration. Les indicateurs relatifs au cadre de l'énonciation originale tels que "aujourd'hui", "demain", "ici",. . .sont remplacés respectivement par "ce jour là", "le lendemain", "ce lieu ci",. . .Le but de ces remplacements est de rendre le discours rapporté compréhensible en-dehors du co-texte dans lequel il s'inscrivait.

L'utilisation du discours indirect nécessite la modification interne de ce qui est rapporté. À cause de cette modification, son utilisation est considérée comme moins objective, moins respectueuse que le discours direct, l'information rapportée ne subit cependant pas de modification, seule la forme est modifiée.

Style narrativisé

Le discours narrativisé est particulier puisqu'il ne s'agit pas réellement de discours rapporté. En effet, le narrateur indique au travers du récit qu'il y a eu un acte d'énonciation, mais le contenu énoncé n'est pas rapporté.

> Papon décrit les arcanes de Vichy sous l'occupation.
>
> *Le Figaro*

Ce style de discours n'a pas réellement d'intérêt dans le cadre de notre étude étant donné qu'il n'est pas utilisé dans les citations.

En respectant à la lettre les paroles rapportées, le discours direct est considéré comme plus objectif que le discours indirect. Cependant, les modifications apportées lors de l'utilisation du discours indirect ne concernent que la forme et pas le fond. L'utilisation du discours direct tend à donner une impression de vérité des paroles rapportées, alors qu'au delà de la modification de la forme de l'énoncé, le placement dans un co-texte différent modifie plus profondément l'énoncé rapporté.

1.2.2 DI avec îlots textuels et DI quasi-textuel

L'utilisation répétée des styles du discours rapporté au sein des articles journalistiques a provoqué une évolution de ces derniers. Ainsi, de nouveaux styles hybrides sont apparus, complétant les styles littéraires classiques, apportant des spécificités propres au domaine journalistique. Nous présentons dans cette section les évolutions du discours indirect (DI) apparus au sein des articles journalistiques : le DI avec îlots textuels et le DI quasi-textuel.

Le linguiste G. Komur traite de la prise de distance des locuteurs dans le genre journalistique par l'utilisation de nouveaux genres de discours indirects [Komur, 2001]. Il traite notamment du DI avec "îlot textuel". Ce style de discours reprend les caractéristiques du discours indirect lié, mais a la particularité d'intégrer des fragments de texte placés entre guillemets. L'exemple 1.2.2 illustre l'intégration d'un tel fragment entre guillemets au sein d'un discours rapporté au discours indirect. Ainsi, le segment "elle répond qu'elle est prête à" présente les caractéristiques d'un discours rapporté au discours indirect au sein d'une proposition subordonnée. Le fragment "ouvrir le débat" n'est lui pas du tout caractéristique du discours indirect, mais plutôt du discours direct. Cet "îlot textuel" correspond bel et bien à une reprise verbatim d'un passage de l'énoncé original. Ce mélange des styles de discours justifie l'appelation de "style de discours hybride".

elle répond qu'elle est prête à "ouvrir le débat"

© Le Figaro - 20 Février 2007

Le DI "quasi-textuel" se différencie du DI avec "îlots textuels" par la taille de l'emprunt à l'énoncé d'origine. Ainsi, d'après G. Komur, le DI quasi-textuel reprend des pans entiers du discours source : "l'intégralité du message d'origine est conservé" [Komur, 2001]. Le linguiste ne nous informe pas plus précisément sur ce qu'il considére comme une conservation complète du message d'origine. La distinction précise entre DI avec "îlots textuels" et "quasi-textuel" n'étant pas primordiale pour notre travail, nous ne cherchons pas à l'affiner. Un travail typologique sur la citation nécessiterait toutefois d'éclaircir ce point. Nous approfondissons toutefois la caractérisation de ces formes de discours dans la section 2.3.1.

La forte consommation par les journalistes du discours rapporté a permis l'émergence de deux formes hybrides du discours indirect. Principalement basées sur le discours indirect, elles intègrent également des caractéristiques du style direct et sont assez présentes au sein des articles de presse, mais nous n'avons pas trouvé de référence de ces constructions au niveau de la littérature. Celà en fait donc une pratique caractéristique de la citation journalistique.

1.3 Techniques de détection automatique

1.3.1 Travaux antérieurs sur le repérage de citation

Deux grandes approches ressortent de notre bibliographie : la première mise au point au LaLIC (Langages, Logiques, Informatique, Cognition), et la seconde au Greyc (Université de Caen). Nous notons l'absence de travaux anglophones sur le sujet. Nous avons délaissé ces derniers car ils se focalisent sur les citations au sein des articles scientifiques et utilisent des éléments extérieurs comme la bibliographie ou une base de données externe pour repérer les citations au sein des textes ([S. Teufel & Tidhar, 2006]).

Détection automatique : identification vs. repérage

On parle de détection automatique de citations sous deux termes au sens distinct : l'identification et le repérage.

Le repérage de citation est la vision la moins exigeante. Elle consiste en effet à isoler des fragments du texte analyser et à les exhiber comme contenant une ou plusieurs citations. Le programme n'est cependant pas censer pouvoir en dire plus sur le contenu ou l'origine de la dite citation.

L'identification est une forme plus aboutie de la détection. La méthode se base très certainement sur une forme de repérage, mais le traitement abouti à une identification des éléments de la citation, et non plus un simple segment de texte qui contiendrait la citation. Comme expliqué dans l'article [Mourad & Desclès, 2002], l'identification permet de répondre aux questions : Qu'est-ce qui est dit ? Par qui ?

Dans le cadre du projet Piithie, la première vision ne nous semble pas pertinente. Toutefois, la seconde nous semble trop difficile à atteindre dans le temps qui nous est imparti. Nous nous proposons donc de mettre en place une méthode intermédiaire qui répond aux questions : Qui dit ? Qu'est-ce qui est dit ?

Approche lexicale par exploration contextuelle

L'exploration contextuelle est la méthode employée par le laboratoire LaLIC dans le cadre de la détection de citation [Mourad & Minel, 2000], [Mourad & Desclès, 2001], [Mourad & Desclès, 2002] et [Mourad, 2001].

L'exploration contextuelle est une méthode très prisée des linguistes. Elle consiste en l'acquisition de régularités lexicales caractéristiques d'une forme linguistique d'intérêt comme la citation dans notre cas. Ces régularités lexicales sont appelées *indices* ou *embrayeurs*. Lorsque l'indice est réellement spécifique à l'objet linguistique étudié et qu'il marque avec une bonne probabilité la présence du dit objet, les chercheurs du LaLIC le nomme *marqueur*.

L'exploration contextuelle présentée dans les articles précédemment citée n'a pas été uniquement appliquée pour des articles journalistiques, mais également sur des textes scientifiques, techniques et au sein de la littérature. De plus, le corpus utilisé contient comme seule source journalistique, les archives du journal "Le Monde Diplomatique" de 1989 à 1998. L'unicité du journal, qui plus est d'une édition spécialisée, est criticable car non représentatif de la grande variation du style d'articles de presse. Cette étude permet toutefois d'extraire un nombre non négligeable d'indices sur lesquels se sont basés des études postérieures.

L'exploration contextuelle sur ce corpus a résulté en l'extraction d'un certain nombre d'indices et de marqueurs, regroupés par Mourad et al. dans trois catégories :

– les marqueurs typographiques
– les marqueurs typographico-linguistiques
– les marqueurs purement linguistiques

Parmis les marqueurs typographiques, on peut notamment noter la séquence ponctuative *"deux-points ouverture des guillemets"*, les incises éliptiques entre guillemets, les points d'exclamation ou d'interrogation aux abords de guillemets fermants ou encore la présence de pronoms personnels au sein d'un couple apparié de guillemets.

Les marqueurs linguistiques sont quant à eux déclinés en trois sous-catégories :

– syntagmes prépositionnels : *selon X, pour X, d'après X, aux yeux de X, ...*
– introducteurs spécifiques : *l'observation de X, les termes de X, ...*
– verbes d'introduction de citations

Le *X* est à substituer dans les textes par une expression locuteur. Ces marques ne sont donc pas repérables à l'aide d'une simple recherche lexicale mais nécessitent également de repérer des segments textuels qui peuvent être considérés comme faisant référence à une source. L'étude n'apporte aucune information sur ce point. Le dictionnaire des verbes d'introduction constitué est publié au sein de [Mourad & Desclès, 2001] et représente près de 800 entrées. Leur travail a principalement porté sur la constitution de cette liste de verbes et la classification de ces derniers selon le degré d'engagment de l'auteur qui les utilise par rapport à ce qu'il rapporte [Mourad & Desclès, 2001].

En résumé, l'approche contextuelle a permis de recenser "3000 formes verbales et introducteurs spécifiques" [Mourad & Desclès, 2002]. Ces formes linguistiques sont spécifiques à la langue française. Bien que les auteurs ne donnent pas d'information précises sur les performances des algorithmes élaborés à partir de leur travail, certaines citations introduites sous des formes plus ou moins originales ne sont pas repérées car elles n'utilisent pas les marques accumulées. L'ajout de nouvelles marques aux dictionnaires n'est pas une solution puisqu'il augmente le taux de faux positifs détectés. Autre écueil à la constitution de grandes listes, le coût du passage à une autre langue : il est nécessaire de recommencer tout le travail de compilation des marques. Pour toutes ces raisons, la constitution de liste est considérée inadaptée pour la mise au point d'un système de détection robuste par le groupe du Greyc.

Approche syntaxique

La grande variabilité des marques introduisant les citations a enclenché une réflexion une méthode alternative ne nécessitant pas la constitution de lexiques de grande taille. L'article [Giguet & Lucas, 2004] introduit les concepts de "constantes" ou "invariants" dans ce but.

Les auteurs considèrent les invariants comme les constantes d'une équation. Les trois invariants citationnels seraient alors *la source*, *le discours rapporté* et *un relateur* "verbal, conjonctif ou prépositionnel" [Giguet & Lucas, 2004]. Pour ce qui est du terme *source*, il est définit comme "nom propre du locuteur et éventuellement ses qualifiants", on considère comme nom propre l'expression linguistique utilisée comme référence inambigue à la source.

Sans utiliser de dictionnaire, il est tout de même nécessaire d'utiliser des ressources linguistiques : les "critères". Ceux-ci peuvent être "interne" s'ils sont intra-phrastiques ou "externes" s'ils sont trans-phrastiques. Les critères internes sont classables dans trois catégories :

– indices typographiques : *ponctuation, casse, ...* ;

– indices morpho-syntaxiques : *morphèmes grammaticaux comme "que", suffixes "ent", ...* ;
– indices positionnels *début, fin d'unités textuelles, ...*

N. Lucas et E. Giguet utilisent le terme "trans-phrastique" pour décrire les éléments qui s'étendents sur les phrases voisines de celle considérée.

Contrairement à l'approche contextuelle, les indices ne sont donc pas des ressources lexicales, mais plutôt syntaxiques, voire même des motifs de quelques lettres. Ces indices ne sont transformés en marques que dans un contexte donné ou leur cooccurence induit leur utilisation comme marqueurs.

Les indices ne servent plus, contrairement à la méthode par exploration contextuelle, à détecter les citations, mais à isoler les invariants de ces dernières. La co-présence et la position relative des indices permet de considérer des segments du texte comme *sources*, *relateurs* ou *discours rapportés*. L'intérêt de cette méthode est la possibilité d'effectuer des déductions positionnelles. En effet, les auteurs proposent un modèle de citation française composé des motifs suivants :

– l'ordre normal : source + relateur + discours rapporté
– l'ordre inversé : discours rapporté + relateur + source

Selon ce modèle, l'identification de deux invariants permet, si les indices ne sont pas assez nombreux, d'identifier le troisième invariant sur critère positionnel.

L'idée des invariants permet de résoudre les cas difficiles où les indices manquent, elle n'a d'intérêt cependant que si le modèle proposé est valide. Ce dernier nous semble cependant incomplet comme nous le montrerons par la suite à partir d'exemples tirés de notre corpus. Ici encore, les auteurs ne fournissent pas d'information quantitative sur les performances de leur méthode. Elle représente toutefois une approche alternative d'intérêt à l'utilisation des listes de ressource de l'exploration contextuelle.

1.4 Synthèse

La citation est à la fois le processus de prélèvement et d'intégration d'un fragment d'un acte d'énonciation au sein d'un autre. Nous nous attachons toutefois au produit fini de ce processus — que l'on nomme également citation — au sein des articles de journaux. La partie du discours source intégrée à l'article est appelée le *texte englobé* alors que la prose l'encadrant est appelée le *texte englobant*. Ce dernier est l'oeuvre de l'auteur, qui décharge la paternité du *texte englobé* à une *source* à laquelle il fait référence à l'aide d'une *expression locuteur* permettant l'introduction d'un *locuteur* et donc un nouveau plan d'énonciation.

Nous nous basons partiellement sur la formalisation en séquence canonique de la citation [Giguet & Lucas, 2004] pour décrire les objets citationnels qui composent les citations à savoir : la source, le relateur et le discours rapporté, mais nous préférons la terminologie : locuteur, relateur et texte englobé. Ce dernier est intégré au texte englobant par le biais de styles du discours. Le *discours direct* étant plus objectif car plus proche du texte source alors que le *discours indirect* impose des modifications morphosyntaxiques du discours repris. À ces deux formes communes de la grammaire française s'ajoute l'utilisation des *îlots textuels* [Komur, 2001] permettant une contraction de l'acte d'énonciation rapporté. La pratique des îlots textuels est particulière au domaine journalistique.

Une autre particularité du domaine journalistique est l'effort de la part de l'auteur pour constamment distinguer son plan d'énonciation de ceux des sources auxquels il fait appel. Il affirme ces prises de distance avec les propos en imposant des marques typographiques et ponctuatives — considérées superflus par la grammaire — pour encadrer le texte englobé. Par ces pratiques, la citation journalistique se détache quelque peu des pratiques grammaticales du discours rapporté. Ces prises de distance avec la grammaire

ont deux buts :

1. marquer clairement la séparation entre le discours de l'auteur et ceux — rapportés — des sources qu'il cite ;
2. apporter un maximum d'éléments extérieurs pour justifier les informations tout en conservant un texte fluide et facile à lire.

Nous avons éprouvé le besoin de trouver des illustrations des différentes notions présentées dans les travaux antérieurs. Ainsi, nous avons décidé de nous lancer dans la constitution d'un corpus qui nous permettrait de constater les différentes formes de citation au sein de leur environnement textuel.

Chapitre 2

Constitution et observation d'un corpus d'étude journalistique

Au fur et à mesure de l'avancement de l'état de l'art, le besoin d'avoir des exemples pour comprendre et comparer les différentes techniques s'est fait sentir. Bien que la plupart des articles se basaient, selon leurs auteurs, sur des corpus, aucun ne donnait d'informations précises sur ce dernier, et encore moins de moyen de les consulter. Seul le groupe "Ci-dit" [Rosier *et al.*,] offrait l'accès à trois corpus sur leur site : `www.ci-dit.com`.

Sophie Marnette (Université d'Oxford), Laurence Rosier (Université Libre de Bruxelles) et Elena Meteva (Université de Sofia), donnent accès à des corpus qu'elles ont constitués dans le cadre de leurs recherches. Le corpus fournit par Sophie Marnette est classé selon les styles du discours, mais n'est pas annoté. Les corpus de Laurence Rosier et Elena Meteva comportent des annotations permettant de repérer source et discours rapporté. Malheureusement les passages citationnels sont fournis sans le contexte dans lequel ils s'inscrivent. Il ne semble pas y avoir à l'heure actuelle d'autre compilation de ressources citationnelles.

La constitution de tout corpus passe par une étape de réflexion sur le type du corpus, son adéquation avec le projet, la possibilité de le réutiliser, sa taille, sa représentativité, l'utilisation de textes complets ou d'échantillons et l'annotation [Marshman, 2003]. Cette section du rapport décrit chacun des points énumérés ci-dessus.

2.1 Constitution du corpus

La constitution du corpus ne doit pas être prise à la légère. Ce dernier a pour rôle de nous fournir les illustrations des concepts que nous avons dans le premier chapitre. Il nous servira également de support d'extraction de nos règles d'apprentissage supervisé. Une grosse partie de notre travail dépendra donc directement de la qualité du corpus. Ainsi, nous définissons tout d'abord nos besoins concernant ce corpus, puis nous présentons les modifications que nous avons apportées aux textes extraits afin de faciliter leur utilisation et leur division.

2.1.1 Définition des besoins

Avant de nous lancer dans l'élaboration du corpus, nous nous sommes attaché à définir les besoins auxquels devraient répondre le corpus. Nous nous sommes basé sur l'article très complet de B. Habert

[Habert, 2001] pour préparer notre cahier des charges et cadrer nos besoins.

Le corpus que nous avons constitué devait nous apporter des éléments de réponses aux questions suivantes qui nous bloquaient dans notre avancée :

- quels types de citations sont les plus présentes dans le domaine journalistique ?
- que doit-on considérer comme citation et comme non-citation ?
- existe-t-il des cas litigieux à la frontière entre le discours rapporté et le reste du texte ?
- est-il possible de définir une "unité citationnelle" ?
- les motifs proposés par [Giguet & Lucas, 2004] sont-ils adaptés ? en existe-t-il d'autres ?

Le corpus devait donc être le plus représentatif possible des citations que l'on peut trouver dans la presse francophone. Mais il devrait également nous permettre de mettre à l'épreuve les méthodes rencontrées dans notre état de l'art, ainsi que nos propres hypothèses.

Représentativité du corpus

Notre travail étant plutôt exploratoire, le corpus sur lequel nous allons travaillé doit être le plus représentatif possible de ce que l'on peut trouver dans la presse francophone en terme de citations. Afin de ne pas trop nous disperser, nous nous sommes limités à la presse généraliste informative. L'avantage de la presse généraliste est qu'elle est également la plus lue et donc a l'impact le plus important sur les façons de faire. De plus, la presse informative utilise au mieux la citation dans le but de conserver une certaine objectivité. En puisant dans les journaux généralistes informatifs nous voulions nous assurer une bonne richesse citationnelle.

Un corpus trop important aurait nécessité une période d'annotation trop importante, mais un corpus trop petit n'aurait pas permis une exhaustivité suffisante des phénomènes de citation , ce qui est délicat à gérer lors de la phase d'apprentissage.

Nous nous limiterons dans un premier temps à la constitution d'un corpus en langue française. Bien que le projet Piithie (cf 3) ait une dimension multilingue, il nous semble nécessaire d'effectuer le travail exploratoire en français tout en conservant la possibilité de compléter le corpus par la suite avec des textes dans d'autres langues.

Finalement, pour s'assurer une bonne représentativité, il est nécessaire de considérer un maximum de styles d'écritures afin de maximiser le nombre de structures de phrases ainsi que des syntagmes introducteurs de citation. Pour ce faire, nous devons sélectionner des textes provenant d'auteurs différents, mais également de journaux différents. En effet, les contraintes éditoriales tendent à niveler les styles d'écriture des auteurs.

En résumé, le choix des textes seront réalisés parmis une sélection d'articles de divers journaux francophones informatifs et généralistes. Le nombre de textes doit être suffisamment élevé pour permettre une bonne variété de citations, une cinquantaine d'articles semble un bon compromis entre richesse et taille du corpus à annoter.

Choix de textes complets

Un critère non négligeable dans la constitution de notre corpus est le choix de sélectionner des textes entiers ou bien seulement des extraits contenant les citations.

Le choix de compiler uniquement des extraits dans des corpus peut se justifier par une densité d'information utile faible au sein des textes complets. Cette faible densité pourrait se traduire en un

apprentissage supervisé médiocre car rencontrant trop de contre-exemples, selon les conditions d'apprentissage choisies.

Dans notre cas, en faisant le choix d'articles riches en citations, nous pensons maintenir la densité d'information utile à un niveau suffisamment élevé. La taille des articles de presse se limitant à quelques centaines de mots, le corpus ne devrait pas être surchargé par des informations non pertinentes.

Un autre argument en faveur des textes complets réside dans la difficulté de déterminer les bornes de certaines citations. Ainsi, si l'on ne peut s'assurer de correctement définir les bornes des citations — qui peuvent potentiellement s'étendre au-delà de la phrase —, la sélection d'extraits citationnels pourraient entraîner le troncage de ces citations et donc la perte d'informations nécessaires à l'apprentissage.

La conservation des textes complets plutôt que des extraits n'implique pas systématiquement une surcharge d'information inutile. Au contraire, le choix des textes complets assurer la conservation de la totalité des informations utiles. Nous avons choisi donc de considérer les textes complets, ce qui de plus à une utilisation future de notre corpus dans le cadres d'autres recherches. Enfin, si jamais il était nécessaire de travailler uniquement sur les segments textuels assimilés à des citations, il est toujours possible d'extraire ces derniers.

Le corpus des citations respectera donc les critères suivants :

- articles/brèves journalistiques francophones ;
- accessibles sous forme numérique ;
- riches en citations ;
- représentant différents thèmes et différents styles d'écriture ;
- textes complets et non limité aux seules citations.

Le besoin défini, nous pouvons nous lancer confiant dans la constitution du corpus, puis dans sa structuration. Cette dernière est importante puisqu'elle va permettre un accès plus ou moins facile à l'information.

2.1.2 Choix des journaux et des articles

La constitution de notre corpus s'effectue à partir de ressources journalistiques présents sur le web. Ce choix est principalement dû à la facilité d'accès à ce genre de ressources. Nous discutons dans un premier temps du choix des journaux, puis de celui des articles au sein de ces journaux.

Choix des journaux

Les journaux choisis doivent être en adéquation avec le projet, à savoir la détection de citation dans un contexte journalistique. Il est donc nécessaire de réunir un panel de journaux qui soit suffisamment représentatif de la presse francophone, en donnant la priorité aux journaux "populaires", ces derniers étant les plus à même de traiter de la plus grande variété des sujets.

Le choix s'est porté sur quatre quotidiens nationaux et un quotidien Belge. Les critères de sélection des titres sont les suivants :

- Nombre d'exemplaires vendus ;
- Éclectisme des sujets ;
- Disponibilités en ligne des articles.

Les journaux choisis pour le corpus sont listés, accompagnés de leurs statistiques de vente, dans le tableau 2.1.

	Ventes moyennes par numéro en 2006	Visites mensuelles du site
Le Monde	312 265	38 262 937
Le Figaro	322 497	24 448 823
Challenges	256 730	7 051 790
Libération	127 687	9 189 585
Le Soir	nc	nc

TAB. 2.1 – Statistiques des ventes de différents journaux (source : www.ojd.com)

Il aurait été intéressant de rajouter un journal local à la sélection, mais nous n'avons trouvé aucun quotidien local ayant un portail internet proposant des articles satisfaisant nos besoins.

Une plus grande variété des titres n'aurait pas forcément apporté une meilleure représentativité. En diminuant le nombre d'articles par journal, nous aurions également réduit le nombre de citations. Ceci aurait pû avoir pour effet de marginaliser les types de citations et aurait rendu l'apprentissage — sur un corpus de petite taille comme celui-ci — plus difficile.

Les articles choisis sont issus de journaux édités au format papier ainsi qu'au format numérique publié sur leur site web. L'existence des textes au format numérique est un avantage important puisque ces derniers peuvent être traités directement dans leur format original. Le choix de l'édition numérique des articles s'est donc imposé par la disponibilité immédiate des dits articles ainsi que l'accès gratuit.

Choix des articles

Nous estimons la taille nécessaire du corpus à une cinquantaine d'articles, soit quelques dizaines de milliers de mots. Celà nous apparaît comme un bon compromis entre le besoin d'une taille importante assurant la richesse du contenu et une taille suffisamment petite permettant une annotation manuelle. Nous considérons cette taille comme permettant d'obtenir un condensé suffisant de citations annotable manuellement en un temps contrôlé.

La taille moyenne d'un article de journaux dans les quotidiens sélectionnés est de 500 à 600 mots. Une dizaine d'articles par journaux était donc suffisant pour constituer notre corpus. Nous avons sélectionné l'ensemble de ces articles à une même période de temps, et selon le même critère : la une en ligne des dits journaux. Ainsi, nous avons sélectionné les articles faisant la une de chacun des quotidiens sur leurs sites respectifs durant la semaine du 19 février 2007.

Les articles à la une sont des articles traitant de sujets populaires et gageurs d'une certaine qualité éditoriale. Ils sont donc "représentatifs" du journal. De plus, en pleine période électorale, la plupart des articles traitent d'interviews ou de discours politiques. Ces types d'articles sont particulièrement riches en citations. Enfin, il est plutôt rare qu'un journaliste publie plus d'un article à la une par jour ; la diversité des journalistes implique également la diversité des styles d'écriture.

Un argument supplémentaire au choix de la "une" des journaux à une même période est la forte probabilité d'obtenir des journaux traitant du même sujet, et donc rapportant potentiellement les mêmes énoncés. Les citations différentes d'énoncés communs nous paraîssent riches d'intérêt.

Voici, ci-dessous, la répartition des articles par titre :

Journal en ligne	Nb. articles	Nb. mots
Libération	10	5555
Le Soir	10	5607
Challenges	11	11332
Le Monde	12	5957
Le Figaro	10	5305
Totaux	53	33756

Le corpus ainsi constitué représente environ 30 000 mots ce qui en fait un corpus de petite taille. Cette petite taille risque d'être handicapante pour la partie exploratoire de cette recherche, mais il ne nous semble pas faisable d'étudier manuellement un corpus de plus grande taille pour le temps qui nous est imparti. Nos critères de sélection doivent toutefois nous permettent d'avoir un nombre intéressant de citations et une variété correcte de ces dernières.

2.1.3 Prétraitement et format des articles

Les différents journaux en ligne sélectionnés publient leurs articles au format HTML dépourvu d'informations de structuration logique des documents. Il est de notre ressort de nettoyer les fichiers récupérés, puis de les structurer en y intégrant des informations complémentaires à l'aide du métalangage XML[1].

Sauvegarde et nettoyage des articles

Les pages web des journaux desquels nous avons extrait les articles de notre corpus n'étaient pas directement utilisable. Les publicités, les menus, les différentes feuilles de style ainsi que les données pour faciliter le référencement, transforment rapidement une source HTML en un amas de code difficilement compréhensible.

Chacun des articles a été sauvegardé dans son format HTML original dans un premier temps. La conservation du format original est toujours une bonne idée, au minimum pour s'assurer par la suite que le nettoyage n'a pas supprimé d'informations utiles. La sauvegarde sur disque permet de prévenir quant à elle, un retrait de l'article du site d'origine et assure la pérénité du corpus.

À propos de la sauvegarde des articles, l'archivage des articles de presse gratuitement mis à disposition par internet est réglementé. L'accord des journaux concernés est nécessaire. Nous avons donc entrepris de contacter chacun des journaux concernés par la réalisation de notre corpus afin de leur demander l'autorisation de reproduire sur notre disque dur certains de leurs articles à des fins de recherche. Seul le journal *Le Figaro* nous a répondu (*cf Annexe A*), nous sommes sans nouvelle des autres journaux, malgré plusieurs relances.

En attendant la réponse des autres journaux, les articles sélectionnés sont conservés en local sur nos ordinateurs. Pour l'organisation de la sauvegarde, nous avons choisi une option simple et efficace étant donné le nombre réduit de fichiers. Ainsi, chaque article est stocké sous la forme d'un fichier xml nommé d'après le titre du journal duquel il est tiré, suivi d'un nombre à deux chiffres variant de "01" au nombre d'article tiré du dit journal. L'ensemble du répertoire est versionné grâce à un dépôt subversion, et les annotations sont stockées dans des sous-répertoires. La figure 2.1.3 illustre cette organisation.

[1] `http ://www.w3.org/TR/2006/REC-xml-20060816/`

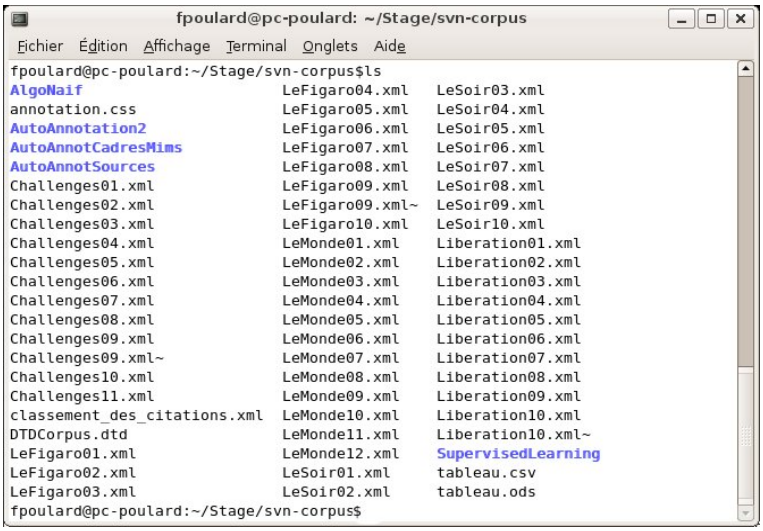


FIG. 2.1 – Organisation des fichiers du corpus

Une fois les articles collectés, organisés et nettoyés, il faut structurer le contenu de chacun des articles pour en faciliter l'intelligibilité.

Format et structuration

Afin de pouvoir utiliser au mieux le corpus, et notamment au delà de ce pour quoi nous l'avons conçu, il est nécessaire que chaque unité élémentaire du dit corpus soit "autonomisable" [Habert, 2001]. En d'autres termes, chaque article utilisé au sein du corpus doit être suffisamment décrit pour qu'on puisse "l'extraire de la base et l'assembler avec d'autres éléments de la même base ou d'autres bases".

De notre point de vue, il était nécessaire de considérer deux types d'information sur chacun des articles : le contenu de l'article et les informations relatives à ce dernier. Nous avons donc défini un format permettant de décrire le contenu d'un article ainsi que les méta informations le concernant. Nous exhibons en annexe l'extrait d'un article structuré selon ce format (cf Annexe D.1).

Chaque article est stocké dans un fichier XML, et chaque fichier se divise en deux parties. La première partie — appelée *metadata* — contient toutes les informations qui concernent la position de l'article au sein du corpus. La deuxième partie — *content* — correspond au contenu informationnel d'intérêt.

Structuration des informations relatives à l'article

Ce qui concerne le contexte de l'écriture de l'article ainsi que sa position dans le corpus est compilé dans la structure finale sous la balise parent <*metadata*>, frère de la balise <*content*> contenant la structure du contenu de l'article.

Les informations qui y sont représentées sont :

– le nom du journal ;
– l'url à laquelle l'article a été récupéré ;

– les auteurs de l'article ;
– la date de publication ;
– la dernière date de révision ;
– le lieu d'écriture de l'article dans le cas d'envoyés spéciaux.

Ces informations permettent ainsi, malgré un renommage éventuel des fichiers, de replacer l'article dans son contexte d'écriture, et récupérer éventuellement en ligne une version complète de l'article avec son environnement graphique.

Structuration du contenu de l'article

Les journaux se limitent bien souvent à mettre en forme le contenu de l'article, mais négligent la structure logique de l'information. Ainsi, bien que d'apparence l'article soit structuré grâce entre autre aux retours à la ligne et à la mise en forme, il n'y a pas de structure logique sous-jacente. Les éléments du texte tels que les titres, sous-titres et paragraphes ne sont alors pas marqués distinctement et sont déduits par le lecteur en fonction d'éléments typographiques tels que la mise en gras, la taille de la police ou encore les sauts de ligne.

La structuration du contenu a pour objectif d'expliciter les titres, sous-titres et paragraphes. De plus nous avons mis en place un système de blocs abstraits permettant de relier les sous-titres à leurs paragraphes. Le choix s'est porté sur l'utilisation de blocs regroupant entre eux les paragraphes liés et pour lesquels il est possible de définir un entête (sous-titre). La possibilité d'imbriquer les blocs permet ainsi de conserver la hiérarchie proposée par l'auteur au moyen de la mise en forme.

Il existe plusieurs standards de structuration des textes tels que le *Text Encoding Initiative* (TEI). L'utilisation de tels standards est un véritable plus pour la diffusion et la réutilisation du corpus. Toutefois, l'encodage de notre corpus dans ce format aurait demandé trop de temps par rapport à la durée du stage. De plus, les capacités du TEI, même pour la version "light", vont bien au-delà de nos besoins et ne semble pas réellement répondre à notre besoin en terme de structuration logique.

Le contenu d'un article se structure donc, selon notre format, en un titre éventuellement accompagné d'un épigraphe et suivi soit par des paragraphes — si aucun lien logique n'unie ces derniers —-, soit par des blocs. Chacun des blocs contient à son tour un header — ce n'est pas obligatoire — ainsi que des blocs ou des paragraphes, et ainsi de suite récursivement.

Il est important de noter que l'on ne peut pas mélanger à un même niveau des paragraphes et des blocs. En effet, la présence de blocs induit des liaisons logiques entre les paragraphes qu'ils contiennent. Si certains paragraphes sont liés logiquement, alors les paragraphes de *même niveau* le sont également, ne serait-ce que par le fait qu'ils n'appartiennent pas aux autres structures logiques. Il est toutefois concevable de créer un bloc ne contenant qu'un seul paragraphe.

