Analyse quantitative et qualitative de citations extraites d'un corpus journalistique

Fabien Poulard

RECITAL 2008, Avignon, 9–13 juin 2008

Analyse quantitative et qualitative de citations extraites d'un corpus journalistique

Fabien Poulard
Université de Nantes, LINA CNRS UMR 6241, 44322 Nantes Cedex 3
Fabien.Poulard@univ-nantes.fr

Résumé. Dans le contexte de la détection de plagiats, le repérage de citations et de ses constituants est primordial puisqu'il peut aider à évaluer le caractère licite ou illicite d'une reprise (source citée ou non). Nous proposons ici une étude quantitative et qualitative des citations extraites d'un corpus que nous avons auparavant construit. Cette étude a pour but de tracer des axes de recherche vers une méthode de repérage automatique des citations.

Abstract. In the plagiarism detection context, finding citations and their components is essential as it may help estimating legal value of a copy (with or without original source specified). We propose here a quantitative and qualitative study of citations we extracted from a corpus we previously built. This study aims at orienting our research towards an efficient automatic citations extraction method.

Mots-clés : citations, contruction et étude de corpus, genre journalistique.

Keywords: citations, corpus creation and analysis, journalistic genre.

1 Introduction

Le problème d'identification des paroles et leurs porteurs a été étudié sous différents axes. En anglais notamment, (Teufel *et al.*, 2006) se sont intéressés aux citations scientifiques alors que (Prasad *et al.*, 2007; Kim & Hovy, 2006) ont travaillé sur les opinions dans des genres plus variés et notamment le genre journalistique. Ces travaux ont montré l'importance de l'identification des paroles et de leurs porteurs dans les tâches courantes du TAL (résumé automatique, extraction d'information, ...). En français, (Mourad & Desclés, 2002; Giguet & Lucas, 2004) ont travaillé sur l'élaboration de modèles généraux de repérage automatique des citations dans le genre journalistique. Leurs travaux ne fournissent toutefois pas de résultats expérimentaux.

La citation est une forme de reprise utilisée dans les articles scientifiques afin de faire référence à des travaux antérieurs et se positionner par rapport à ces derniers (Choi *et al.*, 2005; Teufel *et al.*, 2006). Dans les articles journalistiques elle étaie le propos du journaliste qui recontextualise le discours original (Jackiewicz, 2007). Dans le cadre du projet PIITHIE[1](Plagiats et Impacts de l'Information Textuelle recHerchée dans un contexte InterlinguE), les citations sont des marques de l'impact d'une information. En effet, le contenu des citations reste globalement inchangé parmi les articles traitant d'un même sujet. Contrairement à la typologie proposée par

[1]Ce travail a bénéficié d'une aide de l'Agence Nationale de la Recherche, projet PIITHIE portant la référence 2006 TLOG 013 03. www.piithie.com

(Jackiewicz, 2007) qui nécessite une analyse fine de la citation et son contexte, nous décrivons ici les formes citationnels à partir d'éléments de surface. Nous nous appuyons sur la forme canonique SRD[2] de (Giguet & Lucas, 2004) sur laquelle nous revenons plus en détails dans la section 4.1.2.

Dans cet article, nous cherchons à tracer des pistes pour la mise au point de méthodes efficaces d'extraction automatique des citations, à partir de citations observées au sein d'un corpus que nous avons constitué. Nous nous intéressons à la fois à ce qui est rapporté par l'énonciateur, ce que nous désignons par *discours repris*, et aux termes du discours faisant référence à la source reprise, ce que nous désignons par *expression locuteur*. Ainsi dans l'exemple (1) ci-dessous, *sa mère* est l'expression locuteur tandis que *Julien faisait une sieste dans l'appartement familial lorsqu'il a disparu* est le discours repris.

(1) D'après **sa mère**, *Julien faisait une sieste dans l'appartement familial lorsqu'il a disparu*.