La proposition de structure a fonctionné parfaitement excepté pour les quelques articles — dénombrés à 3 — contenant des listes. L'introduction d'une structure de liste semblant complexifier sans réelle justification la structure des articles, le choix a été fait de considérer les listes comme des blocs logiques et les entrées de ces listes comme des paragraphes. Ce choix ne semble pas réellement influé puisque l'on ne trouve pas de citations au sein de ces listes.

Au delà de la structure logique, certaines informations typographiques présentes dans le HTML nous semblaient importantes à conserver, notamment la mise en italique, en gras et les emphases. Nous avons l'intuition que ces mises en exergues peuvent être utiles dans le repérage des segments citationnels. Cependant, afin de bien séparer la structure logique des éléments de mise en forme, chacun a été placé dans un espace de nom (*namespace xml*) attribué. Ce classement permet ainsi de filtrer facilement les

Journal en ligne	Nb. articles	Nb. citations	Nb. mots	Citations/mots
Libération	10	81	5555	14.10^{-3}
Le Soir	10	44	5607	7.10^{-3}
Challenges	11	87	11332	7.10^{-3}
Le Monde	12	82	5957	$14.10-3$
Le Figaro	10	70	5305	$13.10-3$
Totaux	53	364	33756	

TAB. 2.2 – Répartition des citations par journaux

marques de l'une ou l'autre des catégories et traiter les informations séparemment selon leur dimension.

La structuration des articles par le biais de balises XML nous permet ainsi d'accéder au contexte des phrases en les replaçant dans leur paragraphe, voire leurs groupes de paragraphes. Cette étape de structuration nous a permis de mieux appréhender la structure générale des articles et d'en extraire des informations quantitaties et qualitatives concernant la citation que nous décrivons dans la section suivante.

2.2 Analyse quantitative des citations du corpus

La constitution, le nettoyage et l'organisation du corpus permettent de faire émerger plus facilement l'information contenue dans ce dernier. Il nous a ainsi été possible de repérer manuellement l'ensemble des citations qui y étaient présentes. Dans les cas où il n'était pas évident de délimiter la citation, nous avons procédé empiriquement au cas par cas. Nous présentons une représentation linguistique permettant de résoudre ces problèmes dans la section 2.3.2. Nous présentons et analysons dans cette section les éléments statistiques que nous avons extrait. Nous décrivons dans un premier temps la distribution des citations au sein du corpus. Puis, dans un deuxième temps, nous nous intéressons aux utilisations des styles du discours dans le corpus.

2.2.1 Distribution des citations par article et par journal

D'après [Mourad & Minel, 2000], la citation est engagée. Nous faisons l'hypothèse que cet engagement soit variable selon les journaux et les journalistes. Nous présentons dans cette section les données statistiques que nous avons extraites dans l'objectif de caractériser les citations au sein de leurs articles.

La répartition des citations au sein des journaux est décrite par le tableau suivant :

Il est intéressant de constater que le ratio de citations par mots varie du simple au double sans intermédiaire. On ne peut toutefois tirer aucune conclusion sur un si petit échantillon. L'ordre de grandeur du ratio est d'environ 10^{-2}, soit une citation pour une centaine de mots. On peut raisonnablement dire que l'on a atteint la richesse espérée avec plus de 350 citations réparties au sein du corpus. Sur 53 articles, cela nous fait une moyenne de $5,6$ citations par article. Il faut toutefois noter que le nombre de citations par article peut fortement varier puisque l'on a un minima de 0 citation et un maxima de 31 citations. Les répartitions par article des citations est illustrée par le graphique 2.2.

La grande variation du nombre de citations par article, même pour les articles partageant une même charte éditoriale, nous laisse penser que le nombre de citations au sein d'un article est indépendant de quelconques règles journalistiques et dépend très certainement uniquement du sujet traité et des

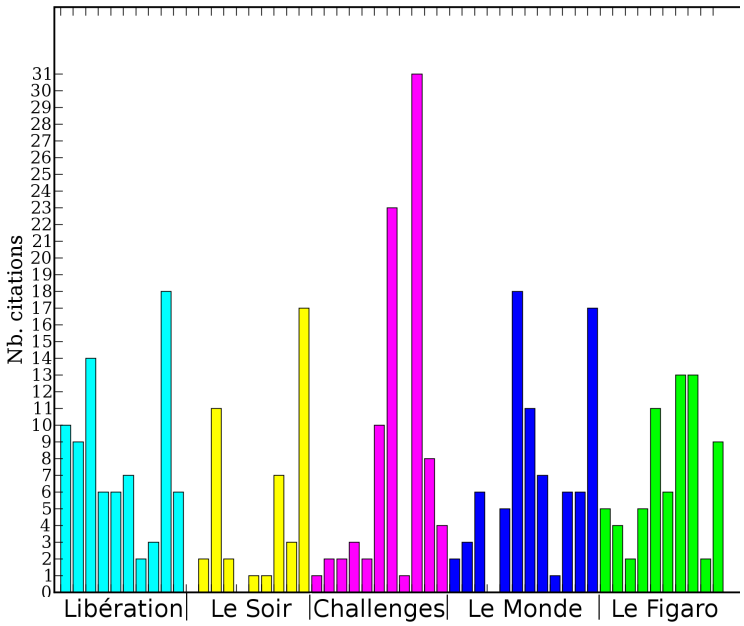


FIG. 2.2 – Répartition des citations par articles et par journaux

Journal	Citations/article	Amplitude	Variance	Écart type
Libération	8.1	16	21.49	4.64
Le Soir	4.4	17	28.44	5.33
Challenges	7.9	30	80.25	8.96
Le Monde	6.83	18	30.81	5.55
Le Figaro	7	11	16	4

TAB. 2.3 – Bilan de la répartition des citations par article

ressources disponibles. Pour nous en assurer, le détail de la répartition des citations pour chaque article est fourni en annexe (cf C).

Le tableau 2.3 résume les données des tableaux en annexe. Nous y avons compilé, par journal, les informations du nombre moyen de citations par article, la différence entre le nombre minimal et maximal de citations (amplitude), la variance du nombre de citations par article.

Les tableaux ne font apparaître aucun motif particulier permettant de modéliser la présence des citations au sein des articles. Si les grands journaux français semblent fournir en moyenne 7 à 8 citations par article, la répartition au sein des articles assez chaotique puisque l'on a pour Challenges par exemple une différence maximale de 30 citations entre deux articles. Si un modèle de répartition des citations au sein d'un article existe, la taille réduite de notre corpus ne permettra pas de le faire émerger.

En résumé, la présence des citations dans les articles ne semble pas dépendre du journal. Le nombre de citations au sein des articles varie de manière chaotique fluctuant au sein des journaux d'une citation tous les 70 à tous les 140 mots. Nous ne nous acharnons donc pas à rechercher un loi de répartition des citations au sein des articles car nous doutons fortement de son existence.

Dans la section suivante, nous nous intéressons aux emplois des différents styles de discours au sein

Citations	372	
Direct entre guillemets	218	58%
Indirect simple	56	15%
Indirect libre	15	4%
Îlots textuels	78	20%
Autres	5	1%

TAB. 2.4 – Répartition des style du discours dans le corpus

de notre corpus.

2.2.2 Distribution des styles de discours rapporté

Chacun des styles du discours possède sa structure et ses indices propres. En d'autres termes, la représentation linguistique de la citation sera différente si celle-ci est au style direct, indirect, . . ., pour une même situation d'énonciation initiale. Cela a pour conséquence immédiate de nécessiter une approche de détection différente selon le style de discours employé. Nous présentons dans cette section la distribution des styles de discours employés au sein de notre corpus. La connaissance de cette distribution nous permettra de nous focaliser sur la mise au point des approches de détection qui permettront de repérer le plus de citations.

Les journalistes prennent quelques libertés avec les règles définissant les styles du discours, de sorte qu'il ne nous est pas aisé de déterminer le véritable style de certains segments textuels. L'intérêt toutefois de connaître la répartition des styles est de pouvoir déterminer quels sont les moyens de repérage qui s'appliqueront au plus grand nombre de citations. Nous considérons ainsi les styles suivants :

- direct guillemets : le texte englobé est placé entre guillemets ;
- indirect simple : le texte englobé est placé dans une proposition subordonnée ou bien placée en incise ;
- indirect libre : le texte englobé ne se distingue du texte englobant que par l'interprétation du texte ;
- îlots textuels : le texte englobé est composé aussi bien de passages avec ou sans guillemet.

Les statistiques ci-dessous ont été extraites juste après la constitution du corpus. Nous considérions alors intuitivement qu'une citation consistait en passage de discours rapporté clairement défini et sa source introductrice. Les cas litigieux des bornes du discours rapporté étaient traités intuitivement au cas par cas. Pour cette raison, le nombre de citations recensées et présenté dans la table 2.4 peut varier légèrement par rapport aux autres chiffres auxquels nous ferons référence par la suite. Toutefois cette légère variation ainsi que l'approximation de ce que l'on considérait comme citation n'influe pas sur les conclusions que nous tirons de ces chiffres.

Sans trop de surprise, les styles utilisés pour les citations au sein du corpus privilégient l'objectivité, la neutralité et la séparation du plan de l'auteur et du locuteur. En effet, le texte englobé de 58% des citations est placé entre guillemets et pour 50% des autres citations, le texte englobé contient des passages placés entre guillemets (îlots textuels). Ces passages entre guillemets indiquent qu'ils sont tirés verbatims du discours original qui est rapporté. Rappelons-nous toutefois du bémol que nous avons soulevé : bien que les textes soient placés entre guillemets, il se peut qu'ils aient subi quelques modifications morphosyntaxiques, le fond reste toutefois identique.

Ces résultats sont encourageants car la plupart des textes englobés semblent respecter fidèlement le

discours original. De plus, les styles employés sont les plus riches en marques discursives, simplifiant l'étape de repérage. En négligeant les considérations concernant la délimitation des bornes et le repérage des sources, près de 80% des segments citationnels devraient être relativement faciles à repérer au sein de notre corpus. Sans vouloir extrapoler sur l'ensemble des textes journalistiques, nous pouvons être confiants pour la suite de notre travail.

La catégorie "Autres" correspond à des segments que nous avons marqués comme citationnels mais pour lesquels nous n'avons pas réussi à prendre de décision quant à leur classement. Nous avons donc préféré les laisser non classés. Nous étudions ces cas problématiques dans la suite de cette section.

> Washington avance une estimation des réserves mondiales "ultimes" de pétrole à 2 275 milliards de barils.
>
> © *Le Monde – 20 Février 2007*

> Elle était bien l'"organisatrice" du concert. Ce concert était une activité de "service public". Les agents qui ont commis des fautes disposaient d'un "pouvoir de représentation" de la ville.
>
> © *Libération – 22 Février 2007*

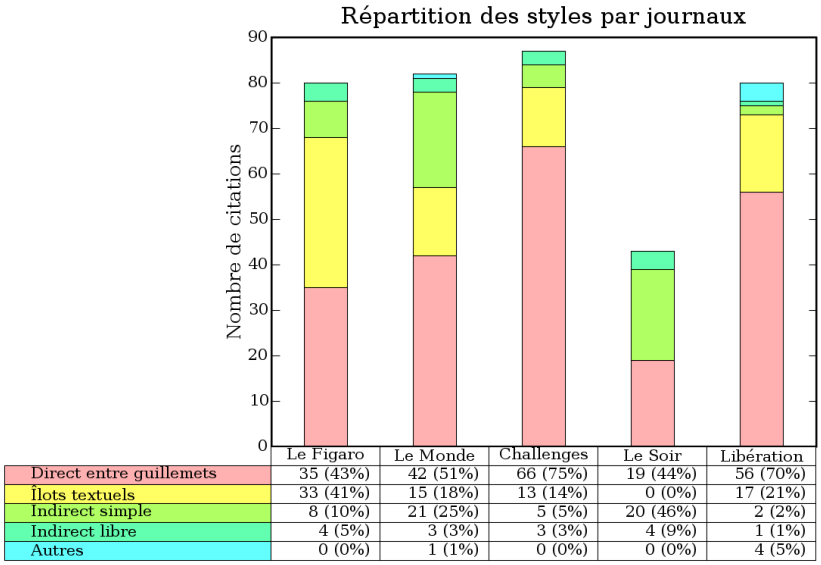
L'utilisation de guillemets permet de rapporter un discours au style direct, ou bien de mettre l'emphase sur un segment textuel. Dans les extraits ci-dessus, un seul mot à chaque fois a été placé entre guillemets. Nous pencherions pour une forme de discours indirect avec îlot textuel, mais la structure des phrases ne s'y prête pas forcément. De plus, aucune marque évidente ne permet de décider. Seule l'interprétation de la phrase dans son co-texte permet de nous faire pencher pour une forme de citation.

Ces cas montrent que l'utilisation abusive des guillemets peut finalement semer le doute sur leur rôle dans la tête du lecteur. Étant donné que ces cas sont isolés, qui ne représentent que 1% des citations du corpus, nous les ignorerons dans la suite de cette étude.

Le graphique de la figure 2.3, illustre la répartition des styles employés pour les citations est relativement semblable au sein des différents journaux. L'utilisation des guillemets pour encadrer le texte englobé ou bien certains mots au sein du texte englobé (îlots textuels) est majoritairement présent au sein de chacun des journaux. On peut noter l'absence dans le journal "Le Soir" d'îlots textuels. Ce journal Belge a en effet la particularité de rendre les citations très diffuses au sein du texte, ceci s'explique partiellement par le fait que les articles obtenus sur le site étaient souvent des brèves. Le journal "Le Monde", utilise plus le style indirect que les autres journaux. Cette caractéristique est peut-être liée à recherche de fluidité accrue dans l'écriture des articles.

En résumé, l'encadrement par des guillemets de l'ensemble du discours rapporté (style direct) ou bien d'une partie uniquement (îlots textuels) est la méthode la plus utilisée par les journalistes pour séparer les plans d'énonciation du locuteur et le leur. Le guillemet n'est toutefois pas une marque fiable de délimitation du texte englobé. En effet, comme toutes les marques de ponctuation, il est ambigu : il peut délimiter un fragment de texte rapporté ou bien mettre l'emphase sur son contenu : "ils [les guillemets] servent soit à inclure en subordonnant, soit à exclure en isolant" [Mourad & Desclès, 2002].

L'ambiguïté des guillemets et la difficulté, d'une manière plus générale, à repérer les bornes introduit une problématique plus vaste, celle de l'unité citationnelle au niveau linguistique. Nous abordons ce sujet dans la section suivante après avoir discuté des citations du corpus d'un point de vue qualitatif.



	Le Figaro	Le Monde	Challenges	Le Soir	Libération
Direct entre guillemets	35 (43%)	42 (51%)	66 (75%)	19 (44%)	56 (70%)
Îlots textuels	33 (41%)	15 (18%)	13 (14%)	0 (0%)	17 (21%)
Indirect simple	8 (10%)	21 (25%)	5 (5%)	20 (46%)	2 (2%)
Indirect libre	4 (5%)	3 (3%)	3 (3%)	4 (9%)	1 (1%)
Autres	0 (0%)	1 (1%)	0 (0%)	0 (0%)	4 (5%)

FIG. 2.3 – Répartition des styles d'intégration du texte englobé au sein des différents journaux

2.3 Analyse qualitative des citations du corpus

2.3.1 Marques et marquages des styles et formes du discours rapporté en corpus journalistique

L'observation des citations présentes au sein de notre corpus nous a permis de repérer certaines régularités concernant ces derniers. L'intérêt d'extraire ces marques et marquages est qu'ils vont nous permettre par la suite de trouver des méthodes adaptées aux différents types de citation présents au sein des textes journalistiques. Nous présentons dans les sections suivantes les marques et marquages repérés sur les citations au style direct, puis au style indirect.

Marques et marquages du style direct

Nous décrivons dans les paragraphes suivants les variations de structure des citations au style direct.

De par la rupture provoquée par l'ouverture et la fermeture des guillemets, le discours direct rigoureux se positionne le plus souvent en incise[2].

> "Avec la fin de la session, les parlementaires sont beaucoup plus disponibles. Cette réorganisation était donc nécessaire", renchérit-on dans l'entourage de Sarkozy
>
> © *Le Figaro – 20 Février 2007*

La citation s'articule, dans le cadre du discours direct, autour du texte englobé placé entre guillemets. L'introduction du locuteur gravite donc autour de ce segment entre guillemets. Nous pouvons noter que

lorsque le locuteur est introduit avant le segment entre guillemets, ce dernier sera souvent précédé d'un deux-points. Lorsqu'au contraire, le locuteur est introduit à la suite — vraisemblablement plus fréquent dans le domaine journalistique — il sera placé en incise à l'aide d'une virgule.

Une dernière variation possible, et assez fréquente, est le placement du locuteur au sein même du segment entre guillemets. Cette structure est notamment employée lorsque que le segment entre guillemets remplace une phrase complète au sein du texte englobant.

> "En 2003 **, explique-t-il,** j'ai fait effectuer douze tests sur des vols en France, dans onze cas sur douze, des armes et explosifs ont pu être introduits dans les avions".
>
> © *Le Figaro – 20 Février 2007*

En résumé, l'utilisation du style direct s'accompagne d'une phrase à incise. Les seules variations étant celles du positionnement de l'incise.

Marques et marquages du style indirect

Nous avons relevé deux structures pour le style indirect. L'utilisation de propositions subordonnées conjonctives dans un premier temps, puis les phrases à incise.

> Arnaud Montebourg, le porte-parole de Ségolène Royal, promet ainsi que, si la candidate de la gauche est élue, la construction de l'EPR ne serait pas interrompue.
>
> © *Libération – 22 Février 2007*

Cette citation au style indirect s'étend sur une seule phrase et met en oeuvre une seule source. On peut remarquer l'introduction du discours rapporté à l'aide d'une formule du type */Verbe d'énonciation*[3] *+ que/* caractéristique des propositions subordonnées. Le verbe utilisé est un verbe d'énonciation, il a la particularité de pouvoir être employé dans le sens de "dire", "énoncer". L'utilisation du présent pour le texte englobant limite les incidences sur les temps des verbes du texte englobé. Sans autre indice, le repérage de type de citation devra se baser sur :

1. la structure syntaxique de la phrase ;
2. le champs sémantique du verbe et son temps.

Il existe toutefois une autre structure de phrase qui ne fait pas appel à la formule *Verbe d'énonciation+que*. Cette seconde structure place l'introduction du locuteur en incise[4], comme l'illustre l'extrait suivant :

> D'après sa mère, Julien faisait une sieste dans l'appartement familial lorsqu'il a disparu.
>
> © *Le Figaro – 20 Février 2007*

Le repérage de ce type de citations semble plus complexe. Il est toutefois possible de modéliser la dite phrase par : */syntagme prépositionnel introduisant le locuteur + texte englobé/*. L'identification de ces phrases ne peut être réalisée seulement sur la présence d'une incise étant donné que ces dernières sont largement utilisées dans les textes. Il semble plus efficace de chercher dans un premier temps les

[3] ou intégrant un caractère énonciatif

[4] incise : proposition généralement de peu d'étendue et syntaxiquement indépendante, intercalée entre virgules dans le corps de la phrase ou rejetée à la fin de celle-ci.

syntagmes prépositionnels potentiellement introducteurs de sources. La constitution d'un dictionnaire de prépositions d'introduction de sources semble nécessaire, nous pouvons déjà y placer les termes *D'après* et *Selon* largement rencontrés dans le corpus.

Finalement, et pour nous compliquer la tâche quelque peu, on peut trouver des citations dont le texte englobé au style indirect s'étend sur plusieurs phrases :

> Edward Lu, physicien et ancien astronaute au centre Johnson de la NASA, a exposé les moyens envisagés pour repousser ces envahisseurs ! Première méthode : envoyer un petit vaisseau spatial de 1000kg pour aller impacter à la vitesse de 5km/sec l'astéroïde menaçant. L'énergie du contact serait plusieurs centaines de fois supérieure à l'énergie gravitationnelle qui colle ensemble les morceaux de l'astéroïde : celui-ci exploserait en morceaux de toutes tailles, à la queue leu leu sur la même trajectoire.
>
> © *Le Figaro – 20 Février 2007*

Le fait que le texte englobé au discours indirect puisse s'étendre sur plusieurs phrases nous indique qu'il faudrait, afin de le repérer, traiter les textes par segments de plusieurs phrases. Nous laissons cette considération en suspens pour le moment.

En résumé, deux structures apparaissent dans le cadre de citations employant le style indirect. La première découle de l'emploi de proposition subordonnées et s'articule autour du motif *verbe + que* où le champ sémantique du verbe contient une entrée à rapprocher de l'acte d'énonciation. La seconde structure, plus caractéristique du genre journalistique, s'articule autour d'un syntagme prépositionnel introduisant le locuteur et d'un texte englobé placé en incise dans la phrase.

2.3.2 Le problème de la délimitation de la citation

L'unité citationnelle au niveau linguistique est une problématique de taille pour la détection automatique. La détection automatique implique que l'on soit capable d'extraire une citation du texte. Pour rappel, ce stage s'inscrit dans le cadre du projet Piithie dont l'objectif est d'automatiser et aider à repérer des plagiats textuels. La reprise sous forme citationnelle n'est pas considérée comme un plagiat, nous devons donc pouvoir extraire les citations pour ne pas qu'elles soient considérées comme du plagiat. Cependant, nous ne sommes pas certain de ce qui constitue *une* citation : l'ensemble des éléments linguistiques du texte qui font parties de la citation et tel qu'on ne peut ajouter ou retirer un élément sans remettre en cause le statut de citation. Nous présentons dans les sections suivantes des situations exhibant cette problématique, puis nous tentons d'apporter des éléments de solution en nous basant sur le modèle de E. Giguet et N. Lucas.

Constatation du problème

Le problème sous-jacent à la définition d'une unité citationnelle est qu'il n'est pas toujours possible de déterminer avec exactitude les limites du texte englobé. Nous présentons dans les paragraphes suivants les cas problématiques, puis nous proposons une alternative au repérage d'unité citationnelle.

La problématique d'une unité citationnelle au niveau linguistique s'illustre particulièrement dans le domaine journalistique. La taille de l'énoncé à rapporter sous la forme d'une citation n'est pas satisfaisante pour un article journalistique. L'auteur doit donc y opérer des transformations. Afin de montrer son

objectivité, il conserve au sein de sa reformulation des passages verbatims. Le texte englobé se constitue alors d'une juxtaposition de segments reformulés et de passage verbatims (cf l'exemple 2.3.2).

> Le Figaro estime, lui, que "les techniciens et les cadres sont en première ligne", notamment ceux de la "Central Entity" de Toulouse.
>
> © *Le Monde – 19 Février 2007*

Les modifications peuvent être plus ou moins importantes, parfois une amorce de remise en contexte suffit (cf 2.3.2 : "L'administration ne peut"). Il arrive cependant que les segments originaux soient noyés au coeur d'une totale reformulation au style indirect (cf 2.3.2).

> L'administration ne peut "utiliser la convocation à la préfecture d'un étranger [...] pour faire procéder à son interpellation en vue de son lacement en rétention", estime-t-elle.
>
> © *Libération – 22 Février 2007*

> L'avionneur européen a indiqué, lundi, que le conseil d'administration de sa maison mère EADS a "interrompu ses travaux" et se réunira "dans les prochains jours" pour tenter de trouver un accord concernant la répartition de la charge de travail liée à l'A350XWB, le futur long-courrier de l'avionneur européen.
>
> © *Le Monde – 19 Février 2007*

Dans certains cas extrêmes, l'auteur tisse des paragraphes entiers reprenant de brefs éléments de discours d'une source. Les modifications de l'extrait à intégrer sont alors si profonds qu'il y a de fortes raisons de penser que le texte englobé et le discours source ne possèdent plus que le sens profond en commun, les détails étant complètement passés sous silence (cf 2.3.2).

> Royal veut aussi créer "une nouvelle génération de dispensaires" pour un meilleur accès au soin, "remettre à niveau" en personnel les hôpitaux publics qui manquent de bras, et elle est contre la fermeture des hôpitaux, qui doivent servir à accueillir les personnes âgées. Sans oublier "la santé gratuite pour les jeunes". Et pour cela, il faudra "desserer le numerus clausus" pour former plus de médecins.
>
> © *Le Figaro – 20 Février 2007*

D'après les exemples rencontrés dans notre corpus, la taille de ce que nous avons considéré comme citation est bornée à minima par le mot et à maxima par le paragraphe. Nous n'avons en effet pas trouvé dans la presse de citations s'étalant au-delà du paragraphe. Plusieurs exemples ont toutefois illustré l'extension d'une citation au-delà de sa phrase d'origine.

La problématique d'aligner le concept de citation sur sa représentation linguistique afin d'extraire une séquence de mots comme *citation* n'est pas résolue. Nous savons toutefois borner au sein du texte l'étendue de cette séquence de mots. Ainsi, dans un texte journalistique, la représentation linguistique de la citation se situe entre l'unité "mot" et l'unité "paragraphe".

Solution partielle avec la séquence canonique de la citation

La formalisation de E. Giguet et N. Lucas [Giguet & Lucas, 2004] se compose de deux éléments :

– la séquence canonique de la citation ;
– le modèle pour la langue française.

Le premier élément nous semble particulièrement intéressant pour apporter des éléments de réponse à la problématique d'unité citationnelle.

Pour rappel, [Giguet & Lucas, 2004] proposent de représenter les citations comme des séquences des trois objets citationnels qu'ils ont identifiés :

– la source : "regroupe le nom propre du locuteur et éventuellement ses qualifiants" ;
– le discours rapporté : correspond à ce que nous appelons le *texte englobé* ;
– le relateur : est le "segment établissant la relation entre la source et le discours rapporté".

La formalisation de Giguet et Lucas pose simplement les choses en abstrayant complètement la partie concernant le discours rapporté et en intégrant la source à la citation. Ces abstractions permettent de prendre du recul sur la notion de citation et nous reconcentrer sur la notion d'unité citationnelle.

En prenant comme modèle la séquence canonique introduite par [Giguet & Lucas, 2004], nous avons considéré dans un premier temps comme unitaires les segments citationnels qui partagent le même locuteur. Rappelons que le locuteur correspond à l'expression textuelle employée par l'auteur pour faire référence à la source. Ce choix permet de regrouper sous une même citation les textes englobés qui ont tous été introduits à l'aide du même locuteur. De ce point de vue, si l'on considère le locuteur comme un point d'ancrage du texte englobé — comprenez comme "une borne potentielle" — de l'unité citationnelle, alors la détection des frontières de la citation ne concerne plus que la frontière qui n'est pas contigüe à cette expression locuteur. En effet, selon la structure de la citation, le locuteur est séparé du texte englobé par des éléments de ponctuation ou bien par un relateur. Dans les deux cas, ces éléments sont identifiables. Leurs bornes étant alors définissables, celle du texte englobé qui s'y rattache l'est également par déduction. Le problème de repérage de la frontière non contigüe au locuteur reste lui entier.

Le problème d'unité citationnelle n'étant que partiellement résolu par le repérage du locuteur, nous avons alors tenté de reprendre l'idée concernant le modèle citationnel, également introduit par Giguet et Lucas [Giguet & Lucas, 2004]. L'idée est de définir des motifs d'agencement des objets citationnels correspondant à la structuration des citations dans les articles journalistiques. Giguet et Lucas proposaient dans leur papier les deux modèles :

– source + relateur + discours
– discours + relateur + source

En tentant d'appliquer la formalisation aux citations de notre corpus, nous nous sommes aperçus que ces deux motifs étaient bel et bien majoritaires puisqu'ils représentaient respectivement 104 (28%) et 148 (40%) occurences. Ainsi, près de 70% des citations de notre corpus correspondaient à l'un de ces deux motifs. Les 30% restant ne sont tout de même pas négligeables.

La première situation pour laquelle les motifs ne s'appliquent pas se caractérise par une citation dépourvue de relateur ou de locuteur. On rencontre ce type de cas particulier lorsque le locuteur est très saillant tout au long de l'article, ce qui est notamment le cas lorsqu'il n'y a qu'une seule source. L'extrait ci-dessous illustre ce genre de cas, représentant près de 9% des citations du corpus.

> Alors Baloua accepte tout, les mois sans jour de repos, les heures sup pas payées, les salaires en dessous des minima. «On est juste une main d'oeuvre moins chère. Ils t'exploitent. Les patrons disent qu'ils ont trop de charges, des dettes. Alors ils piquent à nous, les plus pauvres. On est des victimes.»

La seconde situation pour laquelle les motifs ne s'appliquent pas correspond à des motifs que Giguet et Lucas semblent avoir négligés. Ainsi, l'extrait suivant illustre le motif *relateur + source + discours* qui semble particulièrement approprié aux citations où le syntagme prépositionnel introduisant le locuteur et le texte englobé sont placés en incise. Ces citations représentent également environ 9% du corpus.

> Selon eux, «beaucoup [des saisonniers OMI] auraient bénéficié de CDI en d'autres temps. Relativement qualifiés, ils reviennent régulièrement dans les mêmes exploitations. On dit même de certains que ce sont les véritables chefs d'exploitation».
>
> © *www.liberation.fr – 22 Février 2007*

On trouve également le motif *(discours + relateur + source + discours)* à hauteur de 6%. Le reste des citations correspondant à quelques autres variantes moins fréquentes de la distribution des objets citationnels ou bien au doublement du locuteur, chose fréquente lorsque la citation elle même est reprise d'un autre article ou d'une agence de presse.

L'apparition de nouveaux motifs venant compléter le modèle de Giguet et Lucas réduit malheureusement l'efficacité de leur algorithme basé sur des déductions positionnelles. Nous ne pouvons toutefois pas négliger ces cas particuliers qui représentent tout de même 30% de notre corpus. En résumé, la formalisation de Giguet et Lucas permet partiellement de résoudre le problème d'unité citationnelle définissant comme *une* citation, la séquence comprenant le locuteur et tous les textes englobés qui s'y rattachent. Le problème de bornage des dits textes englobés restent cependant entier, ne permettant pas la mise en place d'une méthode robuste de détection.

Cette formalisation nous semble pourtant être une avancée importante dans la recherche d'une méthode robuste de repérage des citations. Nous choisissons donc de l'adapter d'après les remarques que nous avons soulevé précédemment. Un segment citationnel peut se représenter sous la forme d'une séquence canonique composée, en autorisant les répétitions, des objets citationnels suivants :

- un texte englobé ;
- un locuteur qui peut être omis si la source est suffisamment saillante tout au long de l'article ;
- un relateur qui se présente sous la forme d'une expression textuelle, d'une forme ponctuation ou bien qui est induit par la structure de la phrase et peut donc être omis également.

Toutes les permutations avec répétitions de ces objets citationnels sont possibles.

La structure des objets citationnels *relateur* et *locuteur* est discutée dans les sous-sections suivantes.

2.3.3 Caractérisation des objets citationnels : relateurs et expressions locuteur

Les expressions relateur et les expressions locuteur ont un rôle à jouer non négligeable dans la mise en place d'un algorithme de détection automatique des citations. Associables aux invariants "source" et "relateur", elles sont indissociables de la séquence canonique de la citation proposée par N. Lucas et E. Giguet [Giguet & Lucas, 2004]. Afin de mieux cerner ces formes linguistiques forts mystérieuses, nous avons tenter de les caractériser. Nous présentons donc les caractéristiques extraites de notre corpus concernant les expressions relateur d'abord, et les expressions locuteur ensuite.

Caractéristiques des expressions relateur

L'étude au sein du corpus de l'objet citationnel *relateur* nous a permis de cerner un peu plus cet élément polymorphique au coeur de la relation entre la source et l'acte d'énonciation rapporté. Nous décrivons dans les paragraphes suivants, les caractéristiques linguistiques des expressions relateur, puis

leur distribution au sein du corpus.

En premier lieu, la taille des expressions relateur oscille entre un unique mot et plusieurs subordonnées. Celà se répercute sur leur complexité structurelle. Ainsi, les expressions relateur peuvent se présenter sous la forme d'une simple préposition telle que *selon* ou *d'après*, ou bien comme un composition de verbes et de compléments. Sur les presque quatre cents citations du corpus, nous avons dénombré plus d'une centaine d'expression relateur différentes. Il faut toutefois relativiser ce chiffre par l'aspect exploratoire du dénombrement qui peut certainement s'affiner pour regrouper certaines formes sous une même étiquette.

La variation des relateurs n'est toutefois pas indomptable puisque près de 50% de ceux rencontrés au sein de notre corpus sont des formes verbales simples : verbes d'énonciation ou bien formule *verbe + que*. Suivent par fréquence d'apparition les prépositions *selon* et *pour*. Enfin, les expressions relateur restantes se divisent en deux grandes classes :

- les formes verbales : on y trouve des verbes associés à des indications spatio-temporelles, des prépositions ou encore des expressions d'altération du sens du verbe ;
- les formes adverbiales : expressions composées d'adverbes.

Le détail de l'exploration est fournie dans l'annexe C.

Le polymorphisme des relateurs en font un objet citationnel particulièrement difficile à appréhender. G. Mourad a compilé au sein de [Mourad & Minel, 2000] près de 700 verbes d'énonciation pouvant être utilisés comme relateurs. La taille de ce dictionnaire, ne représentant qu'une partie des relateurs potentiels, nous encourage à délaisser pour le moment l'approche lexicale. En effet, l'augmentation de la taille du dictionnaire a provoqué d'après G. Mourad une augmentation non négligeable de la détection de faux positifs.

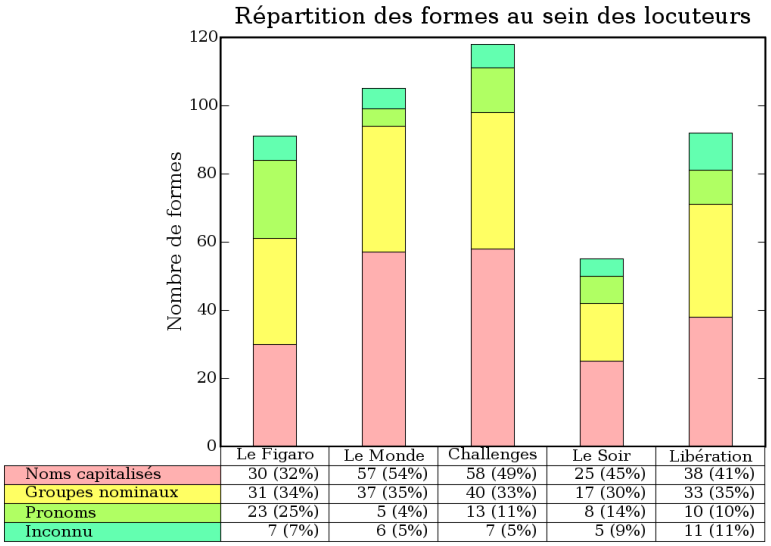
Caractéristiques des expressions locuteur

Les expressions locuteur sont très importantes au sein d'une citation. L'identification de la source est nécessaire dans le cadre du projet Piithie, et dans celui du repérage de citations en général. Nous avons donc tenté de caractériser les locuteurs à partir des exemples que l'on pouvait trouver au sein de notre corpus.

À l'instar des relateurs, les locuteurs sont très hétérogènes. Près d'un quart de ces derniers sont constitués de plusieurs mots. La forme la plus fréquente est constituée d'un nom capitalisé auquel est associé un groupe nominal descriptif.

> "Nous avons surexposé le PC au public", raconte Mark Brailey, directeur du marketing pour l'Europe.
>
> © *Challenges – 21 Février 2007*

Les groupes nominaux et les noms capitalisés sont les entités grammaticales les plus utilisées dans les expressions locuteurs. Toutefois, les auteurs emploient ces formes très précises lorsqu'ils présentent la source pour la première fois au sein de leur texte, puis utilisent ensuite des formes plus réduites. Ce phénomène avait auparavant été décrit par Giguet et Lucas [Giguet & Lucas, 2004] : les "réductions lexicales et les anaphores auront tendance à apparaître après des formes très déterminées".



	Le Figaro	Le Monde	Challenges	Le Soir	Libération
Noms capitalisés	30 (32%)	57 (54%)	58 (49%)	25 (45%)	38 (41%)
Groupes nominaux	31 (34%)	37 (35%)	40 (33%)	17 (30%)	33 (35%)
Pronoms	23 (25%)	5 (4%)	13 (11%)	8 (14%)	10 (10%)
Inconnu	7 (7%)	6 (5%)	7 (5%)	5 (9%)	11 (11%)

FIG. 2.4 – Répartition des formes au sein des locuteurs

> Brigitte Liberman, la directrice générale de Cosmétique Active (La Roche-Posay, Vichy), rétorque : “Les grandes innovations ne peuvent pas naître chaque année.” [...] Mais elle admet que “Jean-Paul nous poussant, nous pouvons faire encore mieux”.
>
> © *Challenges – 21 Février 2007*

Même si des formes réduites sont utilisées après les formes très déterminées, ces dernières dominent largement, puisque les noms capitalisés et les groupes nominaux sont représentés 732 fois dans le corpus en tant qu'éléments constituant un locuteur. Nous avons disséqué chaque locuteur trouvé au sein de notre corpus en "formes constituantes" : noms capitalisés, groupes nominaux, pronoms, et recensé chacune de ces formes.

Nombre de citations	372	
Noms capitalisés	208	55%
Groupes nominaux	59	15%
Pronoms	158	42%
Aucun locuteur	72	9%
Non classé	5	1%

La répartition des formes au sein des journaux est à peu près constante comme l'illustre le graphique de la figure 2.3.3.

Comme le montrent les résultats précédents, nous avons également tenu compte des cas où il n'y avait pas de locuteur dans le contexte phrastique du texte englobé, le lecteur devinant la source grâce à la saillance du dernier locuteur comme le montre l'extrait suivant :

> Il a insufflé une culture d'entreprise à la Kennedy. "Aujourd'hui, on pense d'abord à Philips, ensuite à son business, et enfin à soi."
>
> © *Challenges – 15 Février 2007*

Le pronom "Il" au début de la phrase précédente ravive la saillance concernant la source citée tout au long de l'article — à savoir ici, le patron de Philips — de sorte que, bien qu'aucun locuteur ne soit présent au sein de la phrase : "Aujourd'hui, [...] enfin à soi.", le lecteur est capable de relier le texte cité à sa source.

Lorsque la source est identifiable sans ambiguïté à partir du contexte — et pas seulement du co-texte —, les auteurs prennent parfois la liberté de ne pas expliciter le locuteur, comme dans l'exemple 2.3.3. Le lecteur faisant appel à ses connaissances globales sur le thème de l'article, l'identification de telles sources nécessiterait de la part de l'algorithme de repérage une acquisition de ce genre de connaissances. Nous laissons cette difficulté de côté pour le moment.

> "Nul ne peut être condamné à la peine de mort" : cet article unique du projet de loi constitutionnelle modifiera le titre VIII de la Constitution, consacré à l'autorité judiciaire.
>
> © *Le Figaro – 19 Février 2007*

Les formes employées pour les locuteurs sont assez variables, mais cloisonnées dans trois catégories : noms capitalisés, noms communs ou pronoms. De plus, l'emploi fréquent de plusieurs formes pour un même locuteur, notamment lors de la première référence à la source, augmente le nombre d'indices permettant son repérage. Les sources saillantes pour lesquelles les locuteurs ne sont pas introduits dans la phrase du texte englobé sont problématiques. L'emploi associé de résolveurs anaphoriques ainsi que le maintien d'une liste des sources les plus saillantes sont des solutions pouvant apporter des éléments de réponse à ce problème.

2.4 Schéma d'annotation des citations

Une première étape d'annotation expliquée précédemment a consisté à structurer logiquement les articles constituant le corpus. Nous avons ensuite prélevé au sein de ces articles des phrases dont nous considérions qu'elles contenaient des citations. Afin de permettre l'utilisation et la diffusion du corpus, il a fallu intégrer ces informations au sein du corpus à la main, aucun outil n'étant encore assez fiable pour le faire. La première section expose notre première tentative et les conclusions que nous en avons tiré pour finalement aboutir au format de la seconde section.

2.4.1 Première tentative d'annotation des objets citationnels

Découvrant la formalisation de Giguet et Lucas [Giguet & Lucas, 2004], nous avons tenté de l'appliquer telle quelle au corpus afin d'annoter les citations.

Schéma d'annotation

Nous avons fait le choix depuis le début de l'utilisation du métalangage XML pour structurer le corpus, et nous conservons ce choix pour ce qui est de l'annotation. Le premier avantage d'XML est qu'il est lisible et compréhensible par l'humain si l'on définit avec précision les noms des balises et des attributs. Étant donné que le XML est basé sur des balises textes, et qu'il supporte par défaut l'encodage

unicode, il semble prédestiné à la structuration des textes. Un autre avantage du XML, lié à sa popularité, est le nombre de bibliothèques performantes et libres qui sont disponibles pour la manipulation de ce type de fichiers. Cela nous permet d'économiser un temps précieux lors du traitement automatisé des fichiers. Finalement, le XML étant largement utilisé par les autres chercheurs en TALN, il nous permettra d'échanger le corpus avec d'autres équipes ou d'intégrer d'autres corpus au notre.

L'annotation reprend la séquence canonique de la citation. Elle se répartit donc en trois balises XML : un pour la source (<*source/*>), un pour le relateur (<*linker/*>) et un dernier pour le discours rapporté (<*speech/*>). La réunion en objet citationnel s'effectue grâce à l'attribut commun aux trois balises : *citation*. L'affectation d'un identifiant identique aux balises d'une même citation permet de regrouper les objets citationnels d'une même séquence canonique. Cela permet de reconstituer les citations une fois l'annotation terminée. De plus, afin de palier le problème d'éclatement des objets citationnels, notamment des relateurs qui peuvent s'étaler autour d'une expression locuteur, chaque balise est accompagnée d'un attribut *id* permettant de relier sous un même objet citationnel des segments textuels balisés. Finalement, afin de cloisonner l'annotation des citations de la structuration logique, nous avons choisi de placer les balises dans un espace de nom différent.

```
  <cite:source citation="1" degree="0" id="1">
    Gerard Kleisterlee
  </cite:source>
  <cite:linker citation="1" source="1" id="1">
    parle avec fiert\'e␣de␣ce␣site
␣␣</cite:linker>
␣␣cens\'e illustrer la r\'esurrection␣du␣groupe␣qu'il dirige depuis 2001 :
  <cite:speech citation="1" id="1">
    ``Il y a six ans, cet endroit \'etait␣entour\'e de grillages, peu de
    gens y avaient acc\`es... et peu de choses en sortaient.''
  </cite:speech>
```

Les balises *source* et *linker* possèdent des attributs supplémentaires. Ainsi la balise *source* s'accompagne de l'attribut *degree* particulièrement utile lorsque plusieurs locuteurs sont présents pour une même citation. Sa valeur correspond à la valeur supposée du degré de la source correspondante, à savoir 0 pour la source originale, 1 pour la source ayant rapporté la source originale, ...

L'attribut *source* de la balise *linker* permet de préciser à quel balise *source* il est rattaché. Cela est notamment utilisé lorsque plusieurs sources sont citées pour un même texte englobé.

```
  <cite:speech citation="2" id="2">
    ``L'orientation␣privil\'egi\'ee␣de␣l'enqu\^ete est d'ordre␣familial''
␣␣</cite:speech>
␣␣<cite:linker␣citation="2"␣source="2">
␣␣␣␣,␣indique␣mardi
␣␣</cite:linker>
␣␣<cite:source␣citation="2"␣degree="0"␣id="2">
␣␣␣␣une␣source␣proche␣de␣l'enqu\^ete
  </cite:source>
  <cite:linker citation="2" source="3">
    , cit\'ee␣par
␣␣</cite:linker>
```

```
␣␣<cite:source␣citation="2"␣degree="1"␣id="3">
␣␣␣␣␣l'AFP
  </cite:source>
```

Un exemple d'article annoté est présent dans l'annexe D.2.

Essai d'annotation sur un échantillon du corpus

Une fois le schéma d'annotation défini, nous l'avons essayé sur deux articles du corpus afin de tester son efficacité. Trois personnes se sont donc attelées à annoter les objets citationnels de ces articles. Aucune communication sur l'annotation n'a été effectuée entre les personnes jusqu'à ce que l'on compare les résultats. Le schéma d'annotation s'est révélé inadapté.

La séquence canonique *source+relateur+discours rapporté* fonctionne bien pour les citations dont le texte englobé est clairement délimité par des marques typographiques ou ponctuatives et tel qu'un locuteur introduise ou conclue la citation. En d'autres termes, l'annotation fonctionne correctement lorsque les séquence canoniques correspondent au modèle proposé par Giguet et Lucas [Giguet & Lucas, 2004] comme dans l'extrait ci-dessous :

> "Philips était très compartimenté, avec des métiers très différents. Les gens se rentraient dedans par hasard", raconte-t-il.
>
> ©*Challenges - 15 Février 2007*

Les choses se compliquent lorsque la structure de la phrase ne nous fournit pas de locuteur. Ainsi, dans l'exemple ci-dessous, l'on peut supposer que la source est *le tribunal administratif de Marseille*, sous la forme très certainement d'une communication par le biais d'un représentant officiel, mais aucun indice ne nous l'indique, et notamment pas un relateur.

> le tribunal administratif (TA) de Marseille a "enjoint au préfet des Bouches-du-Rhône de délivrer à M. Aït Baloua un titre de séjour" de dix ans dans les deux mois.
>
> ©*Libération - 22 Février 2007*

La difficulté croît avec la diffusion du texte englobé dans le texte englobant. En effet, les limites du texte englobé étant plus difficiles à définir, les balises sont proportionnellement plus difficiles à positionner dans le texte. L'éclatement du relateur autour du locuteur, ou l'abscence pure et simple de ce dernier compliquent d'autant l'annotation.

En résumé, le premier choix d'annotation se voulait trop optimiste sur notre capacité à définir distinctement les bornes des objets citationnels au sein des articles. L'expérience de l'annotation par différentes personnes a mis en valeur une définition différente des bornes par les différentes personnes sur certaines citations. Nous avons donc décidé d'adapter cette annotation en réduisant sa précision.

2.4.2 Schéma d'annotation retenu : le segment citationnel

Le premier schéma d'annotation se voulait trop complet pour être efficacement appliqué à notre corpus. Le relateur est par exemple un véritable problème par la grande variation de ses formes.

La première décision prise pour le nouveau schéma d'annotation du corpus fût de supprimer la balise *linker* destinée au repérage du relateur. Nous avons en effet décider de laisser de côté cet élément afin de nous concentrer plus particulièrement sur le repérage du locuteur et du texte englobé. La collection de balises s'est donc réduite aux deux balises :

- source : destiné à marquer les expressions locuteurs. Avec le recul le nom de la balise paraît mal choisi ;
- discours : destiné à marquer les segments de textes qui contiennent globalement la totalité du texte englobé ;

Les attributs de la balise *source* ont tous été supprimés à part l'attribut *id* qui permet d'identifier l'expression locuteur et éventuellement de segmenter l'annotation d'une expression locuteur. Cette capacité n'a cependant pas été extrêmement utilisé au sein du corpus.

Les attributs de la balise *discours* ont également tous été supprimés à l'exception cette fois de *source* qui permet toujours de relier le texte englobé à son expression locuteur associée. Il n'est plus question désormais d'annoter le texte englobé, mais plutôt de relier tous les segments de texte rapportés à l'expression locuteur à laquelle ils se réfèrent. Si le cas l'impose, il est possible de spécifier que le segment de texte n'est relié à aucune expression locuteur avec l'identifiant *?*.

La reconstitution de la citation selon que l'on considère comme telle la combinaison d'une expression locuteur et du texte qui y est rattaché s'effectue alors en trois temps :

1. réunification des balises sources de même identifiant ;
2. compilation des balises discours référant à la dite source ;
3. englobement des balises récoltés sous une forme approximative de la citation : le *segment citationnel*.

Ce schéma d'annotation, illustré par l'extrait ci-après, a parfaitement fonctionné pour l'annotation complète du corpus. Les prises de décision quant aux bornes du discours englobé sont toujours présentes, mais contournées par la possibilité de sélectionner au plus large lors de l'apposition des balises ainsi que la segmentation en fragments de texte englobé reliés à une même expression locuteur. L'approximation de la citation en *segment citationnel* nous a donc permis d'annoter la totalité des informations qui nous intéressaient au sein du corpus. Cette approximation ne semble pas avoir causé la dégradation de ces dites informations.

```
  <cite:discours source="1">
    ``La question est de savoir si l'\'economie a aujourd'hui␣un
␣␣␣␣probl\`eme␣de␣demande␣ou␣des␣difficult\'es du c\^ot\'e␣de␣l'offre''
  </cite:discours>
  ,  r\'esume
␣␣<cite:source␣id="1">
␣␣␣␣Lionel␣Fontagn\'e, professeur \`a Paris-I et membre du Conseil
    d'analyse␣\'economique
  </cite:source>
  , pour qui
  <cite:discours source="1">
    ``la plupart des \'economistes␣estiment␣qu'il y a d'abord␣un
␣␣␣␣probl\`eme␣d'offre''
  </cite:discours>.
```

Petit effet de bord ennuyeux, l'ajout des balises délimitant les locuteurs et le texte englobé a parfois entraîné un mauvais chevauchement avec les balises utilisées pour la typographie. Il nous a été nécessaire

de repositionner quelques un des balises de marquage typographique (italique, gras et emphase). Ces décalages se sont toutefois réduits à décaler la fermeture ou l'ouverture des balises de typographie avant ou bien après un signe de ponctuation afin de concorder avec l'ouverture ou la fermeture d'un des balises *discours*.

La simplification du format d'annotation des citations en s'appuyant uniquement sur les expressions textuelles concernant le texte englobé et les sources, nous a permis de compléter plus efficacement la phase d'annotation. De plus, en délaissant la reconstitution des citations à la charge de l'utilisateur du corpus, nous lui donnons un certain degré de liberté quant à ce qu'il veut considérer comme citation.

2.5 Synthèse

La constitution du corpus a marqué une étape importante de l'avancée du stage apportant une illustration concrète aux éléments théoriques récoltés lors de l'état de l'art. La cinquantaine d'articles, tirés de journaux francophones généralistes en ligne, s'est révélé riche d'un peu plus de 350 formes citationnelles. Après nettoyage et structuration en XML du contenu et des informations concernant les articles, il nous a alors été possible d'extraire de ce corpus les formes citationnelles afin de nous en faire une première impression statistique.

L'utilisation de guillemets s'est révélé la méthode la plus employée pour l'intégration des textes englobés, bien que cet encadrement ne soit parfois que partiel, sous la forme d'îlots textuels.

Nous nous sommes attachés, avant de nous lancer dans l'annotation du corpus, à déterminer ce à quoi pouvait correspondre l'unité citationnelle. Sans trouver de réponse claire, nous avons choisi de considérer comme *une* citation, le regroupement au sein du texte d'une expression locuteur et des segments de texte englobé qu'elle introduit.

La séquence canonique de la citation proposée par E. Giguet et N. Lucas [Giguet & Lucas, 2004], bien que posant certaines bases essentielles, ne permet pas de représenter une partie trop importante des citations de notre corpus. Nous l'avons donc simplifiée afin de l'utiliser pour intégrer au sein du XML les informations concernant les segments citationnels et ainsi préparer notre corpus pour les phases d'apprentissage à venir.

Chapitre 3

Méthodologie suivie au long du stage

Nous avons vu dans la section 1.3 que les méthodes de repérage automatiques existantes offraient des approches intéressantes, bien que ne répondant pas exactement à nos critères. Nous proposons une approche basée sur l'idée des espaces de recherche présentée dans la méthode de G. Mourad [Mourad & Minel, 2000].

Notre méthode se base ainsi sur la segmentation du texte en cadres offrant chacun un contexte particulier dans lequel nous recherchons des objets particuliers proches de ceux de la séquence canonique de la citation [Giguet & Lucas, 2004] : le texte englobé ou du moins des îlots textuels, ainsi que l'expression locuteur à y rattacher. Le repérage de la présence de ces objets se déduit de la cooccurrence de certaines marques. Les marques présentes dans plusieurs cadres et recoupent les indices proposés dans les méthodes précédentes. L'originalité réside ici dans l'interprétation différente qui est faite de la cooccurrence de ces indices selon l'espace de recherche auquel ils se rapportent.

Nous donnons dans un premier temps un apperçu de notre approche. Puis, dans un second temps nous présentons les différentes marques considérées pour l'identification des objets citationnels au sein des différents espaces de recherche. Nous rassemblons finalement les pièces du puzzle dans la dernière section où nous discutons les différents scénarios auxquels il est possible d'être confronté une fois l'analyse de la cooccurrence des articles effectuée.

3.1 Apperçu de notre approche

Nous proposons une approche basée sur la considération des marques propres à chaque objet citationnel dans des espaces de recherche distincts. Les sections suivantes rappellent les méthodes existantes en approfondissant l'aspect algorithmique de chacune, puis présente notre proposition d'algorithme. Finalement, nous décrivons les différents espaces de recherche que nous mettons en place pour notre méthode.

3.1.1 Recherche d'un algorithme robuste

Nous avons décrit brièvement dans les chapitres précédents (1.3), les deux grandes approches mises en oeuvre dans le cadre du repérage automatique de citations. Il ne semble pas y avoir de publications concernant l'efficacité des ces méthodes (ce qui est fortement ennuyeux pour se donner une idée de leurs performances respectives). Nous tentons cependant dans cette section de présenter les méthodes sous un plan beaucoup plus algorithmique, puis de montrer ce qui semble être leurs points forts et leurs points faibles.

Afin d'avoir une idée objective des performances, il serait nécessaire d'implémenter chacune des méthodes et de réaliser un ensemble de tests sur un corpus comme le notre. Malheureusement, la durée du

stage ne nous permet pas ce genre de choses, et nous devons nous contenter des suppositions énumérées ci-après.

Algorithme basé sur l'exploration contextuelle

La première approche est celle du laboratoire LaLIC. Cette dernière relève du travail de thèse effectué par G. Mourad [Mourad, 2001]. L'exploration contextuelle consiste en l'observation d'une grande quantité d'éléments linguistiques dans le but d'extraire des éléments caractéristiques de ces objets. Une grande partie du travail de thèse de G. mourad a donc été de compiler des éléments caractéristiques des citations qu'il a ensuite classés en trois catégories : les marqueurs typographiques, les marqueurs linguistiques et les marqueurs typographico-linguistiques à mi-chemin entre les deux catégories précédentes.

G. Mourad décrit avec précision sa méthode de repérage dans l'article [Mourad & Minel, 2000]. Cette dernière consiste en l'application de règles[1] constituées des éléments suivants :

- un en-tête qui permet de définir les espaces de recherche qui vont être utiles dans l'exécution de la règle ;
- un déclenchement qui décrit dans quelles conditions doit être lancée ;
- une liste d'indices qui ont un rôle dans l'exécution de la règle ;
- des conditions qui correspondent aux affirmations nécessaires à la validation de la règle et donc à l'exécution de l'action ;
- des actions qui sont exécutées si la règle s'applique.

L'auteur indique qu'il a implémenté dix règles de ce type qui permettent d'affecter l'étiquette "citation directe" aux segments textuels d'un texte passé à l'algorithme. Il ne décrit toutefois pas précisément la manière dont ces règles ont été écrites.

Lors de la recherche de citations au sein d'un texte passé à l'algorithme, ce dernier commence par construire une structure hiérarchique reflétant l'organisation structurelle du texte (sections, titres, paragraphes, phrases). Le segmenteur ne segmente pas à l'intérieur des guillemets afin de "garder l'autonomie de ces segments textuels". Cette structure est ensuite enrichie afin d'y modéliser les "chaînes de liage" — permet de résoudre les références anaphoriques — et "les cadres du discours" — univers d'énonciation dans lequel s'applique un contexte particulier [Charolles, 2000]. Cette structuration permet de rendre le texte intelligible à la plateforme *ContextO* utilisée par l'auteur. Il s'agit du prétraitement du texte.

Nous ne sommes pas certains de la chronologie et de l'enchaînement des étapes précédentes car elles ne sont pas précisément décrites dans l'article. Nous avons déduit l'enchaînement à l'aide des dépendances des étapes les unes par rapport aux autres.

L'algorithme repère au sein de la structuration hiérarchique extraite du prétraitement les indices compilés lors de la phase d'exploration du corpus. L'algorithme repère d'un côté les indices simples tels que les guillemets, entités nommées,... et d'un autre côté les marqueurs qui vont servir de déclencheurs des règles.

Lorsqu'un marqueur est repéré, on recherche les règles qui l'utilisent en l'ayant stipulé au sein de leur champ *déclenchement*. S'ensuit alors la vérification de la règle :

1. l'espace de recherche est isolé du texte et considéré par l'algorithme, l'ensemble du travail sera réalisé au sein de cet espace de recherche. L'espace de recherche est défini au sein de la règle par le champ d'*en-tête* ;

[1] Terme employé par G. Mourad

2. on extrait de cet espace de recherche les indices listés dans le champ *liste d'indices*. Cette sélection des indices d'intérêt est certainement effectué à des fins d'optimisation ;
3. on vérifie l'application des conditions du champ *conditions*, chacune des conditions se basant sur la présence, l'absence ou le positionnement d'un indice listé précédemment dans le champ *liste d'indices* au sein de l'espace de recherche ;
4. si toutes les conditions sont vérifiées, on applique les actions listées dans le champ *actions*. Elles décrivent l'étiquette à apposer, et la partie du texte concernée par cette apposition.

Finalement, la dernière étape consiste à extraire du texte en sortie de l'application des règles, les passages de textes où ont été apposées les étiquettes correspondant aux citations.

L'avantage majeur de cette méthode est de pouvoir modifier simplement l'algorithme en ajoutant ou retirant des règles. De plus, il est également possible au sein de chaque règle de modifier les conditions, les indices, ... En d'autres termes, la configurabilité de l'algorithme est un avantage indéniable. La définition d'un espace de recherche permet un traitement optimisé puisqu'on ne manipule qu'un extrait du texte, la projection sur un ensemble restreint d'indices permet également de considérer d'une manière générale un grand nombre d'indices, puis en considérer un nombre restreint au cas par cas. Toutefois, l'algorithme définit plus une plateforme de traitement qu'un réel repérage des citations. Nous n'avons ainsi pas d'information sur la constitution des règles sans lesquelles l'algorithme ne peut être implémenté.

Algorithme des invariants

L'approche du laboratoire GREYC, et notamment des chercheurs N. Lucas et E. Giguet se base sur la recherche d'invariants (source, relateur, discours rapporté) à partir exclusivement d'indices surfaciques. Ces indices sont classés en trois catégories : typographiques, morpho-syntaxiques et positionnels. Ces éléments de surface n'ont que peu de valeur pris un par un, mais les chercheurs misent sur leurs cooccurrences et leurs positions relatives pour exprimer des indices de citation.

D'après les auteurs [Giguet & Lucas, 2004], la stratégie lexicale, basée sur l'utilisation de listes, est contradictoire avec un système de détection fiable et robuste pouvant s'adapter à la grande variabilité des formes. Ainsi les formes lexicales rares alourdissent les listes sans apporter d'augmentation significative de l'efficacité des systèmes, entraînant même parfois une augmentation sensible de la détection de faux positifs. Dans cette optique, ils choisissent de n'utiliser que des marques de surface afin d'identifier les invariants de la citation, "la source, le discours rapporté et un relateur", selon eux.

Parmis les éléments de surface utilisés, on retrouve des éléments typographiques tels que les guillemets, la virgule ou encore les mots capitalisés ; des morphèmes grammaticaux comme le suffixe "ent" ou la conjonction "que". Finalement, la position de ces éléments les uns par rapport aux autres participe à la prise de décision sur l'affectation des valeurs source, relateur et discours rapporté. L'énorme avantage de l'utilisation des marques de surface sur la méthode syntaxique précédente est qu'il n'y a pas d'étape de prétraitement nécessaire, les marques de surface sont directement repérées au sein du texte.

Une fois les éléments de surface repérés, les règles obtenues lors de l'apprentissage automatique sur le corpus sont appliquées et permettent de déterminer des valeurs pour les invariants. Les règles sont appliquées phrase par phrase, et l'on considère dans un premier temps les indices de chaque phrase indépendamment des indices des autres phrases voisines. Plusieurs cas se posent alors. Si trois valeurs cohérentes (source, relateur, discours rapporté) sont déterminées, les trois invariants sont identifiés au sein de la phrase. Si seulement deux valeurs sont identifiées, alors la troisième est déduite sur critère

positionnel en se basant sur le modèle proposé par les auteurs : *source + relateur + discours rapporté* ou bien *discours rapporté + relateur + source*. Ainsi, si l'on se trouve en présence d'un motif du type : *source + ? + discours rapporté*, on peut en déduire la présence du relateur entre les deux. Si toutefois les indices étaient insuffisants pour arriver à une déduction positionnels, il est possible de faire appel aux indices des phrases voisines afin de situer la citation dans un contexte plus large. Ces indices complémentaires ne semblent toutefois apporter des informations uniquement sur la résolution de l'invariant "source" : ce qu'ils nomment "chaîne de citation" et qui semble correspondre à une résolution des références anaphoriques.

L'utilisation des formes de surface est un avantage indéniable en terme d'efficacité de traitement. En effet, contrairement à la méthode précédente, il n'est pas nécessaire ici de réaliser de prétraitement visant à structurer le texte. De plus, la méthode a été expérimentée et semble fonctionner sur des textes en anglais et en français, ne nécessitant qu'une adaptation du modèle et une traduction des éléments de surface repérés. Les auteurs reconnaissent que cette approche passe sous silence un certain nombre de citations au discours indirect. Il nous semble également que le modèle proposé, limité à deux motifs, ne soit pas adapté à l'intégralité des cas de citations que nous avons pû rencontrer au sein de notre corpus. Enfin, l'approche des éléments de surface n'est pas aussi adaptable que la méthode précédente étant donné que les règles extraites sont globales et donc que l'apprentissage doit être complètement répété lors de la modification d'un critère.

Notre proposition : vision générale

Nous proposons de rapprocher les deux méthodes décrites en section 1.3 en trouvant un compromis entre la création de longues listes de formes introductrices de citations et la limitation des marques à des éléments de surface.

Nous reprenons l'idée de spécifier un espace de recherche proposé dans la méthode lexicale, mais nous l'adaptons en nous basant sur la proposition de la méthode syntaxique de rechercher des invariants. Ainsi, nous définissons un espace de recherche[2] spécifique à chaque invariant :

- le cadre propos correspondant à l'invariant "discours rappporté" ;
- le cadre phrastique pour l'invariant "source" ;
- le segment citationnel est un regroupement de cadres des types précédents liés les uns aux autres.

Comme nous l'avons discuté auparavant, nous n'essayons pas de repérer le relateur, et pour cette raison aucun espace de recherche ne lui est dédié. La mise en place de ces cadres permet de désambiguïser les différentes marques, en leur donnant un rôle différent selon le cadre auquel elles appartiennent. Notre approche se rapproche de la stratégie syntaxique en tentant de repérer les composants *locuteur* et *texte englobé* dans un premier temps avant d'extraire un *segment citationnel* selon l'extraction de ces deux "invariants".

Notre algorithme se divise donc en quatre étapes successives :

1. segmentation du texte en constituants élémentaires de la citation ;
2. repérage des indices au sein de chacun des constituants ;
3. étiquetage des constituants d'après la coocurrence des indices interprétés selon leur positionnement et le constituant au sein du quel ils s'inscrivent ;

[2]Nous employons le terme cadre au lieu d'espace de recherche. Cependant ce choix est discutable puisqu'il ne fait pas référence aux cadres du discours tel que définit par [Charolles, 2000]

4. extraction des segments citationnels selon l'étiquetage des constituants et leur position dans le texte.

Notre algorithme se veut un bon compromis entre la stratégie lexicale "tout liste" et la stratégie syntaxique "tout règle". Nous interprétons le rôle d'indices — issus des travaux précédents — en fonction des segments du discours, et donc du co-texte, dans lesquels ils se positionnent, nous permettant de considérer les segments comme prenant part à du texte englobé, ou bien à une expression locuteur. L'approche en quatre étapes indépendantes nous permet une certaine adaptabilité, certes pas aussi bonne que celle de la méthode lexicale, mais meilleure que l'approche syntaxique puisqu'il est possible de jouer sur la définition des constituants, des indices considérés et de la façon d'extraire les segments citationnels. De plus, il nous est possible d'ajouter des constituants ou bien des étapes sans devoir recommencer l'apprentissage pour les étapes précédentes.

3.1.2 Définition des cadres pour notre approche

Nous interprétons les indices guidant le repérage de segments citationnels différemment selon le co-texte dans lequel ils se trouvent. Ainsi, nous considérons comme cadre, un bloc contigu de texte aux frontières clairement définies respectant les caractéristiques d'un des cadres énumérés ci-après.

Nous limitons l'étendue de ces cadres au paragraphe. Une fin de paragraphe entraîne donc la clôture de tous les cadres encore ouverts.

Cadres mimétiques

Selon Platon [Platon, 360 av J C], la *mimésie* est une forme de discours oratoire où l'auteur "s'efforce de donner l'illusion que ce n'est pas lui qui parle". Le parallèle entre la mimésie et le discours rapporté placé entre guillemets par l'auteur est intuitif. En effet, par la marque des guillemets, l'auteur prend une certaine distance avec le texte, il montre que les mots ne sont pas de lui. Nous supposons que les cadres mimétiques candidats contiennent tous les énoncés rapportés au style direct ainsi que les îlots textuels.

Pour rappel, l'auteur est l'entité à l'origine du texte écrit que nous appelons *texte englobant*. La source est à l'origine de l'énoncé capté par l'auteur — au travers d'éventuels intermédiaires — et que ce dernier rapporte : *le texte englobé*. Nous ajoutons à cette terminologie notre vision de deux autres termes. Le *discours rapporté* est un extrait des paroles (ou des écrits) de la source retranscrites par l'auteur, alors que le *propos rapporté* est un extrait des idées énoncées par la source et retranscrites par l'auteur.

Les cadres mimétiques se définissent comme des segments textuels délimités par des guillemets appariés. Leur taille varie d'un mot à plusieurs phrases. Les exemples 3.1.2 et 3.1.2 illustrent cette notion. Les cadres mimétiques y sont délimités à l'aide de paires de crochets numérotés.