Nous présentons dans un premier temps notre démarche pour la constitution du corpus d'observation. Nous décrivons ensuite quantitativement les citations afin d'apprécier globalement leur distribution et leur structure dans le corpus. Puis, nous les décrivons qualitativement afin de faire émerger des marques utilisables par des méthodes de repérage automatique. Finalement, nous proposons un schéma d'annotation des citations.

2 Construction du corpus et repérage des citations

Il nous semble nécessaire de travailler sur des exemples attestés, mais il n'existe pas à notre connaissance de corpus français où les citations sont annotées, contrairement à l'anglais (Prasad *et al.*, 2007). Seul le groupe Ci-dit[3] propose des corpus de citations extraites de leur contexte. Nous présentons tout d'abord la définition des besoins, puis la sélection et la structuration des textes.

Le corpus doit être représentatif, en terme de citations, de la presse francophone. Nous choisissons de puiser dans la presse généraliste qui présente un large panel de styles d'écriture, tout en privilégiant la variété des journaux et des auteurs. Le corpus constitué doit avoir une taille raisonnable pour être annoté à la main, mais être suffisamment riche en citations. Afin d'éviter de tronquer des citations, étant donné la difficulté à déterminer leurs bornes, nous conservons l'intégralité des textes. Finalement, les textes sélectionnés doivent être stockés dans un format numérique pour faciliter leur manipulation automatique.

Nous avons choisis quatre quotidiens nationaux (Libération, Challenges, Le Monde, Le Figaro) et un quotidien Belge (Le Soir) comme panel représentatif. La taille des articles dans ces journaux est d'environ 600 mots. Nous avons sélectionné une dizaine d'articles par journal, publiés la semaine du 19 Février 2007, et placés en *Unes*, ce qui est gageure d'une certaine qualité éditoriale. Le corpus totalise ainsi 53 articles issus de cinq journaux différents, représentant plus de 36000 mots répartis au sein de 1200 phrases (*cf. Tableau 1*).

Le contexte graphique (publicités, menus, ...) ne nous intéressant pas, nous avons sauvegardé chaque article original, puis stocké une version épurée de ces artefacts sous la forme d'un arbre XML. Les informations de mise en forme comme l'italique, le gras et les emphases ont été

[2]Source, Relateur, Discours

[3]`http://www.ci-dit.com`

Journal	Nb. articles	Nb. mots	Nb. phrases
Libération	10	7336	246
Le Soir	10	4583	177
Challenges	11	11636	427
Le Monde	12	7035	193
Le Figaro	10	5637	159
Totaux	**53**	**36227**	**1202**

TAB. 1 – Répartition des éléments du corpus par journal.

conservées puisqu'elles représentent des marques potentielles pour le repérage automatique des citations. Les versions HTML en ligne n'offrant pas de structuration du contenu, nous avons également balisé, au sein de l'arbre XML, les titres, les sous-titres et les paragraphes.

Les articles du corpus sont "autonomisable" (Habert, 2000), afin d'être utilisables en dehors du cadre initial pour lequel ils ont été collectés. Chaque article est accompagné de son contexte de publication : nom du journal, url à laquelle il a été récupéré, auteurs, date de publication, ... Ces informations permettent de replacer l'article dans son contexte d'écriture, et en pérennisent l'accès en cas de retrait éventuel de la version en ligne. Des standards tels que le *Text Encoding Initiative* (TEI) proposent un tel format de structuration, nous avons toutefois choisi une DTD simplifiée qui semblait mieux répondre à nos besoins.

3 Analyse quantitative des citations du corpus

Une fois le corpus constitué, nettoyé et organisé, nous en extrayons manuellement les citations. Nous décrivons dans un premier temps la distribution des objets citationnels en son sein, puis les styles de discours rapporté reconnus.