> Les élus UMP n'en reconnaissent pas moins que $[^1$ **“** le souvenir de 2002 (l'élimination de Lionel Jospin au premier tour) est toujours là **”** $]^1$, prêt à provoquer $[^2$ **“**une mobilisation de dernière heure de l'électorat de gauche **”** $]^2$.
>
> © *Le Monde (LeMonde03)*

[1] “ Un ami m’a téléphoné de Hawaï pour m’informer que Converse cherchait un distributeur français. Moi, j’avais remarqué que les Japonaises branchées avaient des Converse aux pieds. Tout le monde m’a déconseillé d’y aller, mais je croyais au retour du vintage. ” [1]

© *Challenges (Challenges05)*

Les cadres mimétiques se composent exclusivement de discours rapportés, placé entre guillemets, et pouvant s’étendre sur plusieurs phrases. Tout ce qui est entre guillemets ne relève pas forcément d’un espace mimétique, c’est ainsi le cas des emphases (segments mis en valeur par un placement entre guillemets). Cependant, d’un point de vue opérationnel, nous choisissons dans un premier temps de délimiter des cadres mimétiques candidats. Il s’agit de cadres mimétiques potentiels, mais qui peuvent également s’avérer être de simples emphases. Plusieurs couples de symboles peuvent être considérés comme des guillemets, la norme unicode[3] définit notamment les symboles suivants :

- quotation mark
- apostrophe
- left-pointing double angle quotation mark
- single high-reversed-9 quotation mark
- double high-reversed-9 quotation mark
- heavy double turned comma quotation mark ornament
- heavy single turned comma quotation mark ornament
- heavy single comma quotation mark ornament
- reversed double prime quotation mark
- double prime quotation mark
- fullwidth quotation mark

Nous nous contenterons pour le moment des seules marques que nous avons rencontrées au sein du corpus, à savoir :

- guillemets typographiques français (chevrons imbriqués) ;
- guillemets typographiques anglais (doubles virgules hautes) ;
- guillemets typographiques allemands (double virgules basses en ouverture et hautes en fermeture) ;
- le guillemet droit ou guillemet dactylographique.

Contrairement aux trois premières marques qui sont asymétriques (symbole ouvrant et fermants distincts), le guillemet dactylographique est utilisé indifféremment pour l’ouverture et la fermeture. Cette marque, bien que typographiquement incorrecte, est encore largement utilisée, notamment sur internet.

Le caractère symétrique du guillemet dactylographique pose problème lors de multiples imbrications.

"On travaille sur plusieurs scénarios. Pour l’instant, nous n’avons que le but, qui est que les pays "moyens" soient mieux représentés. Le reste est à l’étude. Nous avons le temps : l’échéance, c’est la saison 2009-2010."

© *Le Monde (LeMonde03)*

Dans l’exemple ci-dessus, il n’est en effet pas possible de déterminer si le guillemet est fermant ou bien ouvrant sans "comprendre" le texte. Un algorithme trivial considérant l’alternance ouverture/fermeture des guillemets retournerait les deux espace de recherches candidats : "On travaille sur [...] que les pays " et " soient mieux [...] saison 2009-2010." Il n’existe pas à notre connaissance d’algorithme robuste permettant de traiter ces problèmes.

[3] http ://www.unicode.org

En résumé, les cadres mimétiques sont des segments de discours placés entre guillemets par l'auteur pour marquer le fait que celui-ci n'est pas le sien. D'un point de vue opérationnel, nous délimiterons dans un premier temps des espace de recherches mimétiques candidats en repérant les segments de texte placés entre guillemets.

Le discours rapporté ne se limite pas aux segments entre guillemets. Nous appelons *cadre du discours rapporté* les segments textuels contenant le discours rapporté. Ces cadres ne peuvent pas forcément être associés à chaque discours rapporté. En effet, tous les discours rapportés n'ont pas de représentation linguistique explicitement délimitées au sein du texte. Par déduction, pour des besoins opérationnels, les *cadres du discours rapporté candidats* correspondent aux segments textuels qui sont potentiellement des cadres du discours rapporté. Ceux-là englobent les cadres mimétiques candidats mais pas seulement. Malheureusement, il n'est pas possible de caractériser précisément les cadres du discours rapporté candidats et nous sommes donc contraints d'utiliser ce cadre sur un plan contextuel et non linguistique.

Cadres phrastiques

Le découpage en phrases est une étape importante de nombreuses applications de traitement du TAL [Mikheev, 2003]. Pour le repérage de citations, il est nécessaire de bien définir les limites de la phrase afin d'y englober complètement les cadres mimétiques candidats. Cela semble évident d'un point de vue linguistique, mais pose quelques problèmes d'un point de vue technique. Nous considérons le cadre phrastique comme l'espace de recherche dans lequel se situent les indices concernant l'expression locuteur et l'éventuel narrateur.

Les cadres phrastiques délimitent la portée de la recherche en premier lieu des sources linguistiques rattachées à un discours rapporté. En effet, lors du repérage d'un élément de discours rapporté, nous recherchons une expression linguistique — expression locuteur — à laquelle rattacher le discours. Celà permettra par la suite de rattacher les propos repérés à leur locuteur, et donc potentiellement leur source.

Le rôle du cadre phrastique dans notre approche est d'associer une expression locuteur à un discours rapporté. Il est donc nécessaire que la segmentation en cadres phrastiques ne sectionne pas les cadres mimétiques candidats. D'un point de vue linguistique, la phrase contient entièrement les passages entre guillement, même si ces derniers contiennent des marques de fin de phrase en leur sein. D'un point de vue technique, nous ne devons donc pas considérer comme marques de fin de cadre phrastique les signes de ponctuation terminaux (point, points de suspension, point d'exclamation, point d'interrogation, ...) présents au sein d'un cadre mimétique candidat. Seule exception à cette règle, les cas où le signe de ponctuation terminal correspond au dernier caractère — hors guillemet — du cadre mimétique. Dans ce cas particulier, la fermeture du cadre phrastique s'aligne sur la fermeture du cadre mimétique candidat.

> [1] Comme tous les fumeurs, en fait, analyse Odile Lesourne : [2] "La cigarette est un objet symbolique qui, pour aller vite, représente la mère des origines, bonne et chaude, qui réconforte mais qui agresse aussi, qui vous fait du mal à chaque fois qu'elle n'accède pas à votre désir de biberon, de ne pas être seul**...** En écho à ces frustrations de la toute petite enfance, on se constitue un Autre avec un grand A, véritable représentant de la mère archaïque**.**"]2]1
>
> © *Libération (Libe03)*

Dans cet extrait, le cadre phrastique 1 se termine après le guillemet fermant. En effet, les points de suspension (en gras dans l'exemple) sectionneraient le cadre mimétique candidat 2. Le dernier point (en

gras dans l'exemple), quant à lui, bien qu'au sein du cadre mimétique candidat, en constitue le dernier élément, hors guillemet. Il est donc le meilleur candidat à la fermeture du cadre phrastique. Nous nous trouvons en réalité face à deux contextes d'énonciation : pour le citant, il n'y a qu'une phrase délimitée par $[^1$ $]^1$, pour le cité, il y en a deux entre $[^2$ $]^2$.

L'ouverture d'un cadre phrastique peut être déterminé par les marques suivantes :

– fermeture d'un cadre phrastique précédent ;
– commencement d'un paragraphe ;

> [...] parrainages, être candidat**. ”** $]^i$ $[^j$ Selon lui, $[^k$ **“** les Français ont droit à un vrai et grand débat démocratique. La clé est entre les mains des maires de France **”** $]^k$ **.** $]^j$ $[^n$ $[^p$ **“** Ce sont eux qui, [...]
>
> © *Le Figaro (Figaro08)*

Ce passage montre les différentes marques permettant la segmentation du texte en cadres mimétiques candidats (délimités par $[^n$ et $]^n$) et en cadres phrastiques (délimités par $[^n$ et $]^n$). La fin du cadre phrastique *j* marque le début du cadre phrastique *j* qui se termine après le point suivant la fin du cadre mimétique candidat *k*. Nous soulevons également le problème de l'utilisation des marques ponctuatives du type point mais utilisées à d'autres fins. L'utilisation notamment du point pour les abrévitations provoque une segmentation en phrase erronnée. Dans le contexte du repérage de citation, le problème se complexifie à cause de la présence de deux situations d'énonciation. Le tokenizeur en phrase doit imbriquer correctement les cadres mimétiques candidats qui s'étendent sur plusieurs phrases, le discours extrait de la situation d'énonciation du cité n'est en effet qu'une parenthèse au sein de la situation d'énonciation du citant.

En résumé, le cadre phrastique permet de définir un espace de recherche correspondant à la phrase linguistique. Les cadres phrastiques commencent lorsque le précédent s'arrête ou bien lorsqu'aucun cadre phrastique n'est ouvert (comme en début de paragraphe par exemple). Les marques de fin de cadre phratique sont ponctuatives en prenant garde toutefois à ne pas considérer les marques de fin de phrase présentes au sein des cadres mimétiques candidats.

Segment citationnel

Le *segment citationnel* est défini de manière à englober au plus près la notion de citation : un locuteur rapportant le message d'une source. Le segment citationnel regroupe donc sous une même entité le discours rapporté ainsi que les éventuels relateurs et expressions locuteur rattachées à la source du discours rapporté.

Les frontières du segment citationnel sont déterminées par les frontières extrêmes des expressions locuteur, des éventuels relateurs et du discours rapporté, de manière à tous les contenir complètement. D'un point de vue opérationnel, un *segment citationnel candidat* correspond à un segment de texte de taille minimale et regroupant un *cadre du discours rapporté candidat* ainsi que les expressions locuteurs rattachées à ce cadre si elles sont repérées.

3.2 Identification des espaces mimétiques

Les cadres mimétiques candidats sont des segments de texte représentant potentiellement du discours rapporté placé entre guillemets, pouvant s'étendre sur plusieurs phrases. Le repérage des cadres mimétiques

candidats s'effectue par un découpage selon les guillemets ouvrants et fermants rencontrés. Le découpage en cadres mimétiques candidats est discuté dans le chapitre suivant. Nous traitons ici de l'importance des cadres mimétiques dans le processus de repérage des segments citationnels, ainsi que des indices utilisés pour l'apprentissage automatique.

3.2.1 Rôle des cadres mimétiques candidats

Les cadres mimétiques candidats correspondent aux segments du texte placés entre guillemets, ils correspondent donc potentiellement aux cadres mimétiques contenant du discours rapporté. Les guillemets étant les éléments de discrimination les plus utilisés par les journalistes pour séparer le plan d'énonciation de l'auteur de celui du locuteur extérieur, les cadres mimétiques candidats jouent un rôle prépondérant dans le repérage et l'identification des citations.

L'enjeu majeur dans la détermination des cadres mimétiques est de réussir à identifier les cadres mimétiques candidats correspondant en réalité à des emphases, ainsi que ceux correspondant réellement à l'intégration du texte englobé. Les "guillemets, comme tous les signes de ponctuation, sont ambigus" [Mourad & Desclès, 2002], ils ont pour rôle d'inclure en subordonnant ou bien d'exclure en isolant.

Les emphases représentent une part non négligeable d'utilisation des guillemets au sein des articles journalistiques et si elles ne sont pas correctement détectées, elles peuvent entraîner l'algorithme vers une fausse piste. Ainsi, l'utilisation de guillemets dans l'extrait ci-dessous correspond à une emphase mettant en relief le sujet de l'article, à savoir l'installation de fibres optiques par l'entreprise "Orange".

> Orange "fibre" la France à marche forcée
>
> © *Le Monde – 20 Février 2007*

Dans cet extrait, la structure de la phrase peut permettre de résoudre l'ambiguïté sur l'utilisation des guillemets. On trouve cependant, dans le même article, des constructions de phrase qui ne facilitent pas cette désambiguation en plaçant le segment entre guillemets en incise juste avant un verbe, ou bien faisant précédé l'extrait entre guillemets d'un verbe, lui même précédé d'une entité nommée.

> "La fibre", devrait rapidement être proposée dans quelques agglomérations de province.
>
> © *Le Monde – 20 Février 2007*
>
> Orange a choisi le "PON" (Passive Optical Network), qui permet de relier 64 prises sur une seule fibre.
>
> © *Le Monde – 20 Février 2007*

L'identification comme emphase ou cadre mimétique des cadres mimétiques candidats est une étape importante de l'algorithme de par la fréquence des segments entre guillemets et les conséquences sur l'efficacité de l'algorithme. La structuration de la phrase ne permettant pas toujours cette prise de décision, il est nécessaire de considérer d'autres éléments lors de la prise de décision. De plus, l'identification d'un cadre mimétique ne signifie nullement que le texte englobé se limite au cadre mimétique, il peut ne s'agir que d'un îlot textuel. Ceci relève toutefois du travail d'extraction des segments citationnels.

3.2.2 Indices sélectionnés

Nous préservons les décisions basées sur la structure des phrases pour l'étape finale de notre algorithme, à savoir l'extraction du segment citationnel. Nous nous concentrons donc sur des indices linguistiques et

morphologiques pour la prise de décision concernant la résolution de l'ambiguïté concernant les cadres mimétiques candidats.

Le premier indice considéré, et peut-être le plus évident, est la taille du cadre mimétique candidat. Les auteurs, journalistes notamment, se limitent à mettre l'emphase sur un nombre réduit de mots afin de ne pas perdre le lecteur. Si un cadre mimétique candidat composé de deux ou trois mots peut se révéler être un cadre mimétique, il y a peut de chance qu'un segment d'une demi-douzaine de mots se révèle être une emphase. La taille du cadre mimétique se révèle donc être un indice discriminant permettant certainement d'identifier des cadres mimétiques candidats comme n'étant pas des emphases, et donc par déduction de véritables cadres mimétiques. En considérant le fait que les emphases sont placées sur des mots, potentiellement accompagnés de déterminants ou d'adjectif, nous fixons la taille discriminante à trois mots.

Un indice clairement discriminant est la présence au sein du cadre mimétique candidat d'un autre cadre mimétique candidat. Il n'est pas sensé de placer une emphase ou bien un cadre mimétique au sein d'une emphase. La présence d'une telle configuration permet donc d'identifier directement le cadre mimétique candidat comme cadre mimétique. L'efficacité de cet indice est à relativiser toutefois par la fréquence assez faible des cadres mimétiques intégrant une emphase ou bien un autre cadre mimétique.

De la même façon qu'il est peu probable de trouver un cadre mimétique au sein d'une emphase, il est peu probable d'y trouver des incises de sectionnement du texte (*[...]*, *(...)*, ...) ou encore des commentaires placés entre parenthèses. Ainsi, les éléments précédemment énoncés sont considérés comme des indices permettant d'identifier un cadre mimétique candidat comme cadre mimétique.

Les pronoms à la première et deuxième personne sont caractéristiques d'un passage discursif, et donc potentiellement d'un discours rapporté au style direct. En tant que tel, les marques pronominales à la première ou deuxième personne, qu'elles soient personnelles (*je, tu, nous, vous, me, te, moi, toi*), ou bien adjectifs possessifs (*mon, ton, mes , tes, notre, votre, nos, vos, leur*). La présence de pronoms personnels aux deux premières personnes indique donc potentiellement un discours rapporté et donc un cadre mimétique. Ceci doit cependant être relativisé par le fait que cet indice n'a aucune valeur si l'article — fait toutefois assez rare — est écrit à la première personne.

L'écriture à la première personne peut être déjouée par une recherche, à l'extérieur du cadre mimétique candidat, de pronoms à la troisième personne. De plus, il est à noter que le placement en emphase d'un pronom est une chose assez rare, la recherche de pronoms à la troisième personne au sein d'un cadre mimétique candidat complète donc la recherche de pronoms entamée à l'aide de la classe d'indice présentée dans le paragraphe précédent. Nous choisissons toutefois de conserver les deux classes séparées étant donné que les pronoms aux deux premières personnes sont des marques discursives beaucoup plus significatives.

Les verbes d'énonciation, introduits par G. Mourad [Mourad & Desclès, 2001], marquent selon ce dernier l'introduction de discours rapporté. Nous considérons donc ces derniers comme indices, en réduisant toutefois la liste de Mourad. Nous préférons en effet supprimer les verbes ayant une trop forte polysémie sémantique, de peur qu'ils ne rapportent trop de faux positifs. Nous réduisons donc globalement la liste aux verbes se rapportant directement à un acte d'énonciation. Ces derniers sont listés en annexe E.

L'introduction du cadre mimétique candidat comme proposition subordonnée conjonctive à l'aide de la formule *verbe + que* laisse à penser que ce cadre mimétique candidat n'est pas une emphase. Les expressions respectant le motif *verbe + que* précédant les cadres mimétiques candidats sont donc considérés comme indices dans notre démarche de caractérisation des cadres mimétiques.

Finalement, les verbes étant généralement utilisés pour décrire une action, leur présence au sein d'un cadre mimétique candidat nous laisse penser que ce dernier traite d'une action. Le placement en emphase

d'une action n'est certes pas impossible mais toutefois assez improbable à moins de vouloir modérer la dite action. Nous considérons donc la présence d'un verbe comme une classe d'indice.

Les différentes classes précédemment décrites nous permettront par la suite d'extraire des règles de caractérisation des cadres mimétiques, et ainsi d'identifier le plus justement possible les emphases du discours rapporté au sein des cadres mimétiques candidats. Le processus de repérage de ces indices et l'apprentissage automatique qui en découle est discuté au chapitre suivant.

3.3 Identification des expressions locuteur

Les expressions locuteurs marquent plus ou moins explicitement la rupture entre le plan d'énonciation de l'auteur et celui de la source à laquelle fait référence l'expression. La présence d'une expression locuteur induit qu'un texte englobé est présent parmis les fragments de textes avoisinant. D'après les statistiques extraites de notre corpus concernant les références aux sources, la plupart (55%) contiennent une entité nommée et sont constituées d'un groupe nominal. Seulement 15% sont constituées d'un pronom personnel. Nous avons donc choisi de considérer comme expression locuteur potentielle tous les groupes nominaux, ainsi que les pronoms personnels. L'identification d'un groupe nominal ou d'un pronom personnel comme expression locuteur effective est discutée dans la seconde partie de cette section. La première partie discute l'importance de repérer les expressions locuteurs.

3.3.1 Rôle des expressions locuteurs

Si les expressions locuteurs sont au centre de la citation, comme le soulignent G. Mourad et J.P. Desclès [Mourad & Desclès, 2002], considérant comme citation un segment de texte que l'auteur "fait prendre en charge par un locuteur explicite", leur identification est au centre du projet Piithie. En effet, le rattachement de propos à leur source permet de d'en confirmer la paternité et donc l'originalité.

Nous définissons deux étapes dans la reconnaissance des expressions locuteurs. La première étape consiste en le rattachement d'un texte englobé à une expression locuteur faisant référence à la source du dit texte. Cette étape permet de repérer localement les liens unissant le texte englobé et le locuteur auquel il a été associé par l'auteur. La seconde étape correspond au regroupement des différents locuteurs repérés au sein du texte sous des entités sources communes. En d'autres termes, cela revient à identifier les entités sources qui interviennent tout au long de l'article. La réalisation de ces associations est équivalente à la résolution de référence sur des anaphores. La résolution de références anaphoriques est un domaine de recherche très actif, nous avons donc décidé de ne pas traiter de cette étape de résolution dans le cadre du stage.

En résumé, l'identification locale des locuteurs nous permet de rattacher le texte englobé à une expression locuteur. Cependant, si la présence d'un texte englobé *peut* signifier la présence d'une expression locuteur, l'inverse est également vrai. En effet, si l'auteur introduit un locuteur — ce qui est différent d'introduire un personnage —, il y a de fortes chances qu'il lui remette la responsabilité d'un texte englobé présent dans l'article. Nous en faisons du moins l'hypothèse, et dans ce cas, le repérage d'une expression locuteur isolée de tout cadre mimétique nous inviterait à approfondir la recherche de texte englobé, sous la forme d'un discours indirect notamment.

3.3.2 Indices sélectionnés

À la différence des cadres mimétiques candidats dont une majorité (plus de 50%) représentent vraisemblablement des cadres mimétiques avérés, les expressions locuteurs candidates sont en grande

majorité de simples pronoms ou groupes nominaux. La difficulté est de sélectionner des indices relativement faciles à repérer automatiquement au sein des textes et qui soient suffisamment discriminants. Malheureusement pour nous, les expressions locuteurs en tant que telles ne sont pas tellement différentes de n'importe quel pronom ou groupe nominal lambda, si ce n'est au niveau du contexte dans lequel elles s'inscrivent. Nous classons donc les indices concernant la désambiguation en deux catégories : les indices de contexte et les indices de caractérisation de l'expression.

Indices de contexte

Les indices de contexte permettent de décrire le contexte dans lequel se situe l'expression locuteur candidate.

L'auteur se déchargeant de la responsabilité du texte englobé au profit du locuteur, il met en place un scénario d'énonciation impliquant ce dernier. L'expression locuteur se retrouve donc assez souvent aux abords d'un verbe d'énonciation la reliant aux propos qui lui sont délégués par l'auteur. En cas de présence dans le cadre intra-phrastique d'un verbe d'énonciation, la distance minimale entre une extrémité de l'expression locuteur et le verbe constitue un indice. Dans le cas où aucun verbe d'énonciation n'est présent au sein du cadre phrastique, la valeur de l'indice est *inf*.

Une autre méthode des auteurs pour mettre en place le scénario d'énonciation des locuteurs est l'utilisation de syntagmes prépositionnels comme *Selon X* et *D'après X*. Ces formules sont très discriminantes car elles introduisent directement une énonciation. Nous considérons la distance minimale entre l'expression locuteur candidate et ces prépositions, affectant la valeur *inf* lorsqu'il n'y a pas de telle préposition présente dans le cadre phrastique de l'expression candidate. Cet indice est très certainement le plus discriminant dans l'identification des expressions locuteurs.

La dernière méthode significative d'introduction du locuteur par l'auteur est le placement de cette dernière au sein d'une proposition incise elle même placée au coeur du texte englobé, ou bien à une extrémité. L'appartenance d'une expression locuteur candidate à un segment entre virgules est donc le dernier indice de contexte que nous utilisons.

Les indices de contexte concernant les expressions locutives candidates sont assez peu nombreuses, toutefois la distance avec le verbe d'énonciation ou avec la préposition sont assez discriminantes et devraient nous assurer de bons résultats sur l'identification des expressions locuteurs. Nous préférons toutefois compléter cette liste d'indice avec des données de caractérisation des groupes nominaux et des pronoms.

Indices de caractérisation

Les indices de caractérisation permettent de décrire le groupe nominal ou le pronom, extrait de son contexte, et en tant qu'unité linguistique uniquement.

Plus de la moitié des expressions locuteurs extraites du corpus étant composées d'une entitée nommée, nous considérons cette caractéristique comme un indice d'intérêt. Les entitées nommées correspondent à "l'acceptation la plus large que l'on peut faire du nom propre" [Fourour, 2004]. En d'autres termes, il s'agit des expressions textuelles identifiant une entité telle qu'une personne, un lieu, une maladie,... N'ayant pas le temps d'intégrer à la chaîne de traitement l'outil *Némésis* permettant de détecter les entitées nommées dans les textes, nous nous sommes basé sur les résultats plus ou moins fiables de Tree-Tagger[4]. Ce dernier emploie la balise *NAM* pour ce qu'il considère comme entité nommée. Malheureusement,

[4]http ://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

cet étiquetage est assez approximatif puisqu'il semble l'employer uniquement dans les cas d'un nom capitalisé placé en position de sujet du verbe.

Afin de pallier partiellement la catégorisation aléatoire de Tree-Tagger sur les entités nommées, nous avons choisi de considérer également comme indice l'existence au sein de l'expression locuteur candidate d'un nom capitalisé. Ce choix a plusieurs écueils, notamment la confusion entre les majuscules imposées par la typographie (début de phrases, initiales,...) et celles des noms propres de personnes, de lieux,... Nous conservons toutefois l'indice, une évaluation de son impact serait toutefois nécessaire en mettant par exemple en place un protocole expérimental d'annotation en tenant compte puis en ignorant cet indice.

Finalement, le dernier indice qui peut potentiellement nous aider à caractériser les expressions locuteurs avérées est l'utilisation au sein du groupe nominal candidat d'un article défini. En effet, l'utilisation d'un tel article indique une référence précise, et donc potentiellement une entitée nommée. il s'agit là du dernier élément utilisé pour les groupes nominaux, les indices suivants concernant les pronoms.

La caractérisation des pronoms n'est pas chose aisée étant donnée que ces derniers sont isolés du reste du texte et ne forment pas de syntagmes particuliers, du moins pas de syntagmes d'intérêt pour nos besoins. Cependant, nous pouvons constituer deux classes de pronoms : ceux à la troisième personne et ceux aux deux premières personnes. Contrairement à la caractérisation des cadres mimétiques, dans le cas des locuteurs, les pronoms à la troisième personne sont plus discriminants. En effet, l'auteur désirant se détacher du locuteur, il utilisera plutôt une forme à la troisième personne. Dans ce cadre, l'utilisation d'une première ou deuxième personne mène très probablement à un pronom non locuteur. Cette dernière caractéristique peut également être caractéristique envers un pronom non locuteur.

Enfin, en se basant sur le fait que les locuteurs sont très déterminés et en faisant l'hypothèse que ceci s'illustre par un nombre de mots plus importants que les groupes nominaux ordinaires — notamment par la présence d'adjectifs ou d'adverbes —, nous considérons la taille comme un indice supplémentaire.

Certains indices nous semblent inaccessibles pour un repérage automatique efficace mais apporteraient certainement une meilleure précision quant au repérage des expressions locuteurs. Ainsi, lorsque l'auteur fait référence à une source pour la première fois, il utilise "des formes très déterminées, parfois même sur-déterminées" [Giguet & Lucas, 2004]. Cette détermination de la source s'effectue souvent par le biais d'une proposition incise rattachée à l'expression locuteur et décrivant la fonction ou le rôle de la source. Si nous étions capables de repérer effectivement les groupes nominaux auxquels sont rattachées des propositions incises, cela constituerait très certainement un indice très discriminant. Le manque de temps nous a également contraints à délaisser l'utilisation de *Némésis*, développé au sein du LINA[5], et permettant un repérage efficace des entités nommées.

3.4 Identification des segments citationnels

La dernière étape dans notre algorithme est l'extraction des segments citationnels à partir des annotations effectuées. Les segments citationnels sont l'aboutissement du traitement puisqu'ils correspondent à ce que nous avons déterminé comme étant le segment textuel le plus proche de l'unité citationnelle, selon nos critères discutés dans les premiers chapitres. La première sous-section discute ce choix et ses implications. Une fois les étapes de segmentation du texte et de la désambiguation des cadres mimétiques et des expressions locuteurs effectuées, trois scénarios d'intérêt se présentent : un locuteur et un cadre mimétique repérés dans la phrase, un cadre mimétique seulement ou bien un locuteur seulement. Ces différents scénarios sont discutés dans la seconde sous-section.

[5]Laboratoire d'Informatique de Nantes Atlantique

3.4.1 Segments citationnels : adaptation à notre méthode

Ayant rencontré des difficultés pour définir une unité citationnelle, nous avons decidé de considérer une unité approximative : le segment citationnel. Ce choix est motivé par la difficulté de déterminer avec précision les bornes du texte englobé, et discuté plus en profondeur dans les chapitres précédents. Malgré son caractère approximatif, le segment citationnel contient toutes les informations que l'on désire connaître sur la citation.

Les segments citationnels sont généralement constitués de fragments de texte reconstituant le texte englobé et rattachés à un locuteur. Le segment citationnel est le plus petit segment textuel contenant toutes ces informations et aligné sur des signes ponctuatifs. "Généralement", car il y a malheureusement des cas où le locuteur, et donc la source, n'est pas identifiable. Plusieurs solutions sont alors possibles :

- se rattacher au dernier locuteur encore connu, en s'assurant qu'il soit encore suffisamment saillant ;
- se contenter de fragments de texte englobés potentiellement incomplets.

La première solution nous semble la plus probante. Cependant, nous manquons de temps pour l'implémenter et nous limitons donc dans le cadre de ce stage à la deuxième solution.

Le choix de la seconde solution ne nous empêche nullement de discuter la première. En effet, celle-ci nous paraît largement supérieure et mérite quelques approfondissements. Le rattachement à un locuteur présent hors du cadre phrastique n'est pas une idée neuve, elle a notamment été énoncée par E. Giguet et N. Lucas qui avaient introduit le concept de "source de référence" dans [Giguet & Lucas, 2004]. La difficulté de cette solution réside dans l'évaluation de la saillance de la source. Nous proposons une méthode basée sur les résolutions de référence anaphorique. L'idée est d'évaluer la saillance des différentes sources sous forme d'une distance en mots pondérée par les références anaphoriques entre l'introduction du locuteur et le texte englobé détecté.

Notre position est que le lecteur remonte automatiquement à la dernière référence au locuteur présente dans le texte, exceptée si cette dernière est trop éloignée. Il prend alors par défaut le locuteur le plus présent dans le texte. Bien entendu, "trop éloigné" et "le plus présent" sont des concepts à éclaircir et coucher sous forme de mesures rationnelles. La mise au point d'une telle mesure nécessite des expérimentations, aussi nous nous limitons ici à énoncer nos idées, sans pouvoir les soutenir à l'aide d'expérimentations, mais il nous apparaît que la distance entre l'introduction de la source à l'aide du locuteur initial soit relativisée par la présence au sein du même paragraphe d'une référence anaphorique à cette source. Le choix du dernier locuteur s'effecturait donc, par ordre d'importance, par la distance au sein du paragraphe, suivi de la distance en-dehors du paragraphe avec l'introduction de la source divisée par la somme des distances entre les références à la source et le texte englobé.

Le segment citationnel est le regroupement de l'expression locuteur et du texte englobé en alignant les extrémités de chacun sur les éléments de ponctuation les plus proches tout en les contenant complètement. Ceci s'applique aussi bien au cas où le locuteur et le texte englobé appartiennent au même cadre phrastique que celui où les deux éléments sont situés dans des cadres phrastiques différents mais voisins. La difficulté s'inscrit donc lorsqu'il n'y a aucune référence à la source à proximité du texte englobé, nous ne cherchons pas à résoudre ce cas dans le cadre du stage.

3.4.2 Extraction des segments citationnels

Une fois les traitement d'identification des cadres mimétiques et des expressions locuteurs effectués, quatres scénarios sont envisageables pour chacune des phrases de l'article traité. Si aucun élément n'est

identifié au sein de la phrase, alors celle-ci ne contient pas de segment citationnel, ou bien ce dernier est trop diffus dans le texte pour que nous puissions le repérer. Si au moins élément est repéré alors la situation est l'une de celles discutées dans les paragraphes suivants.

Présence d'un locuteur et d'un cadre mimétique avéré

La présence au sein d'un même cadre phrastique d'un locuteur et d'un cadre mimétique avéré correspond au cas d'école. Il suffit d'aligner les extrémités de chacun des éléments sur les éléments ponctuatifs extérieurs les plus proches, le segment textuel ainsi créé correspond au segment citationnel. Ce dernier peut toutefois se voir modifier selon l'apparition des scénarios discutés dans les paragraphes suivants aux abords du segment ainsi créé.

Présence d'un cadre mimétique avéré sans locuteur

La présence d'un cadre mimétique avéré implique inévitablement la présence d'un segment citationnel. Trois cas se présentent alors :

- on peut le rattacher au locuteur d'un cadre phrastique précédent ;
- on peut le rattacher à un segment citationnel voisin ;
- aucun des deux cas précédents n'est réalisable.

Il est concevable de trouver un locuteur avéré isolé, au sein de son cadre phrastique, de tout élément permettant de constituer un segment citationnel. Si un tel locuteur est présent dans les cadres phrastiques précédents ou suivants le cadre phrastique au sein du paragraphe contenant le cadre mimétique et tel qu'aucun segment citationnel n'interfère entre les deux, alors on rattache le cadre mimétique à ce locuteur en formant ainsi un segment citationnel.

Si toutefois aucun locuteur ne satisfait ces conditions, il est possible d'intégrer un segment citationnel voisin. En effet, si le cadre mimétique se trouve dans un cadre phrastique voisin direct d'un segment citationnel, on intègre ce dernier à ce segment citationnel. La saillance du locuteur du segment citationnel voisin rend ce ralliement possible.

Finalement, il est possible — bien que plus rare — qu'aucune des situations précédentes ne soit envisageable. Dans ce cas, on crée un segment citationnel de fait contenant uniquement le cadre mimétique. Ce segment citationnel se rattache implicitement à la source globale au texte dans le cadre d'un texte monolocuteur ou bien il s'agit d'une source déduite d'après le contexte. Nous ne sommes pas en mesure de résoudre ce dernier scénario sans une base de connaissance du domaine dont traite l'article. Ceci nécessite un travail qui sort du cadre du stage de master.

La présence d'un cadre mimétique isolé au sein de son cadre phrastique de locuteur peut se rattacher à un locuteur libre proche ou bien un segment citationnel voisin. En dernier recours, il est possible de constituer un segment citationnel de fait composé uniquement du cadre mimétique. Le cas est plus intéressant, quoi que plus complexe à résoudre lorsque l'on se trouve en présence d'une expression locuteur isolée.

Présence d'un locuteur avéré seul

Nous nous sommes concentré sur l'identification d'expressions locuteurs, ainsi que de cadres mimétiques. Cette approche néglige toutefois les citations dont le texte englobé n'est pas composé de fragments entre

guillemets ou bien tel que les indices permettant d'identifier de tels fragments comme cadres mimétiques n'étaient pas suffisant. L'on se trouve alors face aux scénarios définis ci-dessous.

La premier scénario à explorer est que le locuteur se rattache à un texte englobé ne contenant pas de segments entre guillemets ou tels que ces derniers ne comportaient pas d'indices suffisants pour être reconnus comme cadres mimétiques avérés. Ce cas de figure nécessite de faire appel une nouvelle fois à des indices discursifs. Ces indices sont globalement les mêmes que pour les cadres mimétiques candidats mais font appel à un apprentissage supervisé différent. La présence d'un locuteur avéré est un enclencheur suffisamment important pour utiliser un automate plus laxiste que celui identifiant les cadres mimétiques. De plus, dans ce cas particulier, le texte englobé étant plus diffus, nous n'essayons pas de repérer les bornes du texte englobé, mais simplement décider si le cadre phrastique contenant le locuteur avéré doit être considéré comme un segment citationnel.

Si l'automate ne permet pas d'identifier le cadre phrastique comme un segment citationnel, l'on tombe peut être dans le cas d'un cadre mimétique voisin isolé. Si tel est le cas, le scénario est identique à celui présenté précédemment lorsque l'on se trouve dans la situation où un cadre mimétique est isolé de toute locuteur dans son contexte phrastique.

Notre méthode n'étant malheureusement pas infaillible, la dernière hypothèse est que l'on n'ait pas réussi à détecter de texte englobé aux abords du locuteur ou bien que ce dernier n'en soit pas réellement un. Dans les deux, une fois que l'on s'est assuré que le scénario précédent ne puisse pas être appliqué, on se contente de retirer son étiquette de "locuteur" au segment de texte considéré.

La présence d'un locuteur isolé est le scénario le plus difficile à résoudre. Si les indices de présence d'un texte englobé rapporté sans guillemets manquent, il est alors possible que ce locuteur soit rattachable à un cadre mimétique proche ou bien qu'il ne soit pas réellement un locuteur.

3.5 Synthèse

La recherche de citations passe par l'identification de ses composants : texte englobé et expression locuteur. La considération des indices liés aux cadres mimétiques candidats et aux expressions locuteurs candidates nous permettent parfois de porter ces candidats à l'investiture. Une fois ces différents éléments repérés au sein des textes, il nous est alors possible à l'aide de règles souvent simples d'extraire ou de composer des segments citationnels reflétant au mieux le concept de citation.

Les cadres mimétiques, les cadres phrastiques et les expressions locuteurs occupent un rôle important dans ce processus d'extraction des segments citationnels. Ils permettent de considérer les segments citationnels potentiels selon plusieurs approches : à partir des fragments de texte englobé entre guillemets dans un premier temps, puis à l'aide des locuteurs dans un second temps. L'introduction des cadres candidats nous permettent de considérer les passages potentiellement d'intérêt sous plusieurs angles au lieu de considérer le texte comme un tout. Cette caractéristique fait l'originalité de cette méthode. Nous tâchons de réfléter au mieux l'adaptabilité et l'originalité de cette méthode dans son implémentation sous la forme d'une chaîne de traitement. Les choix d'implémentations et les résultats sont présentés dans le chapitre suivant.

Chapitre 4

Chaîne de traitement, expérimentation et analyse des résultats

Au chapitre précédent, nous avons défini les principes de notre méthode de détection. Dans le présent chapitre, nous passons désormais aux phases d'implémentation et d'expérimentation.

Nos objectifs sont doubles : mettre en place des chaînes de traitement qui permettent l'identification des cadres mimétiques et des expressions locuteurs.

Etant donné la complexité des langues, il est assez difficile de pouvoir définir des règles de fonctionnement de tel ou tel phénomène linguistique à partir d'observations à l'oeil humain. C'est pour cette raison que nous avons choisi d'utiliser des techniques d'apprentissage supervisé pour construire de manière automatique des modèles prédictifs pour classer ensuite nos données.

L'idée est la suivante : dans un premier temps, un modèle est construit en soumettant à un algorithme d'apprentissage des exemples classés de ce que l'on désire apprendre. Puis dans un deuxième temps, nous pouvons à notre guise appliquer le modèle à de nouvelles données non classées afin qu'il prédise leur classe. Les données d'apprentissage sont aussi appelées exemples ou instances.

Dans le cadre de notre travail d'évaluation de nos outils de détection, la phase d'apprentissage consiste à

1. d'abord extraire les cadres mimétiques candidats et les expressions locuteurs candidats,
2. puis à les aligner sur nos annotations manuelles décrites au chapitre sur la constitution de corpus afin de leur attribuer une classe "est un cadre mimétique" ou "pas" et "est une expression locuteur" ou "pas" ;
3. ensuite à caractériser chaque candidat à l'aide de traits pertinents ;
4. et enfin à appliquer un algorithme d'apprentissage sur ces données.

L'utilisation du modèle appris consiste à projeter des classes sur des données caractérisées mais pas encore classées.

L'évaluation de l'apprentissage automatique fait appel aux notions de précision, rappel, faux positifs et faux négatifs. Nous introduisons rapidement ces notions ici. Soit l'ensemble des éléments évalués par un modèle, cet ensemble se partitionne en deux groupes : des éléments pertinents (correspondant à ce que nous recherchons) et des éléments non pertinents. La précision mesure la proportion d'éléments considérés pertinents parmis ceux notés comme tels par le modèle. Une précision de 1 indiquerait donc que tous les éléments considérés pertinents par le modèle le sont effectivement. Le rappel mesure la

proportion d'éléments pertinents retenus par le modèle par rapport à la totalité de ces éléments. Un rappel de 1 indiquerait donc que tous les éléments pertinents ont été repérés par le modèle. Les éléments pertinents qui n'ont pas été classés comme tels par le modèle sont les faux négatifs, alors que les éléments non pertinents qui ont été classés par pertinents par le modèle sont les faux positifs.

Dans une première section, nous présentons d'abord notre chaîne de traitement. Après un bref aperçu, nous indiquons les principes de conception que nous avons adoptés. Puis, dans une deuxième section, nous enchaînons avec la description des composants que nous avons développé pour l'extraction de candidats. Ensuite, nous consacrons une section à la présentation de nos choix d'attributs et des composants qui permettent la caractérisation des données. Enfin, nous rapportons, dans une dernière, nos expérimentations quant à la reconnaissance automatique des cadres mimétiques et des expressions locuteurs. Nous terminons le chapitre en exposant et discutant les résultats obtenus lors de nos évaluations.

4.1 Notre chaîne de traitement

Dans cette section, nous présentons la réalisation logicielle produite en réponse à notre analyse du problème de la détection automatique de citation.

4.1.1 Aperçu général

Suivant la tâche désirée (reconnaissance de cadres mimétiques ou d'expressions locuteurs), notre chaîne de traitement peut avoir différents visages. Globalement elle consiste en un séquençage de composants modulaires réalisant une tâche bien définie.

On dénote trois types de composants :

1. Les composants en amont de chaîne qui réalise un pré-traitement du corpus. Celui-ci consiste à la fois à tokeniser (découper le texte) en mots ainsi qu'à lemmatiser et étiqueter morpho-syntaxiquement les mots ;
2. Les composants centraux qui extraient des candidats à classer (dans notre cas les cadres mimétiques candidats et les expressions locuteurs candidates) ;
3. Et finalement les composants nécessaires ou bien à la construction d'un modèle d'apprentissage ou bien à l'application d'un modèle déjà construit.

Les composants d'apprentissage ou de projection d'un modèle de données reposent tous sur un même composant : le composant de caractérisation qui décrit chaque instance de l'objet à apprendre ou à classer selon une série d'attributs présélectionnés. Ce composant est spécifique à ce que l'on cherche à identifier.

Nous exposons dans la section suivante les principes que nous avons spécifiés afin de définir l'architecture générale de notre chaîne ainsi que les contraintes de nos composants d'extraction de candidats.

4.1.2 Principes de conception

Lors de la conception de notre plate-forme de traitement, nous avons voulu mettre l'accent sur plusieurs points :

- l'utilisation directe du format XML du corpus ;
- la possibilité d'utiliser différents tokenizers, lemmatiseurs, . . . ;
- la modification de l'ordre d'exécution des tâches du traitement ;
- la possibilité d'ajouter des tâches à différents endroits du traitement sans perturber ce dernier.

En privilégiant ces caractéristiques, nous avons négligé les aspects de complexité et de rapidité. Nous pensons toutefois que cette approche se prête plus à notre démarche de recherche, justifiant les choix de la conception. Nous décrivons dans un premier temps l'interface, puis dans un second temps le modèle XML utilisé pour traiter les textes.

Architecture de composants

Nous nous sommes tourné vers l'architecture de composants car elle semblait apporter la meilleure flexibilité, ce que nous recherchions. Cette architecture nous offre la possibilité de modifier l'enchaînement des traitements très simplement en modifiant les connexions entre les composants.

L'architecture de composants est une forme de développement orienté objet où le code est divisé en modules. Cette division permet de faciliter le maintien du code et surtout sa réutilisabilité. Ainsi, il est possible de modifier un composant sans que cela influence sur le reste de l'application, à moins de changer son interface. Il est aisé de développer de nouveaux composants pour l'application sans avoir une connaissance profonde de l'application.

Dans le cadre de notre travail, les composants rappelent quelque peu les règles utilisés par G. Mourad [Mourad & Desclès, 2002]. En effet, ils offrent, à l'instar des règles, la possibilité d'adapter le traitement du texte en modifiant l'enchaînement des tâches : inversion de composants, ajout, suppression. Cette architecture nous permet ainsi de nous rapprocher quelque peu de la modularité offerte par la plate-forme ContextO utilisée par G. Mourad et qui ne nous était malheureusement pas accessible. Les composants s'ils peuvent offrir une certaine souplesse sont cependant soumis à un certain nombre de règles.

Chaque composant de l'architecture est défini par une interface. Cette interface définit la manière dont il communique avec les composants qui l'entourent. Dans le cadre du stage, les composants dérivent de quatres formes de composants principaux étudiés dans la section suivante :

- parsers : permettent un parcours des données XML par balises ou segments de texte (*cf. section 4.1.2 sur le modèle XML*) ;
- segmenters : offrent la possibilité de modifier la structure des données XML traitées ;
- charsegmenters : permettent un parcours par caractère ;
- lemmatisers : offrent les méthodes nécessaires à la lemmatisation et l'attribution des catégories grammaticales (balises POS – *Part Of the Speech*) définissant le rôle des mots dans les phrases.

À ces quatre composants, s'ajoute le composant particulier *Tokenizer* permettant de tokeniser par mots les textes si ils ne le sont pas déjà. Ce composant est particulier puisqu'il est nécessaire à l'utilisation de la couche d'abstraction d'accès aux fichiers (*cf. figure 4.1.2*).

Certains composants, comme le *Tokenizer*, doivent nécessairement être placés en amont de certains autres dans la chaîne de traitement. Cela s'explique tout simplement par le fait que les premiers produisent des résultats indispensables à l'exécution des seconds. Il existe donc une relation de dépendances entre les composants à laquelle il est nécessaire de prendre garde lors de la mise en place d'une chaîne de traitements.

Chacun des composants, défini par son interface, communique avec les autres composants de manière asynchrone par l'intermédiaire de fichiers XML. Cette technique nous apparaît comme la plus simple à mettre en oeuvre. De plus, elle nous permet de connaître le résultat des traitements intermédiaires par chacun des composants impliqués. Elle souffre toutefois d'un manque d'efficacité lors du traitement de données de taille importante à cause des écritures et lectures répétées. La figure 4.1.2 illustre ce fonctionnement. La forme rectangulaire entre les fichiers et les composants représente les différentes classes d'abstraction des fichiers XML. Ces classes permettent de traiter directement avec les concepts de segments textuels ou d'articles sans passer par le modèle XML. Dans cette illustration, le *composant*

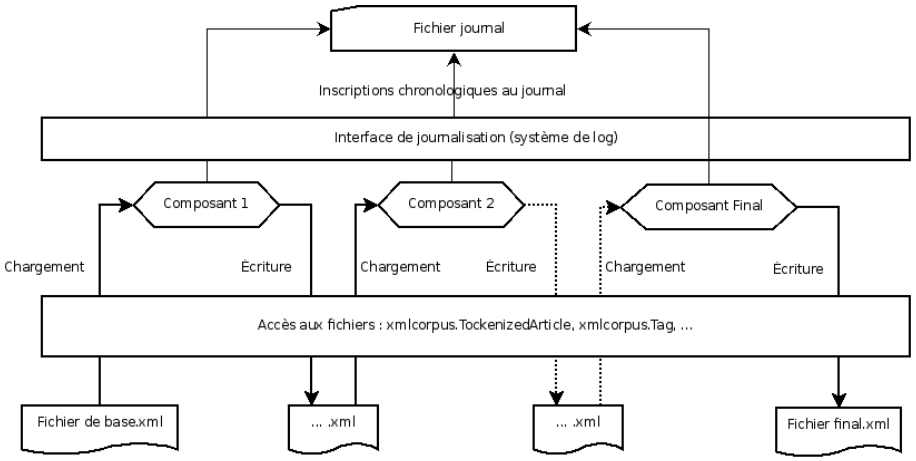


FIG. 4.1 – Vision globale du fonctionnement de la chaîne

1 initie le traitement sur le fichier *Fichier de base.xml* et sauvegarde le résultat de ce traitement dans un fichier repris par le *composant 2* et ainsi de suite jusqu'au dernier composant qui écrit le *Fichier final.xml*. Le résultat du traitement global réalisé correspond donc au contenu de ce dernier fichier.

Nous choisissons d'utiliser le langage Python pour l'implémentation de cette chaîne de traitement. Toutefois, le choix de notre architecture nous permet d'utiliser n'importe quel langage pour chacun des composants, sous réserve de la possibilité de lancer l'exécution du composant à partir de Python. La première raison justifiant ce choix est l'excellente gestion par Python de l'unicode, et des chaînes de caractères en général. De plus, Python offre une orientation objet souple mais performante, permettant de mettre en place l'architecture de composants assez aisément afin de se concentrer sur les composants en eux même. Finalement, les langages de script permettent un développement rapide et une mise au point efficace, ce qui est un atout majeur pour un développement comme le notre.

En résumé, nous choisissons une approche très modulable qui nous permettra non seulement de tester différentes approches de notre méthode. Ainsi, le travail pourra également être utilisé par la suite, dans d'autres travaux sans devoir redévelopper les composants. La communication entre composants par le biais des fichiers respectant notre modèle XML permet également, sous réserve de modifications mineurs, de rendre chacun des composants autonomes et donc directement utilisables sans la plate-forme. Nous discutons le modèle XML dans la sous-section suivante.

Modèle XML

Dans la lignée de notre architecture modulable, nous faisons le choix de la souplesse pour notre modèle XML. Cette souplesse s'illustre tout d'abord par la liberté dans le choix des balises, mais également par la possibilité d'entremêler les balises tout en conservant le XML bien formé.

Premièrement, la liberté de balisage se définit selon nous par les caractéristiques suivantes :

- libre choix du nom de la balise ;
- libre choix de l'espace de nom de la balise ;
- libre choix des attributs et de leurs espaces de nom ;
- liberté d'ignorer ou de considérer certaines balises.

Par libre choix du nom des balises et des espaces de noms, nous entendons que les composants doivent s'adapter aux choix des noms réalisés par les composants précédents. Ces choix doivent bien entendu respecter les règles[1] définies par le World Wide Web Consortium[2].

Par libre choix des attributs, nous voulons dire que le nom de l'attribut doit être libre au même titre que le nom de la balise. Toutefois, la nécessité d'un nombre *n* d'attributs contenant des données tirés des domaines D^n par un composant doit être respecté et ne peut être tronqué. Ainsi, si un composant X nécessite la présence d'un attribut auquel est affecté une valeur de l'ensemble $rouge, noir, vert, bleu$, un composant présent en amont dans la chaîne doit se charger de le placer, peu importe l'identifiant et l'espace de nom qu'il utilise.

Par libre choix d'ignorer ou de considérer certaines balises, nous stipulons que les balises qui ne concernent pas un composant ne doivent pas polluer l'exécution de ce dernier. En d'autres termes, celui-ci n'est pas conscient des balises qui n'interviennent pas dans son traitement.

Deuxièmement, le choix du modèle ne doit pas contraindre les possibilités d'annotation. La structure arborescente du XML ne doit donc pas être une contrainte pour les composants. Il doit ainsi être possible d'entremêler les balises sans devoir se soucier de la validité de la structure, fonction déchargée à la plate-forme.

Nous abstrayons cette fonctionnalité de la plate-forme sous l'entité "segment de texte". Un "segment de texte" au sein de la plate-forme est une suite de tokens mots, éventuellement accompagné d'un balisage, délimité au sein de l'ensemble des token du texte traité par un balisage XML. Ce balisage correspond au couplage standard *<tag>donnée</tag>* si il ne remet pas en cause la structure arborescente du fichier. Si jamais cette structure est remise en cause, la plate-forme "marque" les extrémités du segment de texte à l'aide d'une balise de type "empty-element tag"[3] accompagné d'un attribut dont le domaine de valeur se limite à deux valeurs : une représentant l'ouverture, la seconde la fermeture. Le listing ci-dessous illustre ce cas de figure :

```
  Ceci
  <tag>
    est <bornesegment borne="debut"/>une illustration
  </tag>
  de
  <tag>
    l'entrem\^elage␣possible
␣␣␣␣<bornesegment␣borne="fin"/>␣\'a l'aide␣de
␣␣␣␣l'empty-element tag et d'un
␣␣</tag>
␣␣attribut␣\'a domaine de valeur binaire.
```

Il est toutefois nécessaire de prendre garde à la remarque "Unique Att Spec"[4] de la spécification XML qui stipule que deux attributs ne peuvent avoir le même nom au sein d'une même balise ouvrante, ou vide. Cette remarque s'oppose alors à la liberté du choix des attributs que nous avons précédemment énoncé. Nous sommes donc contraint de réduire cette liberté à l'utilisation de n'importe quel nom respectant les spécifications XML et qui ne provoque pas de collision avec l'attribut utilisé pour le bornage des segments textuels.

[1] http ://www.w3.org/TR/xml/#NT-Name
[2] http ://www.w3.org
[3] http ://www.w3.org/TR/xml/#NT-EmptyElemTag
[4] http ://www.w3.org/TR/xml/#uniqattspec

Le modèle XML que nous avons choisi se veut le plus souple possible, et ne diffère du métalangage XML lui même seulement par la notion de "segment de texte". Cette proximité nous permet d'utiliser au mieux les bibliothèques performantes existantes pour le traitement des fichiers XML. Un fichier de traitement intermédiaire illustrant la mise en oeuvre de ce modèle XML est présent dans l'annexe D.3.

Justification d'une réimplémentation

De nombreux *toolkits* accessibles gratuitement permettent de prendre en charge certaines opérations élémentaires liées au traitement du langage, comme des opérations plus complexes. Nous avons cependant délibéremment fait le choix de réimplémenter une grande partie de ces outils, dans un but éducatif tout d'abord et de pérennité ensuite.

Réimplémentation dans un but éducatif

Le premier argument en faveur d'une réimplémentation est le cadre dans lequel ce travail est effectué. En effet, le but de ce stage, au delà de prouver au jury les capacités d'une personne à réaliser un travail de recherche, est de permettre une découverte plus approfondie du domaine de recherche. Dans cette optique, la réimplémentation d'outils tels qu'un tokeniseur de mots ou de phrases permet de se confronter à certaines difficultés à côté desquelles nous serions certainement passé autrement.

Réimplémentation pérenne

Le deuxième argument en faveur d'une réimplémentation est la volonté de posséder des outils pérennes. Les laboratoires, et d'une façon plus générale les auteurs des toolkits ont tendance à protéger jalousement les outils qu'ils développent en ne permettant pas un accès aux source, ou du moins un droit de regard et de modification des dites sources. Les logiciels verrouillés de ce style peuvent, notamment dans le cadre de recherches, se refermer sur leurs utilisateurs en verrouillant leurs résultats dans un format incompréhensible par un autre logiciel, ou tout simplement en ne permettant plus l'utilisation libre du dit logiciel. Ces situations ne sont pas des fabulations et on en trouve des illsutrations régulièrement.

Notre choix d'une plate-forme développée par nos soins **et** placée sous une licence assurant un accès au code source et sa mise à disposition pour une utilisation libre telle que la GPL, se place dans ce contexte. En développant nous même cette plate-forme, nous sommes certain que celles-ci correspondra à nos besoins. En plaçant notre développement sous licence GPL, nous offrons la possibilité à des personnes extérieures de participer au développement ou tout simplement d'utiliser librement la plate-forme. Nous ne sommes d'ailleurs pas les seuls, de plus en plus de projets libres à destination du traitement des langues naturelles voient le jour. On peut ainsi saluer les projets OpenNLP[5] et Weka[6] pour leur travail en ce sens.

En résumé, nous considérons le développement de cette plate-forme comme un besoin, mais également une sorte de voyage initiatique à la découverte de quelques sujets classiques de la recherche en traitement automatique du langage naturel.

Les sections suivantes sont une mise en application de ces principes au travers le développement de composants.

[5] `http ://opennlp.sourceforge.net/`
[6] `http ://www.cs.waikato.ac.nz/ ml/weka/`

4.2 Composants d'extraction de candidats

Reprenant les conclusions de notre démarche méthodologique, nous avons cherché à implémenter trois traitements distincts qui se retrouvent chacun encapsulé dans un composant. Nous décrivons d'abord notre segmenteur en cadres mimétiques candidats, puis notre segmenteur en cadres phrastiques et enfin notre extracteur d'expressions locuteurs.

4.2.1 Segmenteur en cadres mimétiques candidats

La segmentation en cadres mimétiques candidats est une composante très importante de notre algorithme. Elle consiste à repérer et isoler les fragments de textes placés entre guillemets. Les difficultés majeures rencontrées lors de l'implémentation de ce segmenteur sont l'appariemment des guillemets et la gestion des irrégularités introduites par l'auteur.

Nous avons été confronté en premier lieu au problème d'appariemment des guillemets. Ce problème consiste à regrouper les marques typographiques de type guillemets par deux. Certaines marques typographiques utilisées comme guillemets possèdent font la différence, au niveau du symbole, entre le guillemet ouvrant et le guillemet fermant. L'utilisation de ces marques simplifie le problème puisque l'utilisation d'une structure FILO permet de déterminer les bornes des segments entre guillemets. Toutefois, l'utilisation de guillemets "symétriques", comme les guillemets dits droits : ″ ″, complexifient lourdement le problème. En effet, l'algorithme naïf consistant à considérer les guillemets dans leur ordre chronologique donne de très mauvais résultats.

> Aujourd'hui, William Gaillard, le porte-parole de l'UEFA, précise qu'aucun projet n'a été arrêté : "On travaille sur plusieurs scénarios. Pour l'instant, nous n'avons que le but, qui est que les pays "moyens" soient mieux représentés. Le reste est à l'étude. Nous avons le temps : l'échéance, c'est la saison 2009-2010."
>
> *Le Monde - 20 Février 2007*

Le problème, illustré par l'extrait ci-dessus, est que le placement en emphase au sein d'un passage entre guillemets trompe l'algorithme naïf. Ici, la segmentation s'arrête au guillemet précédant *moyens* et reprend après. La segmentation correcte correspond cependant au segment complet intégrant *moyens*. Nous avions pensé utiliser le positionnement des espaces aux alentours des guillemets pour résoudre ce cas. Cependant, cela ne s'est pas révélé fiable en raison de la variété de positionnement de ces derniers au sein des articles journalistiques. Une brève recherche ne nous a permis d'isoler un algorithme simple et efficace permettant de résoudre les ambiguïtés d'appariemment des guillemets de ce style.

L'utilisation d'une structure FILO pour l'appariemment des guillemets nous permet de gérer l'imbrication de ces marques lorsqu'il n'y a pas d'ambiguité sur l'appariemment (utilisation de guillemets asymétriques ou bien alternance des guillemets droits avec des guillemets asymétriques). L'exemple 4.2.1 n'est donc pas géré, mais le situations telles que celle illustrée par l'exemple 4.2.1 le sont.

> « Il l'a caché aux 36 saisonniers. Je l'ai appris sur un journal. Je lui ai dit : "C'est vendu, M. Edouard ?" "C'est vendu." "Et alors ?" "Alors, je ne peux rien faire pour vous." »
>
> *© Libération - 22 Février 2007*

Le problème apparu en second lieu concerne les irrégularités des auteurs dans le placement des textes entre guillemets. Ainsi, l'oublie d'un guillemet fermant, bien que rare, provoque un certain chaos dans l'exécution de l'algorithme sur les guillemets suivant l'oublie. Ne possédant pas non plus d'algorithme

simple et efficace permettant de pallier ce problème, nous avons utilisé l'heuristique suivante : tous les passages entre guillemets non clôs à la sortie d'un bloc de type titre, paragraphe, . . .sont automatiquement fermés. Cette heuristique s'est révélé plutôt efficace.

Nous avons conservé notre algorithme naïf basé sur une structure FIFO appareillant les guillemets asymétriques de manière efficace, mais biaisé dans l'appariemment des guillemets symétriques. L'impact de l'utilisation d'un algorithme permettant un appariemment efficace dans tous les cas serait à évaluer.

4.2.2 Segmenteur en cadres phrastiques

Le repérage des phrases au sein de textes est un problème récurrent. On pourra s'en rendre compte en consultant le nombre de mails traitant régulièrement du sujet sur la liste de diffusion `corpora@ubi.no`.

Le problème principal auquel nous avons été confronté lors de la mise au point de notre algorithme de segmentation des articles en phrases est l'ambiguïté du point. En effet, notre algorithme se base sur la détection des éléments ponctuatifs *"."*, *"!"* et *"?"*. Nous les considérons comme marqueurs de fin de phrase. Les éléments ponctuatifs *"!"* et *"?"* sont peu ambigus, mais le point peut également être utilisé comme marqueur décimal ou bien comme marque de la fin d'une abréviation.

Nous avons résolu le problème du point comme marqueur décimal par l'ajout d'une règle de contexte dans le tokeniseur en mots, puis en considérant les mots au lieu des caractères dans la segmentation en phrase. La règle de contexte utilisée pour désambiguïser est la suivante : si le point est entouré de chaque côté par un symbole assimilable à un chiffre, alors on ne le considère pas comme un marqueur de fin de phrase. La désambiguation des points utilisés pour les abréviations nécessite de passer par une phase d'apprentissage, ou bien de faire appel à un dictionnaire. Nous n'avons pas pris le temps de constituer de telles ressources.

Il semble fréquent en TALN que les algorithmes naïfs permettent d'obtenir de bons résultats, mais dès que l'on tente de résoudre les cas problématiques, les algorithmes se complexifient très rapidement.

4.2.3 Extracteur d'expressions locuteurs candidats

Initialement, nous projetions de mettre en place le logiciel Nemesis, développé au sein de l'équipe [Fourour, 2002], pour reconnaître les entités nommées dans nos textes. Les entités nommées sont des expressions linguistiques qui désignent des personnes ou des organismes. Ces expressions devaient servir d'expressions locuteurs candidats.

Malheureusement la mise en place et la prise en main de Nemesis se sont avérées trop coûteuses dans le temps imparti.

Nous avons donc décidé de réaliser une première version de notre chaîne avec un extracteur de termes candidats qui se fondent uniquement sur un patron syntaxique simple. Le patron qui a été retenu était de la forme *déterminant+adjectif ?+nom*.

Dans la section qui suit nous présentons les composants qui permettent la caractérisations des candidats.

4.3 Composants de caractérisation des données

Nous avons présenté au chapitre précédent les différents indices que nous considérions d'intérêt pour l'identification des cadres mimétiques et des sources. Dans cette section, nous expliquons la manière dont nous avons choisi nos attributs pour décrire nos exemples et présentons nos composants en charge de la caractérisation des données.

4.3.1 Choix des attributs et composant pour la caractérisation des cadres mimétiques

La section 3.2.2 a introduit et discuté le choix des indices intervenants dans la prise de décision concernant l'identification des cadres mimétiques. Afin de réaliser un apprentissage supervisé, nous devons extraire des exemples (ou instances) à partir de ces indices. Nous discutons dans cette section le choix des attributs de ces instances.

Choix des attributs constituant les instances

Nous décrivons ici les attributs caractéristiques que nous allons utiliser au sein des instances d'apprentissage. Les attributs sont regroupés par indice auquel ils font référence.

Les attributs que nous avons considéré et faisant référence à l'indice "taille du cadre" sont les suivants :

taillecadremim : taille du cadre mimétique candidat en nombre de mots bruts

distmimphrast-gauche : distance entre le bord gauche du cadre phrastique et le bord gauche du cadre mimétique

distmimphrast-droit : distance entre le bord droit du cadre phrastique et le bord droit du cadre mimétique

Les attributs faisant référence à un cadre mimétique candidat au sein du cadre candidat considéré sont :

cadremimwithin : présence de cadres mimétiques au sein du cadre mimétique

autrescadresmim : nombre d'autres cadres mimétiques candidats au sein du cadre phrastique

distplusprochecadre : distance du cadre mimétique le plus proche dans le contexte phrastique

L'attribut rattaché à l'indice de la structure en incise est **presenceincise** il indique la présence d'une incise au sein du cadre mimétique candidat.

Les attributs liés aux pronoms à la première et deuxième personne sont :

pronoms12-within : présence de pronoms personnels à la 1er et 2e personne au sein du cadre mimétique candidat

pronoms12-out : présence de pronoms personnels à la 1er et 2e personne en-dehors du cadre mimétique candidat

Toujours à propos des pronoms, mais rattaché à la classe d'indices des pronoms à la troisième personne :

pronoms3-within : présence de pronoms personnels à la 3e pers au sein du cadre mimétique candidat

pronoms3-out : présence de pronoms personnels à la 3e personne en-dehors du cadre mimétique candidat

Les attributs reliés à la classe d'indices des verbes d'énonciation sont les suivants :

enoncverb-within : présence de verbes d'énonciation au sein du cadre mimétique candidat

enoncverb-out : présence de verbes d'énonciation en-dehors du cadre mimétique candidat

distenoncverb-within : distance du verbe d'énonciation interne au cadre mimétique par rapport au bord le plus proche du cadre mimétique candidat

distenoncverb-out : distance du verbe d'énonciation externe au cadre mimétique par rapport au bord le plus proche du cadre mimétique candidat

Enfin, les attributs finalement rattachés aux indices verbaux sont :

verbque-within : présence de "verbes + que" au sein du cadre mimétique candidat

verbque-out : présence de "verbes + que" en-dehors du cadre mimétique candidat

distverbque-within : distance du "verbe + que" interne au cadre mimétique par rapport au bord le plus proche du cadre mimétique candidat

distverbque-out : distance du "verbe + que" externe au cadre mimétique par rapport au bord le plus proche du cadre mimétique candidat

verbwithin : présence de verbe au sein du cadre mimétique

Finalement, s'ajoute au sein de chaque instance, en dernière position, l'attribut **vraiementmimetique** qui définit si le cadre mimétique candidat se rattache à un segment citationnel d'après notre corpus.

Caractérisation automatique des instances

La segmentation du corpus en cadres du discours a extrait 463 cadres mimétiques candidats, présents au sein de 1327 cadres phrastiques. À raison de 21 attributs par cadre mimétique candidat, cela représente près de 10 000 caractérisations à effectuer. Nous ne pouvons raisonnablement pas réaliser la compilation des données d'apprentissage manuellement.

Nous avons développé un composant "ReglesCadreMimetique" qui prend en entrée des cadres mimétiques candidats accompagnés de l'article et de la phrase auxquels ils appartiennent. Le composant extrait alors de ces données les valeurs des attributs du cadre mimétique, les ordonne et génère ainsi une règle d'apprentissage. Le composant a la capacité, lorsque l'utilisateur le juge nécessaire, de compiler toutes les règles qu'il a généré sous la forme d'un fichier *arff* importable dans Weka.

L'interface de ce composant se compose donc de deux méthodes dédiées aux données d'apprentissage : *computeNewRule* et *writeARFF*. Plusieurs méthodes de configuration du composant permettant de décrire les différentes balises à utiliser complètent cette interface. Les méthodes permettant de calculer les valeurs des différents attributs sont privées et donc inacessibles à l'extérieur du composant.

Étant donné l'importance de la qualité de l'extraction automatique des attributs, il serait nécessaire de tester efficacement les méthodes de calcul des attributs. En effet, une extraction erronnée de ces derniers impacterait directement l'apprentissage supervisé, et le modèle extrait de cet apprentissage. Nous pensons qu'il est nécessaire d'effectuer des tests unitaires sur l'ensemble de ces scripts d'apprentissage afin de s'assurer de leur bon fonctionnement. Nous nous sommes, faute de temps, limités aux tests manuels sur des exemples tirés du corpus.

Le composant "ReglesCadreMimetique" nous a permis de générer à partir de notre corpus annoté un fichier *arff* contenant les règles d'apprentissage nécessaires à la génération d'un classifieur à partir de Weka. L'exécution du composant nous a permis de récupérer des statistiques intéressantes sur les indices discutés au sein du chapitre précédent. Ainsi, on trouve 736 occurrences de verbes d'énonciation et 319 syntagmes prépositionnels répartis au sein de 1327 phrases. De plus, à peine 15% des pronoms (tous styles confondus) sont à la première ou deuxième personne. Enfin, la détection d'uniquement 4 incises au sein du corpus nous pousse à penser que la méthode permettant leur détection est défaillante. Nous ne sommes toutefois pas en mesure d'évaluer l'impact de cette défaillance sur nos résultats.