3.1 Distribution des éléments citationnels dans le corpus

Nous avons repéré près de 800 objets citationnels (discours repris, expressions locuteur) parmi les 53 articles du corpus, soit environ 400 citations (8 citations par article en moyenne). La figure 1 montre la distribution des citations parmi les articles : les lignes verticales marquent l'amplitude de la variation du nombre de citations par article tandis que les cadres représentent l'intervalle de confiance[4]. Les variations au sein même des journaux sont importantes, l'appartenance à un même journal ne semble donc pas influer sur la présence des citations.

Nous nous concentrons dans la section suivante sur la structure des citations.

3.2 Distribution des styles de discours

Étant donnée une situation d'énonciation initiale, un discours extrait de cette énonciation aura une représentation linguistique différente selon le style auquel il est rapporté. Nous avons pris en considération les styles suivants :

[4]Intervalle où l'on retrouve la majorité de la population pour une distribution Gaussienne : $[Moyenne - \sigma; Moyenne + \sigma]$

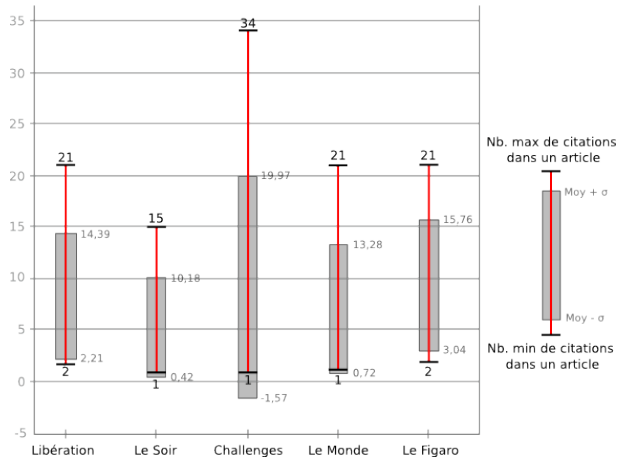

FIG. 1 – Répartition des citations par articles et par journaux

– style direct entre guillemets : le discours repris est placé entre guillemets sans que l'auteur n'apporte de modifications morpho-syntaxiques importantes ;
– style indirect simple : le discours repris a subis des modifications morpho-syntaxiques importantes pour s'intégrer à la nouvelle situation d'énonciation ;
– style indirect libre : seule l'interprétation du texte permet de différencier le discours repris du reste du texte ;
– style indirect avec îlots textuels : style hybride entre le style direct et le style indirect libre.

Les journalistes utilisent très largement le style indirect avec îlots textuels. Les guillemets "servent soit à inclure en subordonnant, soit à exclure en isolant" (Mourad & Desclés, 2002). Dans le discours indirect à îlots textuels, les journalistes excluent en isolant les segments qui correspondent à des reprises *verbatim*. La difficulté est alors de distinguer ces passages *verbatim* des emphases (*cf. exemples 2 et 3*).

> **(2)** Washington avance une estimation des réserves mondiales **"ultimes"** de pétrole à 2 275 milliards de barils.
> **(3)** 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.

Le graphique de la figure 2 illustre la répartition des styles par journaux. L'utilisation des guillemets pour encadrer tout ou partie du discours repris est majoritairement présent. D'une manière générale, les journalistes privilégient les styles limitant les modifications morpho-syntaxiques. Nous faisons l'hypothèse que ces styles sont privilégiés car ils miment la neutralité, l'objectivité.

Ces résultats sont encourageants d'un point de vue opérationnel puisque la majorité des discours repris respectent fidèlement le discours original et héritent ainsi d'un nombre important de marques discursives. La mise au point de techniques de repérage automatique se focalisera donc sur ces styles les plus fréquents.