4.3.2 Choix des attributs et composant pour la caractérisation des expressions locuteurs

Le choix des attributs de description des segments textuels candidats à l'investiture d'expression locutrice est beaucoup moins aisé. En effet, nous n'avons pas trouvé de travaux linguistiques ou informatiques traitant de la description des locuteurs. Nos choix concernant les attributs sont donc totalement intuitifs

et sans lien avec quelconques travaux antérieurs. Nous présentons donc ces attributs dans les chapitres suivants en insistant particulièrement sur le caractère intuitif de leur sélection.

Les attributs reposant sur la morphologie du candidat sont les suivants :

nominalcap : le candidat contient ou est contenu dans un groupe nominal possédant des noms capitalisés ;

size : correspond à la taille du candidat en nombre de tokens ;

Nous utilisons également des attributs syntaxiques :

nb-defnom : nombre de groupes "article défini + nom" présents ;

dist-enoncverb : distance avec le verbe d'énonciation le plus proche ;

dist-syntprep : distance avec la préposition du syntagme prépositionnel le plus proche ;

within-segcoma : le candidat se trouve au sein d'un segment entre virgules ;

within-mimfrm : le candidat se trouve au sein d'un segment au sein d'un cadre mimétique candidat.

Enfin, nous utilisons également des attributs ayant trait à la nature du candidat :

propers3 : présence de pronom personnel à la 3e personne au sein du candidat ;

propers12 : présence de pronom personnel à la 1e ou 2e personne au sein du candidat ;

nompropre : présence d'un nom propre — d'après Tree-Tagger — au sein du candidat.

Une fois de plus nous insistons sur le fait que ces attributs sont purement intuitifs et ne reposent nullement sur des travaux antérieurs sur le sujet. Il serait intéressant d'effectuer une étude linguistique plus poussée visant à caractériser l'introduction des locuteurs au sein des textes.

De la même manière que nous avions développé un composant *ReglesCadreMim* pour l'extraction automatique et la génération des données d'apprentissage pour les cadres mimétiques, nous avons développé un composant *ReglesSources*. Ce dernier a une interface similaire au composant *ReglesCadreMim*.

Une fois le choix des attributs définis et les quelques 3600 instances d'apprentissage extraites du corpus, nous pouvons passer à la phase d'apprentissage.

4.4 Expérimentations et évaluation de nos chaînes d'identification

Les dernières étapes dans la réalisation de l'apprentissage sont le choix de l'algorithme d'apprentissage, l'exécution de cet algorithme sur l'ensemble de nos données et finalement l'évaluation de l'efficacité du modèle généré.

Choix de l'algorithme d'apprentissage

Étant donné la nature de nos attributs qui prennent tous des valeurs discrètes et dont les domaines sont binaires, nous penchons pour l'utilisation d'un algorithme de type "arbre de décision". L'un des algorithmes les plus performants actuellement dans la génération d'arbres de décisions à partir de valeurs discrètes et **C4.5**.

Nous faisons appel, à ce stade du stage, à l'excellent logiciel libre Weka [Witten & Frank, 2005] pour la phase d'apprentissage. Weka est un ensemble d'outils permettant d'organiser des ensembles de données d'apprentissage, de réaliser l'apprentissage sur ces données à l'aide d'implémentations libres d'une grande quantité d'algorithmes d'apprentissage, et également de visualiser les résultats de ces apprentissages. Cet outil est une référence dans son domaine. Weka propose une implémentation libre de cet algorithme : **J48**.

Nous nous sommes tourné vers cet algorithme, et après avoir fait tourner plusieurs algorithmes de type arbre de décision tels que **Id3** ou **NBTree** sur un échantillon du corpus aux côtés de **J48**, ce dernier s'est effectivement révélé le plus efficace.

Données d'apprentissage

Chaque instance soumise à un algoritme d'apprentissage de type "classifieur" décrit un objet que l'on veut classer à l'aide de deux types d'attributs : des caractéristiques que l'on suppose d'intérêt pour la classification et l'attribut indiquant la classe à laquelle l'objet appartient réellement.

Pour la phase d'apprentissage, nous déduisons cette dernière valeur de l'annotation des segments citationnels au sein du corpus. Nous considérerons ainsi les cadres mimétiques candidats prenant part à un segment textuel annoté au sein du corpus à l'aide de la balise <*cite :discours*> comme appartenant à la classe "cadre mimétique". Par "prenant part", nous entendons que le candidat est entièrement contenu ou bien contient complètement un segment balisé par <*cite :discours*>.

Il en va de même avec les expressions locuteurs et les balises correspondantes.

Mode d'évaluation

L'*approche de base (ou baseline)* que nous choisissons et qui nous servira de repère pour discuter nos résultats d'apprentissage a pour comportement de classer toutes les données selon la classe la plus représentative des données d'apprentissage. Nous comparerons nos techniques avec cet algorithme naïf peut coûteux à mettre en place.

Pour des raisons de taille de données d'apprentissage nous choisissons d'évaluer nos techniques par *validation croisée (ou K-fold cross-validation)*. La validation croisée est une méthode permettant d'évaluer le taux d'erreur d'un apprentissage. Cette technique consiste à partitionner notre jeu de données en K partitions dont les éléments sont sélectionnés de manière aléatoire. L'apprentissage est ensuite effectué en prenant chacune des partitions comme ensemble de test et les éléments complémentaires comme ensemble d'entraînement. Le résultat de l'évaluation correspond alors à la moyenne des résultats pour chacune des partitions.

Weka intègre une fonction de validation croisée. Nous choisissons de partitionner notre jeu de données en dix partitions. Ce chiffre est choisi plus ou moins arbitrairement. Ce partitionnement est toutefois le plus rencontré dans les articles publiés traitant d'apprentissage supervisé sur des données linguistiques et semble donner des résultats satisfaisant.

4.4.1 Réconnaissance des cadres mimétiques

Pour cette tâche, l'approche de base (ou *baseline*) obtient des mesures de précision et de rappel respectivement de 0.793 et de 1.

L'algorithme J48 permet de générer un classifieur de type "arbre de décision" à partir de notre ensemble de instances. Les noeuds d'un tel arbre correspondent aux attributs et l'on suit un chemin partant de la racine vers les feuilles en suivant les valeurs des arcs correspondant à la valeur de l'attribut du noeud précédent pour la description de notre objet à classer.

La figure 4.4.1 représente l'arbre de décision généré par J48 d'après les instances extraites de notre corpus. Les noeuds les plus proches de la racine représentent les attributs les plus discriminants dans la prise de décision. Ainsi, sans trop de surprise, la racine correspond à la taille du cadre mimétique. Les noeuds suivants sont moins évidents mais correspondent à des attributs qui nous semblaient réellement

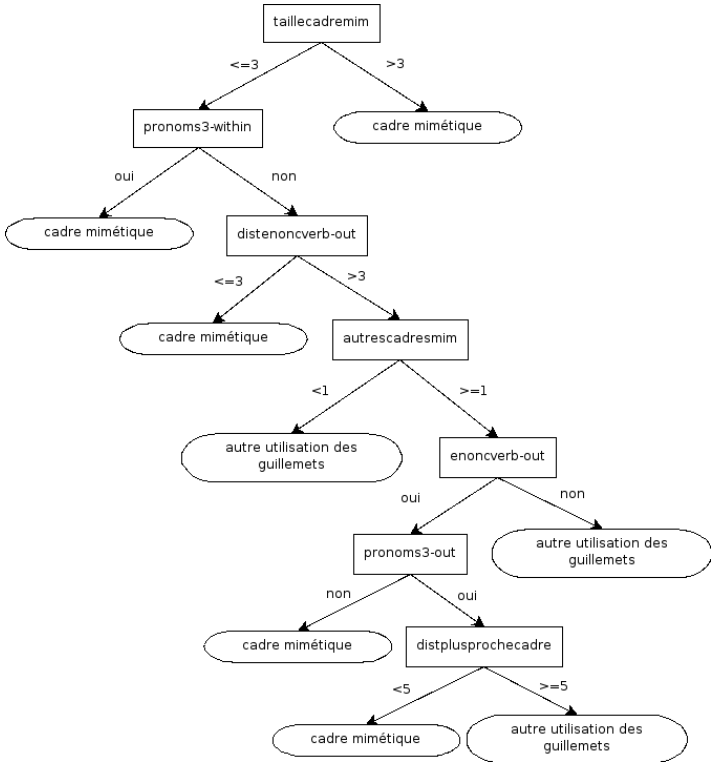


FIG. 4.2 – Arbre de décision classant les cadres mimétiques candidats

intéressants pour caractériser les cadres mimétiques avérés. Nous sommes toutefois surpris de l'absence d'attribut lié aux indices de type pronoms personnels aux deux premières personnes.

L'expérimentation du modèle sur une portion représentant 10% des éléments du corpus est assez concluante. Ainsi, plus de 87% des cadres mimétiques candidats sont correctement classés. La précision correspondante est *0.916* et le rappel *0.926*.

Le modèle déduit par J48 sur notre ensemble d'instances semble valide et performant. L'expérimentation du paragraphe précédent n'est cependant pas suffisamment valide à nos yeux, aussi nous choisissons de réaliser une validation croisée dans le but d'obtenir un taux d'erreur plus fiable.

Les résultats obtenus par validation croisée sont présentées ci-dessous.

Correctly Classified Instances	390	84.2333%
Incorrectly Classified Instances	73	15.7667%
Total Number of Instances	463	

D'après ces résultats, en moyenne, 84% des cadres mimétiques candidats sont correctement identifiés comme cadres mimétiques ou bien comme emphases. A contrario, 16% de ces cadres mimétiques candidats sont mal classés. Une instance est mal classée si jamais il ne s'agit pas réellement d'un cadre mimétique mais que la cadre candidat soit classé comme tel, ou alors lorsque que l'on est face à un véritable cadre mimétique et que l'on ne parvient pas à le classer comme tel.

Du point de vue de notre méthode, nous préférons obtenir des faux positifs plutôt que des faux négatifs car nous pouvons ignorer les premiers par la suite. En effet, s'il ne s'agit pas réellement d'un cadre mimétique, il n'y aura probablement pas d'expression locutrice à laquelle le rattacher. Cela permettrait de rattrapper quelque peu un taux d'erreur trop important. La valeur de la précision et du rappel pour chacune des classes nous permet ainsi d'évaluer la performance par classe.

Classe	Precision	Recall	F-Measure
"cadre mimétique"	0.895	0.907	0.901
"autre utilisation des guillemets"	0.626	0.594	0.61

La classe "cadre mimétique" correspond à la détermination des cadres mimétiques avérés. Il s'agit donc de la classe ayant l'impact le plus important sur notre méthode. Sa F-measure (somme harmonique pondérée de la précision et du rappel) de 0.901 est plutôt bon signe. Notre modèle permet de repérer correctement près de 90% des cadres mimétique candidats avérés.

Par rapport à notre approche de base offrant une précision de 0.793, ce résultat présente un gain de près de 13% et nous semble par conséquent intéressant. Toutefois il nous semble possible de faire mieux.

Notre modèle de repérage des cadres mimétiques permet de repérer environ 9 cadres mimétiques avérés sur 10, à supposer que le repérage des cadres mimétiques candidats soit correct. Nous pourrions affiner ce résultat en estimant manuellement le taux d'erreur de la détection automatique des cadres mimétiques candidats, toutefois nous nous en contenterons.

4.4.2 Reconnaissance des expressions locuteurs

Nous sommes tout d'abord reparti sur une démarche en tous points identiques à celle que nous avions abordée pour la classification des cadres mimétiques candidats. L'objectif est similaire, à savoir classer des expressions locutrices potentielles comme "expressions locutrices effectives" ou non.

L'apprentissage, très similaire à celui des cadres mimétiques, a pour objectif d'extraire un modèle permettant de classer les candidats dans l'une des catégories "expression locutrice" ou bien "ne correspond pas à une expression locutrice".

Nous relatons ci-dessous quelques expériences que nous avons menées successivement en faisant varier le rapport “nombre d’exemples par classe” dans le corpus d’apprentissage afin d’obtenir des résultats un tant soi peu significatif.

Expérience 1 : 105 valides pour 3499 non valides

Satisfaits des résultats renvoyés par l’algorithme J48, nous l’avons de nouveau utilisé dans un premier. Son inefficacité, comme celle des autres algorithmes de type "arbre de décision" nous a poussé à revoir notre approche.

À noter que pour cette tâche, l’approche de base obtient des mesures de précision et de rappel respectivement de 0 et de 0 pour la classe "locuteur" et 0.971 et 1 pour la classe "non locuteur". En d’autres termes, l’approche de base ne détecte aucune expression locuteur.

Nous avons été partiellement surpris du résultat de l’apprentissage sur notre ensemble de instances fraîchement extrait du corpus. Non seulement l’apprentissage sur ces instances à l’aide de J48 nous retourne une précision nulle pour la classe "source", mais ce résultat est identique avec d’autres algorithmes de type “arbre de décision” algorithmes LMT, REPTree, DecisionStump, ADTRee ou encore RandomTree. Tous les modèles générés à partir de ces algorithmes se résument à une feuille "ne correspond pas à une expression locutrice". Les algorithmes RandomForest, Id3 ou encore NBTree quant à eux, même s’ils offrent une meilleur précision pour cette classe, ne proposent toujours pas un modèle viable. Les résultats des différents algorithmes sont résumés dans le tableau ci-dessous :

Algorithme	Classe	Précision	Rappel	F-Measure
J48	“locuteur”	0	0	0
	“non locuteur”	0.971	1	0.985
LMT	“locuteur”	0	0	0
	“non locuteur”	0.971	1	0.985
REPTree	“locuteur”	0	0.001	0
	“non locuteur”	0.999	1	0.971
DecisionStump	“locuteur”	0	0	0
	“non locuteur”	0.971	1	0.985
ADTree	“locuteur”	0	0	0
	“non locuteur”	0.971	1	0.985
RandomTree	“locuteur”	0	0	0
	“non locuteur”	0.971	0.998	0.984
RandomForest	“locuteur”	0.143	0.01	0.018
	“non locuteur”	0.971	0.998	0.984
Id3	“locuteur”	0.273	0.029	0.052
	“non locuteur”	0.972	0.998	0.984
NBTree	“locuteur”	0.5	0.01	0.019
	“non locuteur”	0.971	1	0.985

Nous avons, dans un premier temps, émis l’hypothèse d’une implémentation erronée des méthodes du composant permettant l’extraction automatique des exemples à partir du corpus. Cependant, ces méthodes d’extraction se basent exactement sur le même code que les méthodes du composant d’extraction automatique des exemples pour les cadres mimétiques, seul la description des balises est modifiée. Nous

pouvons toutefois remettre en cause les valeurs de l'attribut *nompropre*. En effet, le repérage de ces éléments est basé sur l'étiquetage de Tree-Tagger qui ne semble pas réellement performant dans cette tâche. L'attribut *nompropre* est, avec *nominalcap*, l'un des attributs les plus discriminants selon nous. Un mauvais repérage à la base de ces noms propres peut effectivement avoir des conséquences importantes sur la qualité de l'apprentissage.

En nous concentrant sur l'échantillonnage de nos exemples, nous pensons toutefois avoir trouvé un élément plus flagrant expliquant l'échec de l'apprentissage. Ainsi, le rapport du nombre de exemples pour les expressions locutrices valides par rapport aux autres est de 105 pour 3499, soit à peine 3%. Les algorithmes de type arbres de décision ayant tendance à niveler les aspérités de cet ordre, ils décident grossièrement que considérer la totalité des instances comme n'étant pas des expressions locutrices avérées offre une précision suffisamment bonne.

Expérience 2 : 105 valides pour 300 non valides

Nous décidons donc de réitérer l'apprentissage automatique sur les expressions locutrices en tentant de rééquilibrer le nombre d'instances correspondant à des expressions locutrices par rapport aux autres.

À noter que pour cette tâche, l'approche de base obtient toujours des mesures de précision et de rappel respectivement de 0 et de 0 pour la classe "locuteur" et cette fois 0.741 et 1 pour la classe "non locuteur". L'approche de base ne détecte donc toujours pas la moindre expression locuteur.

Nous discutons dans un premier temps les modifications apportées à l'ensemble des instances, puis dans un deuxième temps les résultats obtenus à l'aide de ce nouvel échantillon.

Rééchantillonnage des exemples extraites du corpus

L'hypothèse que nous avons émise sur les causes de l'échec de la première séance d'apprentissage était le nombre d'instances de chacune des classes était bien trop déséquilibré. Ce déséquilibre provoquerait le nivellement, par les algorithmes d'apprentissage, en un ensemble constitué uniquement de classe majoritaire. Nous proposons donc de réitérer l'apprentissage en rééquilibrant les deux classes.

Notre ensemble d'instances est constitué de 105 exemples pour la classe "locuteur" et 3499 pour "non locuteur". Nous proposons de rééquilibrer ce nombre à 105 éléments pour la classe "locuteur" et 300 pour la classe "non locuteur". Nous conservons une majorité d'instances non locutrices étant donné que le corpus en est majoritairement constitué. La sélection des 300 instances parmis les 3499 s'est faite de manière totalement aléatoire.

Le rééquilibrage de l'échantillon, même si il ne fixe pas les écueils présentés plus haut, devrait permettre d'obtenir un meilleur modèle de classifieur pour les expressions locutrices candidates.

Réitération de l'apprentissage

Après l'échec de l'apprentissage supervisé sur notre premier échantillonnage, nous avons décidé de rééquilibrer le nombre d'instances de chaque classe en réduisant drastiquement, et de manière aléatoire, le nombre d'instances correspondant à la classe des "non locuteurs".

Nous avons de nouveau utilisé l'algorithme J48 pour l'apprentissage sur l'échantillon nouvellement constitué. Contrairement aux phases d'apprentissage sur l'échantillon précédent, l'algorithme a généré un arbre de décision à plusieurs feuilles (*cf figure refarbresdec_sources*).

Il est intéressant d'observer le premier noeud de l'arbre. Ce dernier correspond à l'attribut que nous avons utiliser pour compléter le mauvais repérage des entitées nommées par Tree-Tagger. Sur les 405

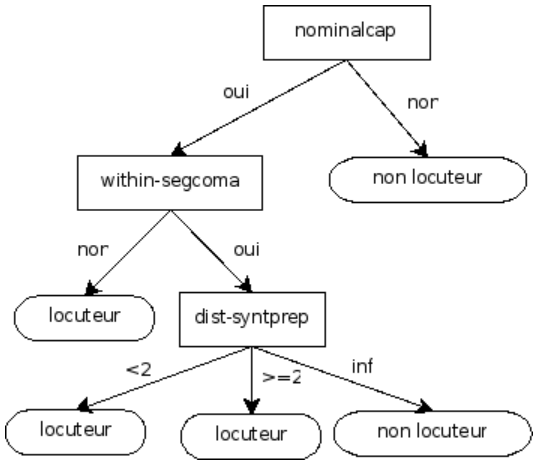


FIG. 4.3 – Arbre de décision généré sur l'échantillon rééquilibré

instances que comporte notre nouvel échantillon, 348 sont directement classées dans "non locuteur" au passage par ce noeud, soit près de 86%. Cela fait de l'attribut *nominalcap* un attribut extrêmement discriminant alors qu'il est partiellement erronné. Ainsi, la majuscule peut provenir du fait que le mot se situe en début de phrase. L'utilisation d'un véritable outil permettant l'identification d'entités nommées tel que Némesis devrait donc permettre d'augmenter notablement la qualité des instances et donc du modèle généré lors de l'apprentissage.

De la même façon que pour les cadres mimétiques, nous effectuons une validation croisée sur l'échantillon pour obtenir une information plus fiable sur le taux d'erreur de l'algorithme. Les résultats de cette validation croisée sont présentés dans le tableau ci-dessous :

Correctly Classified Instances	335	82.716%
Incorrectly Classified Instances	70	17.284 %

La précision moyenne obtenue lors de cette validation croisée est plutôt moyenne pour les deux classes, soit 0.807 pour "locuteur" et 0.83 pour "non locuteur". Le rappel est toutefois très bon pour "non locuteur" (0.963) et franchement mauvais pour "locuteur" : 0.438. La tableau ci-dessous représente la matrice de confusion, ie la classification des différentes expressions locutrices candidates, selon leur classe réelle :

	classé dans "locuteur"	classé dans "non locuteur"
"locuteur"	46	59
"non locuteur"	11	289

La matrice de confusion révèle clairement la mauvaise classification des véritables expressions locutrices puisque plus de la moitié d'entre elles sont classées comme comme "non locuteur". Contrairement aux cadres mimétiques, nous préférons privilégier une juste reconnaissance des vrais locuteurs quit à ce que cela se reporte par une augmentation des faux positifs. En effet, ces derniers devraient être supprimmées par la suite si ils ne sont pas mis en concordance avec un texte englobé.

Expérience 3 : 105 valides pour 150 non valides

Nous avons expérimenté un nouvel échantillonnage afin de ramener le nombre d'instances de "non locuteur" à 150, le tableau ci-dessous présente la matrice de confusion qui en ressort :

	classé dans "locuteur"	classé dans "non locuteur"
"locuteur"	59	46
"non locuteur"	10	140

Les résultats semblent meilleurs. L'approche de base quant à elle ne détecte toujours pas de locuteur.

Expérience 4 : projection des différents modèles sur tout le corpus en données de test

Si les résultats semblent meilleurs, nous préférons nous en assurer en appliquant à chacun des modèles la liste complète des 3604 instances telle qu'elle a été originalement extraite du corpus. Le résultat de cette évaluation pour le modèle extrait de l'échantillon de 405 instances est décrit dans le tableau ci-dessous :

Correctly Classified Instances	3415	94.7558 %
Incorrectly Classified Instances	189	5.2442%
"locuteur" class precision	0.267	
"locuteur" class recall	0.457	
"non locuteur" class precision	0.983	
"non locuteur" class recall	0.962	

Ainsi que celui de l'échantillon de 255 instances :

Correctly Classified Instances	3121	86.5982 %
Incorrectly Classified Instances	483	13.4018 %
"locuteur" class precision	0.126	
"locuteur" class recall	0.61	
"non locuteur" class precision	0.987	
"non locuteur" class recall	0.874	

Soit les matrices de confusions respectives du modèle issue de l'échantillon de 405 instances et de celui de 255 instances :

	classé dans "locuteur"	classé dans "non locuteur"
"locuteur"	48	57
"non locuteur"	132	3367

	classé dans "locuteur"	classé dans "non locuteur"
"locuteur"	64	41
"non locuteur"	442	3057

La réduction de l'échantillon si elle semble fournir un modèle offrant de meilleurs résultats au premier abord génère une augmentation significative (330%) du nombre de faux positifs alors qu'elle ne permet qu'une amélioration de 130% pour le repérage correct des instances de la classe "locuteur". Nous préférons donc conserver le modèle issue de l'échantillon de 405 instances.

L'échec partiel de l'apprentissage supervisé sur les expressions locutrices s'expliquer par plusieurs écueils tout au long de l'extraction. Tout d'abord, la sélection des candidats nécessite d'être affiné. En effet, plus de 3600 candidats sont proposés sur un ensemble de 1327 phrases, soit près de trois candidats par phrase. Les prochaines recherches doivent permettre de réduire ce nombre en ne s'intéressant par exemple uniquement aux phrases présentant au moins un indice indiquant la présence d'un locuteur. De plus, seules 105 instances appartiennent à la classe "locuteur" cela nous laisse penser que de nombreuses expression locutrices avérées ont été laissées de côté.

L'incapacité de notre chaîne de traitement à repérer les entitées nommées est une seconde explication de cet échec. L'utilisation de Tree-tagger couplée à la prise en compte des noms capitalisés n'atteint nullement la qualité de repérage que pourrait offrir un outil dédié comme Némesis. Étant donné l'importance de l'attribut *nominalcap* dans les modèles générés, nous ne pouvons qu'être confiant dans l'apport d'une technologie.

Expérience 5 : les attributs les plus discriminants

Finalement, nous avons voulu évalué l'apport de chacun des attributs proposé à l'aide de l'algorithme *AttributeSelection* de Weka. Ce dernier algorithme extrait de la liste des attributs sont qui semblent avoir un intérêt pour un apprentissage automatique supervisé. Il en ressort que seuls les attributs *nominalcap*, *dist-syntprep* et bien entendu *reallysrc* sortent indemnes de cette sélection. Nous en déduisons donc que les attributs que nous avons sélectionné pour la caractérisation des expressions locutrices ne parviennent pas à décrire assez précisément les instances extraites du corpus.

4.5 Synthèse

En raison des résultats médiocres obtenus sur la génération d'un modèle sur la reconnaissance des expressions locutrices, il ne nous a pas semblé pertinent de continuer l'expérimentation sur le repérage des segments citationnels. En effet, la présence trop importante de faux positifs au niveau des locuteurs ainsi que l'incapacité à repérer convenablement les véritables expressions locutrices nous laissent penser que la considération unique des cadres mimétiques repérés offre de meilleurs résultats que ce que nous pourrions obtenir à l'aide de nos exemples définies dans le chapitre précédents.

Chapitre 5

Conclusion

5.1 Synthèse

Le terme citation est généralement utilisé à la fois pour le processus de prélèvement et d'intégration d'un fragment d'un énoncé au sein d'un autre, et également pour le produit fini de ce processus. Notre étude a porté uniquement sur le produit fini au sein des textes journalistiques et donc toute référence au terme citation prend ce sens. Les travaux antérieurs sur le sujet emploient des terminologies distinctes, nous avons donc tenté de définir une terminologie nous permettant de désambiguïser les termes utilisés.

Une citation est le résultat de l'intégration au sein d'un texte englobant d'un extrait d'énoncé produit par un locuteur (ou un scripteur) externe à l'aide des styles de discours direct et indirect et leurs variantes (DI avec îlot textuel et DI quasi-textuel) issues du besoin de contraction des extraits par les journalistes. L'énonciation du locuteur externe et celle du locuteur courant se déroulent dans des situations d'énonciation distinctes. Les journalistes tentent de distinguer leurs plans d'énonciation de ceux de leurs sources au sein des articles.

La constitution de notre corpus nous a permis de décrire à l'aide d'exemples réels les différentes formes de citations. Nous avons ainsi pu extraire de la cinquantaine d'articles composant notre corpus un peu plus de 350 formes citationnelles. Nous avons décrit qualitativement ces formes au sein du chapitre 2.

Après avoir tenté de poser et résoudre la problématique de l'unité citationnelle au niveau linguistique, nous avons appliqué un schéma d'annotation XML sur les citations du corpus. Ce schéma permet d'identifier les expressions locuteur et les éléments de discours rapporté rattachés les uns aux autres et ainsi détecter les citations. Nous avons alors cherché à mettre en place une méthode de repérage de ces citations permettant l'identification automatique de ces éléments.

Nous avons introduit les espaces de recherche cadres mimétiques et cadres phrastiques (associés à des réalités linguistiques), que nous avons complétés par l'espace conceptuel cadre du discours rapporté. L'étude de ces espaces de recherche a conduit à la création du segment citationnel : segment textuel ayant une réalité linguistique et représentant au mieux le concept de citation.

La méthode d'identification automatique que nous proposons se base sur la segmentation des textes en cadres mimétiques candidats et cadres phrastiques. S'ensuit la recherche au sein des cadres mimétiques candidats des marques du discours rapporté entre guillemets et au sein du cadre phrastique des marques de la présence d'expressions locuteur. Finalement, l'application des arbres de décision obtenus lors des phases d'apprentissage supervisé associée à des règles simples répondant aux différents scénarios de constitution des cadres phrastiques auraient permis l'identification des segments citationnels.

Si le modèle obtenu par apprentissage pour la classification des cadres mimétiques candidats apporte un gain de plus de 10% par rapport à la *baseline*, le modèle concernant la classification des expressions locuteur candidates offre une précision trop faible pour être considéré viable, et ce malgré différentes tentatives d'amélioration de l'apprentissage.

5.2 Perspectives

Le choix des expressions locuteur candidates et leur caractérisation est directement remis en cause aux vues des résultats de l'apprentissage automatique. L'utilisation effective de Némesis (outil de repérage des entités nommées) devrait nous permettre de réduire le nombre d'expressions locuteur candidat dans un premier temps, puis de permettre une meilleure caractérisation de ces dernières.

L'absence de résultats concernant la détection automatique des citations à l'aide des techniques déjà existantes ne permet pas de réellement juger de l'efficacité de notre algorithme, au moins pour l'identification des cadres mimétiques. L'application d'implémentations de ces méthodes sur notre corpus nous permettrait d'obtenir une précision et un rappel de référence permettant de juger des améliorations à apporter et de comparer l'efficacité de la détection selon les formes de citations.

Finalement, notre méthode se basant sur un prétraitement des textes, il serait intéressant d'évaluer l'impact des erreurs de segmentation imputées à ce prétraitement sur la performance globale de l'algorithme.

Annexes

Annexe A

Réponse du journal *Le Figaro*

Monsieur,

Nous vous autorisons à utiliser les articles suivants dans le cadre
de vos recherches universitaires:

http://www.lefigaro.fr/election-presidentielle-2007/20070220.FIG000000207__deux_
http://www.lefigaro.fr/france/20070220.WWW000000363_disparition_de_julien_la_pis
http://www.lefigaro.fr/election-presidentielle-2007/20070219.WWW000000502_le_fin
http://www.lefigaro.fr/election-presidentielle-2007/20070219.WWW000000587_le_gel
http://www.lefigaro.fr/election-presidentielle-2007/20070220.FIG000000204_le_pen
http://www.lefigaro.fr/france/20070220.WWW000000360_les_memoires_secretes_de_pap
http://www.lefigaro.fr/election-presidentielle-2007/20070220.FIG000000193_nicola
http://www.lefigaro.fr/france/20070219.WWW000000637_pour_jean_luc_delarue_ca_se_
http://www.lefigaro.fr/election-presidentielle-2007/20070220.FIG000000200_segole
http://www.lefigaro.fr/sciences/20070220.FIG000000011_un_plan_mondial_contre_les
http://www.lefigaro.fr/france/20070220.FIG000000215_un_superprefet_au_chevet_des

C'est en effet, la diffusion qui pose le plus de soucis. Nous vous
autorisons dans le cadre de vos recherches universitaires à diffuser
votre corpus au sein de vos collègues, cependant l'édition à plus
grande échelle (étudiants ou grand public) de vos recherches devra
faire l'objet d'une demande ultérieure.

Bien cordialement,

Marie-Céline Courtet

Annexe B

Guidelines de caractérisation du corpus

B.1 Type du discours rapporté

Dans [Mourad & Desclès, 2001], les auteurs proposent de considérer deux catégories de citations. La première, reconnaissable à ces marques typographiques, correspondant au discours direct. La deuxième pour la forme indirecte du discours.

D'autres auteurs ont proposé d'introduire la catégorie du discours indirect libre aux deux précédentes.

Enfin, après avoir parcouru quelques articles de presse, il m'a paru judicieux de considérer la possibilité de voir plusieurs styles mélangés pour une même citation. Une sorte de discours hybride à mi-chemin entre le discours direct et le discours indirect.

B.1.1 Discours direct

Dans les articles de la bibliographie, le discours direct n'est clairement défini que par [Mourad & Minel, 2000] où il indique que le discours direct se différencie du discours indirect par la présence de marqueurs typographiques.

Cette définition est renforcée par la présentation du discours direct par des sites généralistes sur l'enseignement du français tels que [let,]. En effet, ces derniers le définissent comme un type de discours rapporté dans lequel les paroles ou les pensées sont rapportées directement, entre guillemets.

J'ai donc considéré comme citations directes les segments citationnels où les propos rapportés étaient introduits par l'ouverture d'un guillemet et conclus par la fermeture d'un guillemet.

> Mais Brice Hortefeux précise que " le seul objectif de cette fusion, c'est de parvenir à une plus grande synergie et à une plus grande coordination de ceux qui s'expriment au nom du candidat ".
>
> *Exemple de discours direct extrait du corpus (©Le Figaro)*

B.1.2 Discours indirect (ou discours indirect lié)

Le discours indirect n'est pas clairement défini, en réalité, il est souvent assimilée comme la catégorie complémentaire au discours direct. Cependant, cette définition ne peut pas nous satisfaire pleinement étant donné que nous avons quatre catégories de discours définies.

D'après [let,], il s'agit d'un type de discours rapporté par lequel les paroles ou les pensées sont rapportées indirectement, à l'aide de subordonnées. Contrairement au dicours direct, le discours indirect

n'est pas délimité par des éléments typographiques. Si toutefois des guillemets étaient utilisés en son sein, l'utilisation serait celle de mise en valeur de l'entité entre guillemets, et non la délimitation des propos rapportés.

> il a déclaré qu'il pourrait nommer un premier ministre de gauche, s'il était élu président de la République.

Exemple de discours indirect extrait du corpus (©Le Monde)

B.1.3 Discours indirect libre

Le discours indirect libre est une catégorie particulière. Certains ([Mourad & Minel, 2000], [Mourad & Desclès, 2001 ou [Mourad & Desclès, 2002]) la considèrent comme une sous-catégorie du discours indirect. D'un autre côté, le linguiste Roman Jakobson, considère cette catégorie comme une catégorie à part entière (vérifier sources).