4 Analyse qualitative des citations du corpus

Après l'aperçu global des citations tracé dans la section précédente, cette section se concentre sur les cas particuliers. L'analyse qualitative présentée dans cette section pose le problème de

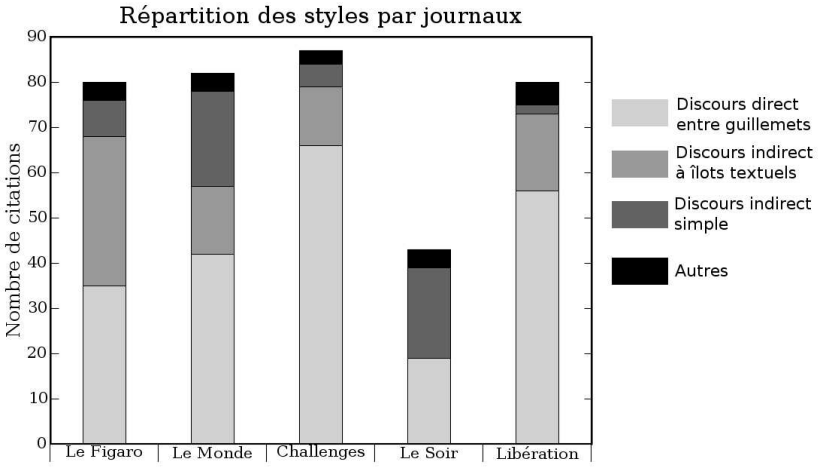

FIG. 2 – Répartition des styles d'intégration du texte englobé au sein des différents journaux

délimitation de la citation puis expose des caractéristiques du discours repris, des expressions locuteur et des relateurs.

4.1 Le problème de délimitation de la citation

Lorsque l'on cherche à extraire automatiquement un objet linguistique, il est nécessaire de pouvoir définir précisément ses bornes. Nous exposons ci-après la difficulté à déterminer les bornes des citations, puis nous présentons une approche alternative basée sur les travaux de (Giguet & Lucas, 2004).

4.1.1 Présentation du problème

La problématique d'une unité citationnelle au niveau linguistique s'illustre particulièrement dans le domaine journalistique. Le discours à rapporter étant trop volumineux, l'auteur doit y opérer des transformations[5]. Afin de montrer son objectivité, il conserve au sein de sa reformulation des passages *verbatim*. Le discours repris est alors constitué d'une juxtaposition de segments reformulés et de passages *verbatim* (*cf. exemple 4*).

(4) Le Figaro estime, lui, que "les techniciens et les cadres sont en première ligne", notamment ceux de la "Central Entity" de Toulouse.

Les modifications peuvent être plus ou moins importantes, parfois une amorce de remise en contexte suffit (*cf. exemple 5* avec "L'administration ne peut"). Il arrive cependant que les segments originaux soient noyés dans une totale reformulation au style indirect (*cf. exemple 6*).

(5) L'administration ne peut "utiliser la convocation à la préfecture d'un étranger [...] pour faire procéder à son interpellation en vue de son placement en rétention", estime-t-elle.

(6) 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.

[5]Dans la typologie de Jackiewicz il s'agit de citation interprétée, l'énonciateur reconstruit ce qui lui semble essentiel dans la situation d'énonciation originale.

Dans certains cas extrêmes, l'auteur tisse des paragraphes entiers reprenant de brefs éléments d'un discours source. Les modifications de l'extrait à intégrer sont alors si nombreuses que le discours repris et le discours source ne possèdent plus que le sens profond en commun, les détails étant complètement passés sous silence (*cf. exemple 7*).

> **(7)** 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.

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.

4.1.2 Proposition d'une unité intermédiaire : la séquence canonique

(Giguet & Lucas, 2004) proposent de représenter une citation comme une séquence de trois objets citationnels élémentaires :
– la source : équivalente à notre expression locuteur ;
– le discours rapporté : équivalent à notre discours repris ;
– le relateur : "segment établissant la relation entre la source et le discours rapporté".
Ces abstractions permettent de prendre du recul sur la notion de citation et nous recentrer sur la notion d'unité citationnelle. L'idée est de définir des motifs d'agencement des objets citationnels correspondant à la structuration des citations dans les articles journalistiques. (Giguet & Lucas, 2004) proposent ainsi les motifs : source + relateur + discours et discours + relateur + source.