Toujours selon [let,], le discours indirect libre se détache du discours indirect par la suppression des verbes introducteurs et des subordonnées. De plus, les pronoms, adverbes et les temps grammaticaux étant ceux du récit, cela en fait un type de discours rapporté ardu à distinguer du récit.

Contrairement aux deux discours précédents qui sont – dans la plupart des cas – introduits dans un contexte phrastique par une subordonnée issue du récit, dans le discours indirect libre les propos rapportés sont les éléments principaux de la phrase (subordonnée principale).

> Il met bas son fagot, il songe à son malheur. / Quel plaisir a-t-il eu depuis qu'il est au monde ?

Exemple tiré de la fable La Mort et le Bûcheron de Jean de La Fontaine

> Selon un policier binchois, habitué de l'événement, 80.000 à 100.000 personnes auraient rejoint la Cité du Gille.

Exemple de discours indirect libre extrait du corpus (© Le Soir)

B.1.4 Mix de styles

La forme des articles de presse et les règles éditoriales spécifiques aux différents journaux contraignent souvent les journalistes à adapter les citations pour pouvoir les réduire en taille, tout en conservant l'idée originale et les propos forts qui illustrent les idées qu'il veulent mettre en avant.

Cette combinaison semble faire apparaître un style de discours particulier où l'idée générale du propos rapporté est reformulée de manière synthétique au discours indirect, mais accompagnée d'expressions exactes entre guillemets – discours direct – afin d'appuyer la véracité et l'authenticité des dits propos.

> c'est depuis vingt-cinq ans " l'une des économies les plus dynamiques du monde ", note l'OCDE.

Exemple de discours rapporté mélangeant les styles extrait du corpus (©Challenges)

B.1.5 Cas litigieux

Différenciation entre discours direct et mélange de styles

Dans certains cas, les extrémités du propos rapporté sont clairement identifiés par des guillemets comme par exemple :

un de ses proches se réjouit : “ Il sait écouter et accorde le droit à l’erreur. ”

Exemple extrait du corpus (©Challenges)

Cependant, dans certains segments citationnels, le début – ou la fin – est moins bien marqué. Dans l’exemple ci-dessous, on peut se demander si "avec fierté" devrait faire partie du discours rapporté et auquel cas il s’agirait d’un mélange de styles plutôt que d’un discours direct :

Agon vante avec fierté “ les résultats spectaculaires ” .

Exemple extrait du corpus (©Challenges)

Cependant, l’expression "avec fierté" décrit la manière dont les propos sont introduits par la source, mais elle n’appartient pas au discours. Il ne faut donc par l’inclure dans le discours rapporté et considérer la citation comme appartenant au discours direct.

Dans le cas ci-dessous, toutefois, l’expression "c’est depuis vingt-cinq ans" appartient potentiellement aux propos tenus par la source :

c’est depuis vingt-cinq ans “ l’une des économies les plus dynamiques du monde ”, note l’OCDE.

Exemple extrait du corpus (©Challenges)

Il est donc nécessaire de classer le discours rapporté comme un mélange de style. En effet, la partie "c’est depuis vingt-cinq ans" est une reformulation des propos de la source, alors que "l’une des économies les plus dynamiques du monde" est une reprise verbatim. La concaténation des deux consiste donc à un mélange des styles du discours dans le même segment citationnel.

Discours direct sans guillemet

L’oublie de la fermeture d’un guillemet est une chose assez fréquente. Cependant, parfois il semble que le journaliste est oublié d’ouvrir et de fermer les guillemets pour rapporter des propos exacts :

Pour faire simple, dans quatre ans, un microprocesseur contiendra 32 milliards de transistors (100 fois plus qu’aujourd’hui) et sera doté d’une puissance phénoménale, nous explique le directeur du management des technologies d’Intel.

Exemple extrait du corpus (©Challenges)

Bien que cette citation ressemble fortement à une reprise verbatim des propos de la source, l’auteur ne l’a pas marqué typographiquement. Dans ce cas précis, d’après les règles présentées plus haut, le discours est considéré comme indirect, car rien ne nous indique qu’il s’agit effectivement des propos exacts, et non pas d’une reformulation s’approchant de ces propos exacts.

B.2 Source de la citation

La source est un élément clef de la citation. En effet, elle permet l’identification de l’origine des propos rapportés par l’auteur. Cependant, les sources sont polymorphiques, ce sont donc des entités dont il nécessaire de réaliser une description.

Un des objectifs de la caractérisation du corpus était de déterminer les proportions de cette polymorphie et donc connaître quelles étaient les formes les plus usités.

Il est important de noter que la source comprend non seulement la référence au locuteur mais également tous les éléments apportant des informations sur ce dernier et étant situé en dehors des propos rapportés et du relateur. Les caractérisations ci-dessous sont non-exclusives, i.e. il est possible qu'une source soit constitué de plusieurs des objets énoncés.

Pour Jacques Delpla, économiste à BNP Paribas, il ne s'agit que de " signaux politiques ",

Exemple extrait du corpus (©Challenges)

Dans l'exemple ci-dessus, l'on dénomme comme source, l'extrait : extit Jacques Delpla, économiste à BNP Paribas.

B.2.1 Source nommée

On entend "source nommée" dans le sens où la source est constituée d'au moins un élément capitalisé considéré comme le nom d'une personne ou d'une entité.

Steve Jobs estimait, ….
…estime Marc Menesguen, directeur général de la division produits de luxe.
Un porte-parole de la Maison Blanche a indiqué …

Exemples extraits du corpus (©Challenges, ©Le Soir)

B.2.2 Source pronominale

Une source est considérée "source pronominale" lorsqu'elle contient au moins un pronom personnel – pas impersonnel – sujet, complément d'objet direct ou indirect. Les pronoms personnels réfléchis ne sont pas pris en considération.

Le prochain chapitre à Bassora sera écrit par les Irakiens eux-mêmes, a-t-il dit
Elle a en revanche confirmé qu'il ne […]
Pour lui, " toute l'administration[…]

Exemples extraits du corpus (©Le Soir)

B.2.3 Source nominale

Une source est considérée "source nominale" lorsqu'elle contient au moins un groupe nominal composé d'un nom commun. Les noms communs peuvent être connectés entre eux à l'aide de conjonctions et/ou de déterminants.

[…] a admis hier, le procureur Brice Raymondeaud-Castanet.
le magistrat avait pointé sans relâche […]
Comme le conclut Françoise, 61 ans et deux paquets par jour depuis quarante-quatre ans, " ça n'est pas drôle de fumer ".

Exemples extraits du corpus (©Libération)

B.2.4 Source inconnue

Dans le cadre de la caractérisation, la recherche de la source se limite au contexte phrastique. Une source est donc considérée inconnue lorsqu'il n'est pas fait mention de cette dernière de manière explicite dans la phrase introduisant le discours rapporté.

> “ Pour la première fois était posée la question des "saisonniers" sous cet angle-là : leur qualité de travailleur permanent. ”
> l'affaire dite des bagagistes de Roissy révélait qu'un groupe de salariés soupçonnés de liens avec l'islamisme radical avait été écarté parce que présentant “ une vulnérabilité incompatible avec une habilitation d'accès en zone réservée ”.

Exemples extraits du corpus (©Libération)

B.3 Motif de la citation

L'expérimenation du corpus fût également une opportunité pour tester les méthodes proposées dans des travaux antérieures comme la méthode des motifs proposée par [Giguet & Lucas, 2004].

B.3.1 Schéma du motif

Contrairement à la proposition de [Giguet & Lucas, 2004] qui incluait la typographie au sein des entités extit source, extit relateur et extit discours. Dans le cadre de l'expérimentation, j'ai quelque peu modifié cet aspect en intégrant aux motifs les éléments typographiques qui faisaient la jointure entre deux objets.

Ainsi, dans l'exemple ci-dessous, le relateur est *Selon*, la source *lui* et le discours rapporté extit les Français ont [...] des maires de France. Les éléments typographiques **,** et **“** joignent la source et le discours alors que les éléments **”** et **.** clôturent le discours. Le motif extrait est donc : extit <relateur><source>, “ <discours> ”. :

> Selon lui, “ les Français ont [...] des maires de France ”.

Exemple extrait du corpus (©Le Figaro)

B.3.2 Schéma du relateur

Le relateur est un objet introduit par [Giguet & Lucas, 2004] qui permet de relier la source et ses propos au sein du texte englobant. Les relateurs peuvent être de différentes tailles et de différentes formes. Le but de la caractérisation était de se faire une idée des formes les plus employées.

Les conventions utilisées pour le schéma du relateur sont de spécifier les catégories linguistiques des mots ou expressions lorsque son utilisation n'est pas spécifique, ou bien le mot lui même – dans sa version lemmatisée – autrement.

Ainsi, dans l'exemple ci-dessous, le relateur est le verbe extit assure, cependant d'autres verbes auraient pû être employés : dire, penser,...On choisit donc de le représenter par "Verbe" :

> “ Il ne s'agit pas d'une augmentation déguisée de la taxe automobile ”, assure le ministre des Transports, Wolfgang Tiefensee,

Exemple extrait du corpus (©Libération)

Cependant, dans le cas ci-dessous, le relateur est extit a déclaré qu' correspondant au verbe extit déclarer suivi de extit que. Bien qu'on puisse remplacer le verbe extit déclarer par extit dire ou extit penser, aucun remplacement n'est évident pour extit que que l'on conserve donc tel quel dans le schéma : "Verbe+que".

> il a déclaré qu'il pourrait nommer un premier ministre de gauche, s'il était élu président de la République.
>
> *Exemple extrait du corpus (©Le Monde)*

B.4 Concordance des temps

La caractérisation des temps employés concerne tous les temps et modes qui peuvent être présents au sein du segment citationnel.

Cette caractérisation du corpus se limitant à l'étude du contexte phrastique, il est possible qu'aucun verbe du récit ne soit présent dans le segment citationnel considéré, ni qu'aucun verbe ne soit présent au sein des propos rapportés. Dans ces cas, il suffit d'appliquer la valeur extit NA.

B.4.1 Temps du récit

On considère comme verbes du récit, tous les verbes situés à l'extérieur des propos rapportés par la source, les éventuels verbes du relateur y compris.

B.4.2 Temps du discours rapporté

On considère comme verbe du discours rapporté tous les verbes présents au sein des propos rapportés par la source, y compris lorsqu'il s'agit du discours indirect.

Annexe C

Résultats qualitatifs de l'analyse du corpus

C.1 Styles du discours et formes des sources par article

C.1.1 "Corpus : Le Figaro"

"Article Fig01" : Discours

Nb citations	5	
Nb discours direct	2	(40%)
Nb discours indirect	2	(40%)
Nb discours indirect libre	0	(0%)
Nb discours mix	1	(20%)
Nb discours inconnu	0	(0%)

"Article Fig02" : Discours

Nb citations	4	
Nb discours direct	3	(75%)
Nb discours indirect	1	(25%)
Nb discours indirect libre	0	(0%)
Nb discours mix	0	(0%)
Nb discours inconnu	0	(0%)

"Article Fig03" : Discours

Nb citations	3	
Nb discours direct	0	(0%)
Nb discours indirect	1	(33%)
Nb discours indirect libre	1	(33%)
Nb discours mix	1	(33%)
Nb discours inconnu	0	(0%)

"Article Fig04" : Discours

Nb citations	5	
Nb discours direct	1	(20%)
Nb discours indirect	2	(40%)
Nb discours indirect libre	1	(20%)
Nb discours mix	1	(20%)
Nb discours inconnu	0	(0%)

"Article Fig05" : Discours

Nb citations	17	
Nb discours direct	4	(23%)
Nb discours indirect	0	(0%)
Nb discours indirect libre	1	(5%)
Nb discours mix	12	(70%)
Nb discours inconnu	0	(0%)

"Article Fig06" : Discours

Nb citations	6	
Nb discours direct	3	(50%)
Nb discours indirect	0	(0%)
Nb discours indirect libre	0	(0%)
Nb discours mix	3	(50%)
Nb discours inconnu	0	(0%)

"Article Fig07" : Discours

Nb citations	13	
Nb discours direct	9	(69%)
Nb discours indirect	0	(0%)
Nb discours indirect libre	0	(0%)
Nb discours mix	4	(30%)
Nb discours inconnu	0	(0%)

"Article Fig08" : Discours

Nb citations	14	
Nb discours direct	6	(42%)
Nb discours indirect	1	(7%)
Nb discours indirect libre	0	(0%)
Nb discours mix	7	(50%)
Nb discours inconnu	0	(0%)

"Article Fig09" : Discours

Nb citations	2	
Nb discours direct	2	(100%)
Nb discours indirect	0	(0%)
Nb discours indirect libre	0	(0%)
Nb discours mix	0	(0%)
Nb discours inconnu	0	(0%)

"Article Fig10" : Discours

Nb citations	11	
Nb discours direct	5	(45%)
Nb discours indirect	1	(9%)
Nb discours indirect libre	1	(9%)
Nb discours mix	4	(36%)
Nb discours inconnu	0	(0%)

"Article Fig01" : Sources

Nb citations	5	
Nb entites nommees	3	(60%)
Nb pronom	1	(20%)
Nb groupes nominaux	4	(80%)
Nb source inconnue	0	(0%)
Nb ?	0	(0%)

"Article Fig02" : Sources

Nb citations	4	
Nb entites nommees	4	(100%)
Nb pronom	0	(0%)
Nb groupes nominaux	2	(50%)
Nb source inconnue	0	(0%)
Nb ?	0	(0%)

"Article Fig03" : Sources

Nb citations	3	
Nb entites nommees	0	(0%)
Nb pronom	3	(100%)
Nb groupes nominaux	0	(0%)
Nb source inconnue	0	(0%)
Nb ?	0	(0%)

"Article Fig04" : Sources

Nb citations	5	
Nb entites nommees	0	(0%)
Nb pronom	0	(0%)
Nb groupes nominaux	5	(100%)
Nb source inconnue	0	(0%)
Nb ?	0	(0%)

"Article Fig05" : Sources

Nb citations	17	
Nb entites nommees	5	(29%)
Nb pronom	7	(41%)
Nb groupes nominaux	2	(11%)
Nb source inconnue	3	(17%)
Nb ?	0	(0%)

"Article Fig06" : Sources

Nb citations	6	
Nb entites nommees	2	(33%)
Nb pronom	1	(16%)
Nb groupes nominaux	2	(33%)
Nb source inconnue	1	(16%)
Nb ?	0	(0%)

"Article Fig07" : Sources

Nb citations	13	
Nb entites nommees	12	(92%)
Nb pronom	0	(0%)
Nb groupes nominaux	6	(46%)
Nb source inconnue	0	(0%)
Nb ?	0	(0%)

"Article Fig08" : Sources

Nb citations	14	
Nb entites nommees	0	(0%)
Nb pronom	11	(78%)
Nb groupes nominaux	2	(14%)
Nb source inconnue	1	(7%)
Nb ?	0	(0%)

"Article Fig09" : Sources

Nb citations	2	
Nb entites nommees	1	(50%)
Nb pronom	0	(0%)
Nb groupes nominaux	2	(100%)
Nb source inconnue	0	(0%)
Nb ?	0	(0%)

"Article Fig10" : Sources

Nb citations	11	
Nb entites nommees	3	(27%)
Nb pronom	0	(0%)
Nb groupes nominaux	6	(54%)
Nb source inconnue	2	(18%)
Nb ?	0	(0%)

C.1.2 "Corpus : Le Monde"

"Article Monde01" : Discours

Nb citations	2	
Nb discours direct	1	(50%)
Nb discours indirect	1	(50%)
Nb discours indirect libre	0	(0%)
Nb discours mix	0	(0%)
Nb discours inconnu	0	(0%)

"Article Monde02" : Discours

Nb citations	3	
Nb discours direct	1	(33%)
Nb discours indirect	1	(33%)
Nb discours indirect libre	0	(0%)
Nb discours mix	1	(33%)
Nb discours inconnu	0	(0%)

"Article Monde03" : Discours

Nb citations	7	
Nb discours direct	5	(71%)
Nb discours indirect	0	(0%)
Nb discours indirect libre	0	(0%)
Nb discours mix	2	(28%)
Nb discours inconnu	0	(0%)

"Article Monde04" : Discours

Nb citations	0	
Nb discours direct	0	(100%)
Nb discours indirect	0	(100%)
Nb discours indirect libre	0	(100%)
Nb discours mix	0	(100%)
Nb discours inconnu	0	(100%)

"Article Monde05" : Discours

Nb citations	5	
Nb discours direct	5	(100%)
Nb discours indirect	0	(0%)
Nb discours indirect libre	0	(0%)
Nb discours mix	0	(0%)
Nb discours inconnu	0	(0%)

"Article Monde06" : Discours

Nb citations	18	
Nb discours direct	7	(38%)
Nb discours indirect	5	(27%)
Nb discours indirect libre	2	(11%)
Nb discours mix	4	(22%)
Nb discours inconnu	0	(0%)

"Article Monde07" : Discours

Nb citations	11	
Nb discours direct	2	(18%)
Nb discours indirect	6	(54%)
Nb discours indirect libre	0	(0%)
Nb discours mix	2	(18%)
Nb discours inconnu	1	(9%)

"Article Monde08" : Discours

Nb citations	7	
Nb discours direct	2	(28%)
Nb discours indirect	4	(57%)
Nb discours indirect libre	1	(14%)
Nb discours mix	0	(0%)
Nb discours inconnu	0	(0%)

"Article Monde09" : Discours

Nb citations	1	
Nb discours direct	0	(0%)
Nb discours indirect	1	(100%)
Nb discours indirect libre	0	(0%)
Nb discours mix	0	(0%)
Nb discours inconnu	0	(0%)

"Article Monde10" : Discours

Nb citations	6	
Nb discours direct	2	(33%)
Nb discours indirect	2	(33%)
Nb discours indirect libre	0	(0%)
Nb discours mix	2	(33%)
Nb discours inconnu	0	(0%)

"Article Monde11" : Discours

Nb citations	5	
Nb discours direct	4	(80%)
Nb discours indirect	0	(0%)
Nb discours indirect libre	0	(0%)
Nb discours mix	1	(20%)
Nb discours inconnu	0	(0%)

"Article Monde12" : Discours

Nb citations	17	
Nb discours direct	13	(76%)
Nb discours indirect	1	(5%)
Nb discours indirect libre	0	(0%)
Nb discours mix	3	(17%)
Nb discours inconnu	0	(0%)

"Article Monde01" : Sources

Nb citations	2	
Nb entites nommees	2	(100%)
Nb pronom	0	(0%)
Nb groupes nominaux	2	(100%)
Nb source inconnue	0	(0%)
Nb ?	0	(0%)

"Article Monde02" : Sources

Nb citations	3	
Nb entites nommees	1	(33%)
Nb pronom	2	(66%)
Nb groupes nominaux	0	(0%)
Nb source inconnue	0	(0%)
Nb ?	0	(0%)

"Article Monde03" : Sources

Nb citations	7	
Nb entites nommees	7	(100%)
Nb pronom	0	(0%)
Nb groupes nominaux	3	(42%)
Nb source inconnue	0	(0%)
Nb ?	0	(0%)

"Article Monde04" : Sources

Nb citations	0	
Nb entites nommees	0	(100%)
Nb pronom	0	(100%)
Nb groupes nominaux	0	(100%)
Nb source inconnue	0	(100%)
Nb ?	0	(100%)

"Article Monde05" : Sources

Nb citations	5	
Nb entites nommees	2	(40%)
Nb pronom	0	(0%)
Nb groupes nominaux	1	(20%)
Nb source inconnue	3	(60%)
Nb ?	0	(0%)

"Article Monde06" : Sources

Nb citations	18	
Nb entites nommees	12	(66%)
Nb pronom	1	(5%)
Nb groupes nominaux	8	(44%)
Nb source inconnue	0	(0%)
Nb ?	0	(0%)

"Article Monde07" : Sources

Nb citations	11	
Nb entites nommees	9	(81%)
Nb pronom	0	(0%)
Nb groupes nominaux	7	(63%)
Nb source inconnue	0	(0%)
Nb ?	1	(9%)

"Article Monde08" : Sources

Nb citations	7	
Nb entites nommees	4	(57%)
Nb pronom	1	(14%)
Nb groupes nominaux	3	(42%)
Nb source inconnue	0	(0%)
Nb ?	0	(0%)

"Article Monde09" : Sources

Nb citations	1	
Nb entites nommees	1	(100%)
Nb pronom	0	(0%)
Nb groupes nominaux	1	(100%)
Nb source inconnue	0	(0%)
Nb ?	0	(0%)

"Article Monde10" : Sources

Nb citations	6	
Nb entites nommees	2	(33%)
Nb pronom	0	(0%)
Nb groupes nominaux	4	(66%)
Nb source inconnue	1	(16%)
Nb ?	0	(0%)

"Article Monde11" : Sources

Nb citations	5	
Nb entites nommees	3	(60%)
Nb pronom	1	(20%)
Nb groupes nominaux	3	(60%)
Nb source inconnue	0	(0%)
Nb ?	0	(0%)

"Article Monde12" : Sources

Nb citations	17	
Nb entites nommees	14	(82%)
Nb pronom	0	(0%)
Nb groupes nominaux	5	(29%)
Nb source inconnue	2	(11%)
Nb ?	0	(0%)

C.1.3 "Corpus : Challenges"

"Article Challenges01" : Discours

Nb citations	1	
Nb discours direct	0	(0%)
Nb discours indirect	0	(0%)
Nb discours indirect libre	0	(0%)
Nb discours mix	1	(100%)
Nb discours inconnu	0	(0%)

"Article Challenges02" : Discours

Nb citations	2	
Nb discours direct	1	(50%)
Nb discours indirect	0	(0%)
Nb discours indirect libre	1	(50%)
Nb discours mix	0	(0%)
Nb discours inconnu	0	(0%)

"Article Challenges03" : Discours

Nb citations	2	
Nb discours direct	2	(100%)
Nb discours indirect	0	(0%)
Nb discours indirect libre	0	(0%)
Nb discours mix	0	(0%)
Nb discours inconnu	0	(0%)

"Article Challenges04" : Discours

Nb citations	3	
Nb discours direct	1	(33%)
Nb discours indirect	2	(66%)
Nb discours indirect libre	0	(0%)
Nb discours mix	0	(0%)
Nb discours inconnu	0	(0%)

"Article Challenges05" : Discours

Nb citations	2	
Nb discours direct	2	(100%)
Nb discours indirect	0	(0%)
Nb discours indirect libre	0	(0%)
Nb discours mix	0	(0%)
Nb discours inconnu	0	(0%)

"Article Challenges06" : Discours

Nb citations	10	
Nb discours direct	10	(100%)
Nb discours indirect	0	(0%)
Nb discours indirect libre	0	(0%)
Nb discours mix	0	(0%)
Nb discours inconnu	0	(0%)

"Article Challenges07" : Discours

Nb citations	23	
Nb discours direct	10	(43%)
Nb discours indirect	1	(4%)
Nb discours indirect libre	2	(8%)
Nb discours mix	10	(43%)
Nb discours inconnu	0	(0%)

"Article Challenges08" : Discours

Nb citations	1	
Nb discours direct	0	(0%)
Nb discours indirect	0	(0%)
Nb discours indirect libre	0	(0%)
Nb discours mix	1	(100%)
Nb discours inconnu	0	(0%)

"Article Challenges09" : Discours

Nb citations	31	
Nb discours direct	30	(96%)
Nb discours indirect	0	(0%)
Nb discours indirect libre	0	(0%)
Nb discours mix	1	(3%)
Nb discours inconnu	0	(0%)

"Article Challenges10" : Discours

Nb citations	8	
Nb discours direct	7	(87%)
Nb discours indirect	1	(12%)
Nb discours indirect libre	0	(0%)
Nb discours mix	0	(0%)
Nb discours inconnu	0	(0%)

"Article Challenges11" : Discours

Nb citations	4	
Nb discours direct	3	(75%)
Nb discours indirect	1	(25%)
Nb discours indirect libre	0	(0%)
Nb discours mix	0	(0%)
Nb discours inconnu	0	(0%)

"Article Challenges01" : Sources

Nb citations	1	
Nb entites nommees	1	(100%)
Nb pronom	0	(0%)
Nb groupes nominaux	1	(100%)
Nb source inconnue	0	(0%)
Nb ?	0	(0%)

"Article Challenges02" : Sources

Nb citations	2	
Nb entites nommees	2	(100%)
Nb pronom	0	(0%)
Nb groupes nominaux	2	(100%)
Nb source inconnue	0	(0%)
Nb ?	0	(0%)

"Article Challenges03" : Sources

Nb citations	2	
Nb entites nommees	2	(100%)
Nb pronom	0	(0%)
Nb groupes nominaux	2	(100%)
Nb source inconnue	0	(0%)
Nb ?	0	(0%)

"Article Challenges04" : Sources

Nb citations	3	
Nb entites nommees	2	(66%)
Nb pronom	1	(33%)
Nb groupes nominaux	1	(33%)
Nb source inconnue	0	(0%)
Nb ?	0	(0%)

"Article Challenges05" : Sources

Nb citations	2	
Nb entites nommees	0	(0%)
Nb pronom	1	(50%)
Nb groupes nominaux	0	(0%)
Nb source inconnue	1	(50%)
Nb ?	0	(0%)

"Article Challenges06" : Sources

Nb citations	10	
Nb entites nommees	5	(50%)
Nb pronom	3	(30%)
Nb groupes nominaux	1	(10%)
Nb source inconnue	2	(20%)
Nb ?	0	(0%)

"Article Challenges07" : Sources

Nb citations	23	
Nb entites nommees	19	(82%)
Nb pronom	2	(8%)
Nb groupes nominaux	12	(52%)
Nb source inconnue	0	(0%)
Nb ?	0	(0%)

"Article Challenges08" : Sources

Nb citations	1	
Nb entites nommees	1	(100%)
Nb pronom	0	(0%)
Nb groupes nominaux	0	(0%)
Nb source inconnue	0	(0%)
Nb ?	0	(0%)

"Article Challenges09" : Sources

Nb citations	31	
Nb entites nommees	15	(48%)
Nb pronom	6	(19%)
Nb groupes nominaux	14	(45%)
Nb source inconnue	4	(12%)
Nb ?	0	(0%)

"Article Challenges10" : Sources

Nb citations	8	
Nb entites nommees	7	(87%)
Nb pronom	0	(0%)
Nb groupes nominaux	4	(50%)
Nb source inconnue	0	(0%)
Nb ?	0	(0%)

"Article Challenges11" : Sources

Nb citations	4	
Nb entites nommees	4	(100%)
Nb pronom	0	(0%)
Nb groupes nominaux	3	(75%)
Nb source inconnue	0	(0%)
Nb ?	0	(0%)

C.1.4 "Corpus : Le Soir"

"Article Soir01" : Discours

Nb citations	0	
Nb discours direct	0	(100%)
Nb discours indirect	0	(100%)
Nb discours indirect libre	0	(100%)
Nb discours mix	0	(100%)
Nb discours inconnu	0	(100%)

"Article Soir02" : Discours

Nb citations	2	
Nb discours direct	0	(0%)
Nb discours indirect	1	(50%)
Nb discours indirect libre	1	(50%)
Nb discours mix	0	(0%)
Nb discours inconnu	0	(0%)

"Article Soir03" : Discours

Nb citations	10	
Nb discours direct	0	(0%)
Nb discours indirect	10	(100%)
Nb discours indirect libre	0	(0%)
Nb discours mix	0	(0%)
Nb discours inconnu	0	(0%)

"Article Soir04" : Discours

Nb citations	2	
Nb discours direct	0	(0%)
Nb discours indirect	2	(100%)
Nb discours indirect libre	0	(0%)
Nb discours mix	0	(0%)
Nb discours inconnu	0	(0%)

"Article Soir05" : Discours

Nb citations	0	
Nb discours direct	0	(100%)
Nb discours indirect	0	(100%)
Nb discours indirect libre	0	(100%)
Nb discours mix	0	(100%)
Nb discours inconnu	0	(100%)

"Article Soir06" : Discours

Nb citations	1	
Nb discours direct	1	(100%)
Nb discours indirect	0	(0%)
Nb discours indirect libre	0	(0%)
Nb discours mix	0	(0%)
Nb discours inconnu	0	(0%)

"Article Soir07" : Discours

Nb citations	1	
Nb discours direct	1	(100%)
Nb discours indirect	0	(0%)
Nb discours indirect libre	0	(0%)
Nb discours mix	0	(0%)
Nb discours inconnu	0	(0%)

"Article Soir08" : Discours

Nb citations	7	
Nb discours direct	0	(0%)
Nb discours indirect	6	(85%)
Nb discours indirect libre	1	(14%)
Nb discours mix	0	(0%)
Nb discours inconnu	0	(0%)

"Article Soir09" : Discours

Nb citations	3	
Nb discours direct	2	(66%)
Nb discours indirect	0	(0%)
Nb discours indirect libre	1	(33%)
Nb discours mix	0	(0%)
Nb discours inconnu	0	(0%)

"Article Soir10" : Discours

Nb citations	17	
Nb discours direct	15	(88%)
Nb discours indirect	1	(5%)
Nb discours indirect libre	1	(5%)
Nb discours mix	0	(0%)
Nb discours inconnu	0	(0%)

"Article Soir01" : Sources

Nb citations	0	
Nb entites nommees	0	(100%)
Nb pronom	0	(100%)
Nb groupes nominaux	0	(100%)
Nb source inconnue	0	(100%)
Nb ?	0	(100%)

"Article Soir02" : Sources

Nb citations	2	
Nb entites nommees	1	(50%)
Nb pronom	0	(0%)
Nb groupes nominaux	2	(100%)
Nb source inconnue	0	(0%)
Nb ?	0	(0%)

"Article Soir03" : Sources

Nb citations	10	
Nb entites nommees	6	(60%)
Nb pronom	3	(30%)
Nb groupes nominaux	5	(50%)
Nb source inconnue	0	(0%)
Nb ?	0	(0%)

"Article Soir04" : Sources

Nb citations	2	
Nb entites nommees	0	(0%)
Nb pronom	0	(0%)
Nb groupes nominaux	2	(100%)
Nb source inconnue	0	(0%)
Nb ?	0	(0%)

"Article Soir05" : Sources

Nb citations	0	
Nb entites nommees	0	(100%)
Nb pronom	0	(100%)
Nb groupes nominaux	0	(100%)
Nb source inconnue	0	(100%)
Nb ?	0	(100%)

"Article Soir06" : Sources

Nb citations	1	
Nb entites nommees	1	(100%)
Nb pronom	0	(0%)
Nb groupes nominaux	1	(100%)
Nb source inconnue	0	(0%)
Nb ?	0	(0%)

"Article Soir07" : Sources

Nb citations	1	
Nb entites nommees	0	(0%)
Nb pronom	0	(0%)
Nb groupes nominaux	0	(0%)
Nb source inconnue	1	(100%)
Nb ?	0	(0%)

"Article Soir08" : Sources

Nb citations	7	
Nb entites nommees	3	(42%)
Nb pronom	3	(42%)
Nb groupes nominaux	2	(28%)
Nb source inconnue	0	(0%)
Nb ?	0	(0%)

"Article Soir09" : Sources

Nb citations	3	
Nb entites nommees	2	(66%)
Nb pronom	0	(0%)
Nb groupes nominaux	2	(66%)
Nb source inconnue	1	(33%)
Nb ?	0	(0%)

"Article Soir10" : Sources

Nb citations	17	
Nb entites nommees	12	(70%)
Nb pronom	2	(11%)
Nb groupes nominaux	3	(17%)
Nb source inconnue	3	(17%)
Nb ?	0	(0%)

C.1.5 "Corpus : Libération"

"Article Libe01" : Discours

Nb citations	10	
Nb discours direct	6	(60%)
Nb discours indirect	0	(0%)
Nb discours indirect libre	0	(0%)
Nb discours mix	3	(30%)
Nb discours inconnu	1	(10%)

"Article Libe02" : Discours

Nb citations	8	
Nb discours direct	4	(50%)
Nb discours indirect	0	(0%)
Nb discours indirect libre	0	(0%)
Nb discours mix	4	(50%)
Nb discours inconnu	0	(0%)

"Article Libe03" : Discours

Nb citations	13	
Nb discours direct	9	(69%)
Nb discours indirect	0	(0%)
Nb discours indirect libre	0	(0%)
Nb discours mix	2	(15%)
Nb discours inconnu	2	(15%)

"Article Libe04" : Discours

Nb citations	6	
Nb discours direct	5	(83%)
Nb discours indirect	1	(16%)
Nb discours indirect libre	0	(0%)
Nb discours mix	0	(0%)
Nb discours inconnu	0	(0%)

"Article Libe05" : Discours

Nb citations	6	
Nb discours direct	5	(83%)
Nb discours indirect	0	(0%)
Nb discours indirect libre	0	(0%)
Nb discours mix	1	(16%)
Nb discours inconnu	0	(0%)

"Article Libe06" : Discours

Nb citations	7	
Nb discours direct	7	(100%)
Nb discours indirect	0	(0%)
Nb discours indirect libre	0	(0%)
Nb discours mix	0	(0%)
Nb discours inconnu	0	(0%)

"Article Libe07" : Discours

Nb citations	2	
Nb discours direct	0	(0%)
Nb discours indirect	0	(0%)
Nb discours indirect libre	1	(50%)
Nb discours mix	1	(50%)
Nb discours inconnu	0	(0%)

"Article Libe08" : Discours

Nb citations	3	
Nb discours direct	2	(66%)
Nb discours indirect	0	(0%)
Nb discours indirect libre	0	(0%)
Nb discours mix	1	(33%)
Nb discours inconnu	0	(0%)

"Article Libe09" : Discours

Nb citations	19	
Nb discours direct	14	(73%)
Nb discours indirect	0	(0%)
Nb discours indirect libre	0	(0%)
Nb discours mix	4	(21%)
Nb discours inconnu	1	(5%)

"Article Libe10" : Discours

Nb citations	6	
Nb discours direct	4	(66%)
Nb discours indirect	1	(16%)
Nb discours indirect libre	0	(0%)
Nb discours mix	1	(16%)
Nb discours inconnu	0	(0%)

"Article Libe01" : Sources

Nb citations	10	
Nb entites nommees	2	(20%)
Nb pronom	2	(20%)
Nb groupes nominaux	4	(40%)
Nb source inconnue	1	(10%)
Nb ?	1	(10%)

"Article Libe02" : Sources

Nb citations	8	
Nb entites nommees	2	(25%)
Nb pronom	3	(37%)
Nb groupes nominaux	2	(25%)
Nb source inconnue	1	(12%)
Nb ?	0	(0%)

"Article Libe03" : Sources

Nb citations	13	
Nb entites nommees	7	(53%)
Nb pronom	0	(0%)
Nb groupes nominaux	5	(38%)
Nb source inconnue	2	(15%)
Nb ?	2	(15%)

"Article Libe04" : Sources

Nb citations	6	
Nb entites nommees	3	(50%)
Nb pronom	0	(0%)
Nb groupes nominaux	6	(100%)
Nb source inconnue	0	(0%)
Nb ?	0	(0%)

"Article Libe05" : Sources

Nb citations	6	
Nb entites nommees	4	(66%)
Nb pronom	0	(0%)
Nb groupes nominaux	2	(33%)
Nb source inconnue	1	(16%)
Nb ?	0	(0%)

"Article Libe06" : Sources

Nb citations	7	
Nb entites nommees	6	(85%)
Nb pronom	1	(14%)
Nb groupes nominaux	5	(71%)
Nb source inconnue	0	(0%)
Nb ?	0	(0%)

"Article Libe07" : Sources

Nb citations	2	
Nb entites nommees	2	(100%)
Nb pronom	0	(0%)
Nb groupes nominaux	2	(100%)
Nb source inconnue	0	(0%)
Nb ?	0	(0%)

"Article Libe08" : Sources

Nb citations	3	
Nb entites nommees	2	(66%)
Nb pronom	1	(33%)
Nb groupes nominaux	0	(0%)
Nb source inconnue	0	(0%)
Nb ?	0	(0%)

"Article Libe09" : Sources

Nb citations	19	
Nb entites nommees	6	(31%)
Nb pronom	3	(15%)
Nb groupes nominaux	5	(26%)
Nb source inconnue	5	(26%)
Nb ?	1	(5%)

"Article Libe10" : Sources

Nb citations	6	
Nb entites nommees	4	(66%)
Nb pronom	0	(0%)
Nb groupes nominaux	2	(33%)
Nb source inconnue	1	(16%)
Nb ?	0	(0%)

C.2 Styles du discours et formes des sources par journaux

C.2.1 Stats "Corpus : Le Figaro"

Nb citations	80	
Nb discours direct	35	(43%)
Nb discours indirect	8	(10%)
Nb discours indirect libre	4	(5%)
Nb discours mix	33	(41%)
Nb discours inconnu	0	(0%)
Nb entites nommees	30	(37%)
Nb pronom	23	(28%)
Nb groupes nominaux	31	(38%)
Nb source inconnue	7	(8%)
Nb ?	0	(0%)

C.2.2 Stats "Corpus : Le Monde"

Nb citations	82	
Nb discours direct	42	(51%)
Nb discours indirect	21	(25%)
Nb discours indirect libre	3	(3%)
Nb discours mix	15	(18%)
Nb discours inconnu	1	(1%)
Nb entites nommees	57	(69%)
Nb pronom	5	(6%)
Nb groupes nominaux	37	(45%)
Nb source inconnue	6	(7%)
Nb ?	1	(1%)

C.2.3 Stats "Corpus : Challenges"

Nb citations	87	
Nb discours direct	66	(75%)
Nb discours indirect	5	(5%)
Nb discours indirect libre	3	(3%)
Nb discours mix	13	(14%)
Nb discours inconnu	0	(0%)
Nb entites nommees	58	(66%)
Nb pronom	13	(14%)
Nb groupes nominaux	40	(45%)
Nb source inconnue	7	(8%)
Nb ?	0	(0%)

C.2.4 Stats "Corpus : Le Soir"

Nb citations	43	
Nb discours direct	19	(44%)
Nb discours indirect	20	(46%)
Nb discours indirect libre	4	(9%)
Nb discours mix	0	(0%)
Nb discours inconnu	0	(0%)
Nb entites nommees	25	(58%)
Nb pronom	8	(18%)
Nb groupes nominaux	17	(39%)
Nb source inconnue	5	(11%)
Nb ?	0	(0%)

C.2.5 Stats "Corpus : Libération"

Nb citations	80	
Nb discours direct	56	(70%)
Nb discours indirect	2	(2%)
Nb discours indirect libre	1	(1%)
Nb discours mix	17	(21%)
Nb discours inconnu	4	(5%)
Nb entites nommees	38	(47%)
Nb pronom	10	(12%)
Nb groupes nominaux	33	(41%)
Nb source inconnue	11	(13%)
Nb ?	4	(5%)

C.3 Styles du discours et formes des sources : distribution au sein du corpus

C.3.1 Stats Globales

Nb citations	744	
Nb discours direct	436	(58%)
Nb discours indirect	112	(15%)
Nb discours indirect libre	30	(4%)
Nb discours mix	156	(20%)
Nb discours inconnu	10	(1%)
Nb entites nommees	416	(55%)
Nb pronom	118	(15%)
Nb groupes nominaux	316	(42%)
Nb source inconnue	72	(9%)
Nb ?	10	(1%)

C.4 Motifs repérés au sein du corpus selon la formalisation de Giguet et Lucas

<discours><relateur><discours>	1	ϵ%
<discours><relateur><source1><relateur><source2>	1	ϵ%
<discours><relateur><source1><source2>	1	ϵ%
<discours><relateur><source><relateur>	1	ϵ%
<discours><relateur><source><source><relateur><discours>	1	ϵ%
<discours><source>	1	ϵ%
<discours><source><relateur><discours>	1	ϵ%
<relateur><discours><relateur><source>	1	ϵ%
<relateur><source><relateur><discours>	1	ϵ%
<relateur><sourcer><discours>	1	ϵ%
<source1><relateur><source2><relateur><discours>	1	ϵ%
<introducteur><discours>	3	env. 1%
<source><discours>	3	env. 1%
<discours><relateur><source><relateur><discours>	6	1.6%
<relateur><discours>	6	1.6%
<discours><relateur><source><discours>	22	6%
<relateur><source><discours>	31	8.4%
<discours>	33	8.9%
<source><relateur><discours>	104	28.3%
<discours><relateur><source>	148	40.2%

C.5 Formes de relateurs présents dans le corpus

<introducteur>+car	1
Assurantque	1
Auxiliaire	1
Auxiliaire+,subordonnée,+indicetemporel+participe+que	1
Auxiliaire+Adverbe+Participe+que	1
Auxiliaire+ainsi+Participe+indicegéographique	1
Auxiliaire+ainsi+Participe+indicetemporel	1
Auxiliaire+déjà+participe	1
Auxiliaire+enrevanche+Participe	1
Auxiliaire+également+Participe+que	1
Auxiliaire+également+ParticipePassé+indicetemporel	1
Comme	1
Commele+Verbe	1
D'yallerdesoncoupletsurlefaitque	1
Eneffet	1
Groupenominal+de	1
Groupenominal+verbepronominal+de	1
Imparfait	1
Indicetemporel	1
Indicetemporel+,+Verbe+COD	1
Introducteur	1
Objectif	1
Parceque+Verbe	1
Participe+par	1
Pronom+Verbe+que	1
Vajusqu'à+Verbe+que	1
Vamêmeplusloin	1
Verb+descriptiondiscours	1
Verbe+Adverbe+COD+,+où	1
Verbe+COD+COI	1
Verbe+CODsubordonnée+ParticipePrésent+surlefaitque	1
Verbe+COI	1
Verbe+adverbe	1
Verbe+ainsi+que	1
Verbe+aussi+sur	1
Verbe+avecfranchise	1
Verbe+comme	1
Verbe+coordonnée	1
Verbe+d'autrepart+que	1
Verbe+de+groupenominal	1
Verbe+descriptiondiscours	1
Verbe+désormais	1
Verbe+désormais+de	1
Verbe+encore	1
Verbe+encore+adjectif	1
Verbe+eneffet+que	1
Verbe+engarde	1
Verbe+ensuite	1
Verbe+indicetemporel+indicelocalisation	1
Verbe+indicetemporel+indicelocalisation+que	1
Verbe+indicetemporel+que	1
Verbe+indicetemporel+surbordonnée	1

Annexe D

Application des schémas XML

D.1 Exemple de structuration d'un article du corpus

```
<?xml version="1.0" encoding="utf-8"?>
<!DOCTYPE article SYSTEM "DTDCorpus.dtd">
<corpus:article
xmlns:corpus="http://www.fabienpoulard.info/xmlns/corpus"
xmlns:typo="http://www.fabienpoulard.info/xmlns/typo">
<corpus:metadata>
<corpus:journal>Le Figaro</corpus:journal>
<corpus:url>
http://www.lefigaro.fr/election-presidentielle-2007/20070220.FIG000000204_le_pen
</corpus:url>
<corpus:authoring>
<corpus:author>Olivier Pognon</corpus:author>
</corpus:authoring>
<corpus:publicationdate>20.02.2007</corpus:publicationdate>
<corpus:revisiondate>20.02.2007</corpus:revisiondate>
</corpus:metadata>
<corpus:content>
<corpus:title>
Le Pen lance un "appel solennel" aux maires de France
</corpus:title>
<corpus:epigraph>
Le président du FN déclare avoir à ce jour " entre 460 et
500 parrainages ".
</corpus:epigraph>
<corpus:bloc>
<corpus:paragraph>
DRAMATISATION calculée, ou<typo:i> " inquiétude "</typo:i>
réelle et fondée, comme il a dit l'éprouver hier ? Peut-être
un peu les deux. Jean-Marie Le Pen, comme souvent, excelle à
utiliser à son avantage une situation périlleuse. Le fait est
```

que celle-ci ressemble de plus en plus à celle de 2002, quand il n'avait disposé des 500 parrainages requis qu'in extremis.
</corpus:paragraph>
<corpus:paragraph>
Cette fois encore, à moins d'un mois du 16 mars, date limite fixée pour le dépôt des parrainages au Conseil constitutionnel, il n'a pas un nombre de promesses de signatures suffisant pour être sûr de pouvoir figurer dans la compétition présidentielle. Au point qu'il a lancé hier un nouvel <typo:i>" appel solennel "
</typo:i> aux élus pour qu'ils lui donnent " <typo:i>dans les meilleurs délais "</typo:i> les parrainages qui lui
<typo:i>" manquent "</typo:i>. Il a actuellement, a-t-il dit,
<typo:i>" entre 460 et 500 "</typo:i> promesses, alors qu'il s'était fixé un objectif de 600 pour tenir compte des désistements. D'où son <typo:i>" inquiétude "</typo:i>, tout de même atténuée par le fait qu'il a, assure-t-il,
" <typo:i>pour règle de vie de se battre, de persévérer et de ne jamais s'avouer vaincu ".</typo:i> Il a d'ailleurs reconnu que <typo:i>" beaucoup de maires attendent les formulaires pour donner leur signature "</typo:i>, ce qui peut lui laisser espérer une nouvelle arrivée de parrainages.
</corpus:paragraph>
<corpus:paragraph>
Jean-Marie Le Pen a fait, avec cette <typo:i>" déclaration solennelle ",</typo:i> d'une pierre deux coups : en même temps qu'il appelait les maires à lui permettre d'être candidat, il présentait sa participation à l'élection comme une nécessité démocratique et nationale. " <typo:i>Le pays est dans une situation difficile qui exige une campagne électorale qui informe, par un large débat démocratique, les électeurs qui vont décider de leur destin "</typo:i>, a-t-il dit dans sa déclaration, lue au cours d'une conférence de presse.
</corpus:paragraph>
<corpus:paragraph>
Et d'ajouter :<typo:i> " Tout le monde ou presque s'accorde à dire qu'il serait scandaleux et dommageable pour la démocratie et la République que je ne puisse pas, par défaut de parrainages, être candidat. "</typo:i> Selon lui,
<typo:i>" les Français ont droit à un vrai et grand débat démocratique. La clé est entre les mains des maires de France </typo:i>". <typo:i>" Ce sont eux qui, par la loi, sont en charge de l'accomplissement de cette formalité. Ce n'est pas une simple option, c'est une obligation ",
</typo:i> a souligné le président du FN, en, rappelant qu'il n'était pas <typo:i>" un candidat fantaisiste ",

```
</typo:i>ayant obtenu 14 % des voix en 1988, 15 % en 1995 et
17 % en 2002.
</corpus:paragraph>
</corpus:bloc>
<corpus:bloc>
<corpus:header>
<typo:b>"Vice fondamental"</typo:b>
</corpus:header>
<corpus:paragraph>
Le président du FN a une nouvelle fois pointé du doigt le
fait que les noms des parrains des candidats - plus exactement
les noms de 500 d'entre eux tirés au sort par candidat - soit
rendus publics. Il s'agit là, selon lui, du <typo:i>" vice
fondamental de cette institution qui est contraire à l'article
3 de notre Constitution, qui prévoit que le scrutin est
toujours secret "</typo:i>. C'est cette publicité qui, selon
lui, dissuade les maires de parrainer, " <typo:i>les élus
indépendants, isolés "</typo:i> craignant <typo:i>" à tort
ou à raison d'être victimes de représailles "</typo:i> et
<typo:i>" l'étant parfois ".</typo:i> Il a assuré que
<typo:i>" beaucoup de maires se concertent pour refuser tout
parrainage à quiconque, pratiquant ainsi une abstention que
pourtant ils déplorent quand elle est le fait des électeurs ".
</typo:i>
</corpus:paragraph>
</corpus:bloc>
</corpus:content>
</corpus:article>
```

D.2 Article annoté avec la première méthode

```
<?xml version="1.0" encoding="utf-8"?>
<?xml-stylesheet href="annotation.css" type="text/css" media="screen"?>
<!DOCTYPE article SYSTEM "DTDCorpus.dtd">
<corpus:article
 xmlns:corpus="http://www.fabienpoulard.info/xmlns/corpus"
 xmlns:typo="http://www.fabienpoulard.info/xmlns/typo"
 xmlns:ht="http://www.fabienpoulard.info/xmlns/ht"
 xmlns:cite="http://www.fabienpoulard.info/xmlns/cite">
 <corpus:metadata>
  <corpus:journal>Challenges</corpus:journal>
  <corpus:url>
   http://www.challenges.fr/business/art_95809.html
  </corpus:url>
```

```
   <corpus:authoring></corpus:authoring>
   <corpus:edition>Challenges.fr</corpus:edition>
   <corpus:publicationdate>21.02.2007</corpus:publicationdate>
  </corpus:metadata>
  <corpus:content>
   <corpus:title>
    <cite:discours source="1">Le programme de Royal co\^uterait
    35 milliards</cite:discours>, selon <cite:source id="1">
    le PS</cite:source>
   </corpus:title>
   <corpus:bloc>
    <corpus:paragraph>
     Un document officiel du Parti socialiste, rendu public
     mercredi 21 f\'evrier, \'evalue \'a 35 milliards le
    "co\^ut net" du Pacte pr\'esidentielle de S\'egol\`ene
    Royal. Puisque les d\'epenses repr\'esenteraient 50
    milliards d'euros, compens\'ees en partie par 15
    milliards d'euros d'\'economies et de red\'eploiement.
    </corpus:paragraph>
    <corpus:paragraph>
    <cite:discours source="3">"Il s'agit d'un exercice de
    v\'erit\'e", de "coh\'erence, de volont\'e et de
    responsabilit\'e"</cite:discours>, a assur\'e
    <cite:source id="3">le Premier secr\'etaire du PS,
    Fran\c{c}ois Hollande</cite:source>, lors de la
    pr\'esentation de ces chiffres, en pleine pol\'emique
    sur le co\^ut des programmes des candidats \'a la
    pr\'esidentielle.
    </corpus:paragraph>
    <corpus:paragraph>
    Objectif pour le PS: r\'eduire la dette \'a 60% en
    2012, stabiliser les pr\'el\`evements obligatoires au
    niveau de 2006 et limiter \'a 1,8% annuels la hausse
    de la d\'epense publique. Le tout sur fond d'une
    croissance de 2,5% par an d\`es 2008.
    </corpus:paragraph>
    <corpus:paragraph>
    De son c\^ot\'e, l'UMP chiffre \'a 31,7 milliards
    d'euros sur 5 ans les mesures propos\'ees par
    Nicolas Sarkozy.
    </corpus:paragraph>
   </corpus:bloc>
   <corpus:bloc>
    <corpus:header>
    <typo:b>Ou 63 milliards et 50 pour Sarkozy?</typo:b>
    </corpus:header>
```

```
␣␣␣␣<corpus:paragraph>
␣␣␣␣␣Pourtant␣,␣l'Institut␣de␣l'entreprise␣avan\c{c}ait␣,
␣␣␣␣␣mardi␣soir␣,␣les␣chiffres␣respectifs␣,␣pour␣les␣100
␣␣␣␣␣propositions␣de␣S\'egol\`ene␣Royal␣et␣le␣programme
␣␣␣␣␣de␣Nicolas␣Sarkozy␣,␣de␣63␣milliards␣d'euros␣et␣entre
␣␣␣␣␣49␣et␣51␣milliards.
␣␣␣␣</corpus:paragraph>
␣␣␣␣<corpus:paragraph>
␣␣␣␣␣Selon␣les␣calculs␣de␣sa␣cellule␣de␣chiffrage␣D\'ebat2007␣,
␣␣␣␣␣les␣mesures␣de␣la␣candidate␣socialiste␣\'a␣la
␣␣␣␣␣pr\'esidentielle␣les␣plus␣co\^uteuses␣sont:␣le␣doublement
␣␣␣␣␣du␣budget␣de␣la␣justice␣(6,2␣milliards)␣,␣le␣revenu␣de
␣␣␣␣␣solidarit\'e␣active␣(6␣milliards)␣,␣l'augmentation␣du
␣␣␣␣␣budget␣de␣la␣recherche␣(7␣milliards)␣,␣l'allocation
␣␣␣␣␣d'autonomie␣pour␣les␣jeunes␣(7␣milliards)␣,␣les
␣␣␣␣␣emplois-tremplins␣(6␣milliards).
␣␣␣␣</corpus:paragraph>
␣␣␣</corpus:bloc>
␣␣␣<corpus:bloc>
␣␣␣␣<corpus:header>
␣␣␣␣␣<typo:b>Exon\'erer␣de␣droits␣de␣succession␣co\^uterait
␣␣␣␣␣5␣milliards</typo:b>
␣␣␣␣</corpus:header>
␣␣␣␣<corpus:paragraph>
␣␣␣␣␣Quant␣au␣pr\'esident␣de␣l'UMP,␣son␣projet␣d'augmenter
␣␣␣␣␣le␣budget␣de␣l'enseignement␣sup\'erieur␣de␣50%␣et␣de
␣␣␣␣␣porter␣l'effort␣de␣recherche␣\'a␣3%␣du␣PIB␣en␣cinq␣ans
␣␣␣␣␣co\^uterait␣10,2␣milliards.␣L'exon\'eration␣de␣charges
␣␣␣␣␣et␣d'imp\^ots␣sur␣les␣heures␣suppl\'ementaires␣vaudrait
␣␣␣␣␣4,6␣milliards.␣Et␣l'exon\'eration␣des␣droits␣de
␣␣␣␣␣succession␣pour␣95%␣des␣Fran\c{c}ais␣capitaliserait␣5
␣␣␣␣␣milliards.␣A␣cela␣s'ajoute␣le␣d\'eveloppement␣du␣sport
␣␣␣␣␣\'a␣l'\'ecole␣(4␣milliards).
␣␣␣␣</corpus:paragraph>
␣␣␣</corpus:bloc>
␣␣</corpus:content>
␣</corpus:article>
```

D.3 Extrait d'un fichier intermédiaire illustrant le modèle XML

```
<?xml version="1.0" encoding="utf-8"?>
<?xml-stylesheet type="text/css" href="annotation.css"?>
<corpus:article xmlns:corpus="http://www.fabienpoulard.info/xmlns/corpus">
 <corpus:metadata>
```

```
  <corpus:journal>
   Le Soir
  </corpus:journal>
  <corpus:url>
   http://www.lesoir.be/culture/cinema/2007/02/21/article_les_des_sont_jetes_pou
  </corpus:url>
  <corpus:authoring>
   <corpus:author>
    AFP
   </corpus:author>
  </corpus:authoring>
  <corpus:edition>
   Le Soir en ligne
  </corpus:edition>
  <corpus:publicationdate>
   21.02.2007
  </corpus:publicationdate>
 </corpus:metadata>
 <corpus:content>
  <citation:sentence anchor="start" xmlns:citation="http://www.fabienpoulard.inf
  <corpus:word id="1"> </corpus:word>
  <corpus:word id="2"> </corpus:word>
  <corpus:word id="3"> </corpus:word>
  <corpus:title>
   <corpus:word id="4" lemme="le" pos="DET" pos-info="ART">Les</corpus:word>
   <corpus:word id="5"> </corpus:word>
   <corpus:word id="6" lemme="&lt;unknown&gt;" pos="NOM">d\'es</corpus:word>
␣␣␣<corpus:word␣id="7">␣</corpus:word>
␣␣␣<corpus:word␣id="8"␣lemme="\^etre"␣pos="VER"␣pos-info="pres">sont</corpus:wor
␣␣␣<corpus:word␣id="9">␣</corpus:word>
␣␣␣<corpus:word␣id="10"␣lemme="&lt;unknown&gt;"␣pos="ADJ">jet\'es</corpus:word>
   <corpus:word id="11"> </corpus:word>
   <marks:syntagmeprepos xmlns:marks="http://www.fabienpoulard.info/xmlns/marks"
    <corpus:word id="12" lemme="pour" pos="PRP">pour</corpus:word>
   </marks:syntagmeprepos>
   <corpus:word id="13"> </corpus:word>
   <corpus:word id="14" lemme="le" pos="DET" pos-info="ART">les</corpus:word>
   <corpus:word id="15"> </corpus:word>
   <corpus:word id="16" lemme="oscar" pos="NOM">Oscars</corpus:word>
  </corpus:title>
  <corpus:word id="17"> </corpus:word>
  <corpus:word id="18"> </corpus:word>
  <corpus:word id="19"> </corpus:word>
  <corpus:epigraph>
   <corpus:word id="20" lemme="le" pos="DET" pos-info="ART">Les</corpus:word>
   <corpus:word id="21"> </corpus:word>
```

```
    <corpus:word id="22" lemme="vote" pos="NOM">votes</corpus:word>
    <corpus:word id="23"> </corpus:word>
    <marks:syntagmeprepos xmlns:marks="http://www.fabienpoulard.info/xmlns/marks"
    <corpus:word id="24" lemme="pour" pos="PRP">pour</corpus:word>
    </marks:syntagmeprepos>
    <corpus:word id="25"> </corpus:word>
    <corpus:word id="26" lemme="&lt;unknown&gt;" pos="VER" pos-info="infi">d\'esi
␣␣␣<corpus:word␣id="27">␣</corpus:word>
␣␣␣<corpus:word␣id="28"␣lemme="le"␣pos="DET"␣pos-info="ART">les</corpus:word>
␣␣␣<corpus:word␣id="29">␣</corpus:word>
␣␣␣<corpus:word␣id="30"␣lemme="vainqueur"␣pos="NOM">vainqueurs</corpus:word>
␣␣␣<corpus:word␣id="31">␣</corpus:word>
␣␣␣<corpus:word␣id="32"␣lemme="du"␣pos="PRP"␣pos-info="det">des</corpus:word>
␣␣␣<corpus:word␣id="33">␣</corpus:word>
␣␣␣<corpus:word␣id="34"␣lemme="&lt;unknown&gt;"␣pos="NAM">Oscars</corpus:word>
␣␣␣<corpus:word␣id="35">␣</corpus:word>
␣␣␣<corpus:word␣id="36"␣lemme="avoir"␣pos="VER"␣pos-info="pres">ont</corpus:word
␣␣␣<corpus:word␣id="37">␣</corpus:word>
␣␣␣<corpus:word␣id="38"␣lemme="prendre"␣pos="VER"␣pos-info="pper">pris</corpus:w
␣␣␣<corpus:word␣id="39">␣</corpus:word>
␣␣␣<corpus:word␣id="40"␣lemme="fin"␣pos="NOM">fin</corpus:word>
␣␣␣<corpus:word␣id="41">␣</corpus:word>
␣␣␣<corpus:word␣id="42"␣lemme="ce"␣pos="PRO"␣pos-info="DEM">cette</corpus:word>
␣␣␣<corpus:word␣id="43">␣</corpus:word>
␣␣␣<corpus:word␣id="44"␣lemme="nuit"␣pos="NOM">nuit</corpus:word>
␣␣␣<corpus:word␣id="45"␣lemme=","␣pos="PUN">,</corpus:word>
␣␣␣<corpus:word␣id="46">␣</corpus:word>
␣␣␣<corpus:word␣id="47"␣lemme="cinq"␣pos="NUM">cinq</corpus:word>
␣␣␣<corpus:word␣id="48">␣</corpus:word>
␣␣␣<corpus:word␣id="49"␣lemme="jour"␣pos="NOM">jours</corpus:word>
␣␣␣<corpus:word␣id="50">␣</corpus:word>
␣␣␣<corpus:word␣id="51"␣lemme="avant"␣pos="PRP">avant</corpus:word>
␣␣␣<corpus:word␣id="52">␣</corpus:word>
␣␣␣<corpus:word␣id="53"␣lemme="le"␣pos="DET"␣pos-info="ART">la</corpus:word>
␣␣␣<corpus:word␣id="54">␣</corpus:word>
␣␣␣<corpus:word␣id="55"␣lemme="prestigieux"␣pos="ADJ">prestigieuse</corpus:word>
␣␣␣<corpus:word␣id="56">␣</corpus:word>
␣␣␣<corpus:word␣id="57"␣lemme="&lt;unknown&gt;"␣pos="NOM">c\'er\'emonie</corpus:v
␣␣␣<corpus:word␣id="58">␣</corpus:word>
␣␣␣<corpus:word␣id="59"␣lemme="hollywoodien"␣pos="ADJ">hollywoodienne</corpus:wo
␣␣␣<corpus:word␣id="60"␣lemme="."␣pos="SENT">.</corpus:word>
␣␣<citation:sentence␣anchor="stop"␣xmlns:citation="http://www.fabienpoulard.info
␣␣</corpus:epigraph>
```

Annexe E

Liste des verbes d'énonciation considérés comme indices

abjurer	aboyer	abréger	accepter
acclamer	accorder	accuser	acquiescer
admettre	affirmer	affliger	ajouter
alerter	amorcer	amplifier	analyser
annoncer	apostropher	appeler	arguer
argumenter	articuler	assurer	avertir
aviser	avouer	bafouiller	balbutier
baragouiner	bavarder	bégayer	bêler
beugleer	blaguer	bourdonner	bredouiller
cafarder	cafouiller	calligraphier	causer
chanter	chuchoter	citer	clamer
commenter	communiquer	compléter	complimenter
concéder	conclure	confesser	confier
confirmer	conjurer	conseiller	constater
conter	contester	contredire	correspondre
crier	critiquer	croasser	débiter
déclamer	déclarer	décréter	décrier
décrire	déduire	définir	demander
démentir	démontrer	dénoncer	déplorer
déposer	détonner	développer	deviner
dévoiler	dialoguer	dicter	diffamer
diffuser	dire	discuter	divulguer
se documenter	ébruiter	s'écrier	écrire
éditer	s'égosiller	engueuler	énoncer
s'enthousiasmer	entonner	épeler	épiloguer
s'époumoner	espérer	s'étonner	exalter
expliciter	expliquer	exprimer	extérioriser
extorquer	faire entendre	féliciter	formuler
fulminer	garantir	geindre	gémir
glapir	glorifier	gratifier	gribouiller

griffonner grogner grommeler gronder
hennir hurler illustrer imiter
indiquer inférer infirmer informer
injurier inscrire insinuer insister
interpeller interpréter interroger invoquer
ironiser jacasser jargonner jurer
justifier lire manifester marmonner
médire mentionner mentir miauler
moucharder mugir murmurer narrer
nier notifier nuancer octosyllaber
ordonner parler penser philosopher
plaider poétiser polémiquer ponctuer
préconiser prédire présager prévenir
proclamer proférer professer promettre
prononcer pronostiquer protester publier
questionner raconter rappeler rassurer
réaffirmer récapituler réciter recopier
rédiger redire réfuter relater
répéter répliquer répondre réprimander
reprocher résumer rétorquer retransmettre
révéler revendiquer rouspéter rugir
s'écrier s'interroger semoncer sermonner
signaler signifier suggérer supposer
tergiverser traduire verbaliser vilipender
vociférer zozoter

Bibliographie

[let,] Étude et enseignement du français.

[Académie Française, 1992] ACADÉMIE FRANÇAISE (1992). *Dictionnaire de l'Académie Française*. éditions Fayard, neuvième edition edition.

[Charolles, 2000] CHAROLLES M. (2000). Les expressions introductrices de cadres de discours et leur portée textuelle. séminaire de recherche.

[Fourour, 2002] FOUROUR N. (2002). Nemesis, un système de reconnaissance incrémentielle des entités nommées pour le français.

[Fourour, 2004] FOUROUR N. (2004). *Identification et catégorisation automatique des entités nommées dans les textes français*. PhD thesis, Université de Nantes.

[Giguet & Lucas, 2004] GIGUET E. & LUCAS N. (2004). *La détection automatique des citations et des locuteurs dans les textes informatifs*, In *Le discours rapporté dans tous ses états : Question de frontières*, p. 410–418. l'Harmattan.

[Habert, 2001] HABERT B. (2001). Des corpus représentatifs : de quoi, pour quoi, comment ?

[Komur, 2001] KOMUR G. (2001). L'ilot textuel et la prise de distance par le locuteur dans le genre journalistique. In *Le discours rapporté dans tous ses états : question de frontiéres* : L'Harmattan.

[Marshman, 2003] MARSHMAN E. (2003). Construction et gestion des corpus : Résumé et essai d'uniformisation du processus pour la terminologie.

[Mikheev, 2003] MIKHEEV A. (2003). *Text Segmentation*, In *The Oxford Handbook of Computational Linguistic*, chapter 10.

[Mourad, 2001] MOURAD G. (2001). *Analyse informatique des signes typographiques pour la segmentation de textes et l'extraction automatique des citations*. PhD thesis, Paris IV Sorbonne.

[Mourad & Desclès, 2001] MOURAD G. & DESCLÈS J.-P. (2001). Identification et extraction automatique des informations citationnelles dans un texte. In *Ci-Dit. Colloque internationnal et interdisciplinaire*.

[Mourad & Desclès, 2002] MOURAD G. & DESCLÈS J.-P. (2002). Citation textuelle : identification automatique par exploration contextuelle. *Faits de Langues*, (19).

[Mourad & Minel, 2000] MOURAD G. & MINEL J.-L. (2000). Filtrage sémantique du texte, le cas de la citation. p. 41–56. 3e Colloque international sur le Document Electronique.

[Platon, 360 av J C] PLATON (360 av. J.-C.). *Livre III, La République*.

[Rosier *et al.*,] ROSIER L., MARNETTE S. & MUNOZ J. M. L.

[S. Teufel & Tidhar, 2006] S. TEUFEL A. S. & TIDHAR D. (2006). An annotation scheme for citation function. In *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*, p. 80–87.

[Witten & Frank, 2005] WITTEN I. H. & FRANK E. (2005). *Data Mining : Practical machine learning tools and techniques*. Morgan Kaufmann, 2nd edition edition.

Sommaire

INSTITUT DE RECHERCHE EN INFORMATIQUE DE NANTES

Repérage automatique de citations dans des documents journalistiques

Mémoire de Master 2 Recherche SAD

Fabien POULARD
(encadré par Nicolas Hernandez et Annie Tartier)

Résumé

Détection de citations dans un domaine journalistique.

Termes généraux : traitement automatique du langage, repérage automatique, citations, domaine journalistique

Mots-clés additionnels et phrases : recherche d'information intra-documentaire,repérage et extraction d'information, analyse linguistique, apprentissage automatique, citation

LINA, Faculté des Sciences & Techniques de Nantes
2, rue de la Houssinière
B.P. 92208 — F-44322 NANTES CEDEX 3