En appliquant cette formalisation sur notre corpus, nous avons repéré vingt motifs distincts. Ceux proposés par (Giguet & Lucas, 2004) sont les plus présents avec respectivement 104 (28%) et 148 (40%) occurences, les autres motifs correspondant à des combinaisons de seulement un ou deux des trois composants (disparition de la source ou du relateur) ou à l'apparition d'une deuxième source. Ainsi, 9% des citations du corpus sont dépourvues de relateur ou de locuteur (*cf. exemple 8*), il s'agit des motifs les plus importants en nombre, ceux de (Giguet & Lucas, 2004) mis-à-part.

> **(8)** 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.»

Les autres motifs sont : *relateur* + *source* + *discours* et *discours* + *relateur* + *source* + *discours*, ils représentent 15% des citations rencontrées dans notre corpus. Le premier motif semble spécifiquement utilisé lorsque le syntagme prépositionnel introduisant l'expression locuteur est propulsé en début de phrase (cf. *exemple 9*), alors que le second est utilisé lorsque l'expression locuteur est placée dans une proposition juxtaposée entre deux fragments du discours repris.

> **(9)** 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».

L'apparition de nouveaux motifs ne permet plus de déduire un éventuel troisième composant d'après sa position, ce qui réduit l'efficacité de l'approche de (Giguet & Lucas, 2004). Nous proposons une adaptation de ce modèle plus proche de la représentation linguistique et tout en conservant l'essence de la forme canonique : *le segment citationnel*. Un segment citationnel se compose, en autorisant les répétitions, des objets citationnels suivants :

– un discours repris — pour lequel le problème de bornage reste entier — ;
– une expression locuteur potentiellement omise si la source est suffisamment saillante ;
– un relateur prenant la forme d'une expression linguistique, d'une ponctuation ou directement induit par la structure de la phrase.

Nous avons désormais un aperçu assez précis de la structuration globale des citations au sein de notre corpus. Les sections suivantes s'attachent à en observer les constituants pour les caractériser.

4.2 Caractérisation du discours repris

L'observation des citations du corpus nous a permis d'extraire des régularités qui devraient permettre de définir des méthodes adaptées aux différents types de citation présents au sein des textes journalistiques. Au delà de la structure, le style de discours rapporté employé influe sur la forme de la citation. De ce point de vue, nous présentons dans un premier temps les régularités concernant les citations au style direct, puis celles concernant le style indirect.

Nous avons observé deux constructions prédominantes pour le style direct dans la manière de lier l'expression locuteur et le discours repris : soit à l'aide d'une proposition juxtaposée, soit uniquement par un relateur. D'une manière générale, on remarque que si l'expression locuteur est en début de phrase, un relateur est utilisé, sinon l'expression locuteur est juxtaposée (*cf. exemple 10*). De plus, lorsque le discours repris entre guillemets se positionne comme une phrase à part entière, l'expression locuteur est introduite au sein même des guillemets (*cf. exemple 11*).

(10) "Avec la fin de la session, les parlementaires sont beaucoup plus disponibles. Cette réorganisation était donc nécessaire", renchérit-**on dans l'entourage de Sarkozy**.

(11) "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".

Nous avons également observé deux constructions majeures pour le style indirect : les propositions subordonnées conjonctives et les propositions juxtaposées. La première construction fait majoritairement appel à un relateur du type *verbe + que* (*cf. exemple 12*), alors que dans le second cas la partie reprise est juxtaposée à un syntagme prépositionnel contenant l'expression locuteur (*cf. exemple 13*). Une difficulté majeure propre au style indirecte est de suivre la continuité d'un discours repris sur plusieurs phrases (*cf. exemple 14*).

(12) 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.

(13) **D'après sa mère,** Julien faisait une sieste dans l'appartement familial lorsqu'il a disparu.

(14) 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* [...] *sur la même trajectoire*.

En résumé, deux grandes structures ressortent aussi bien pour le discours indirect que pour le discours direct. La première fait appel à des relateurs ponctuatifs ou verbaux qui mettent en relation l'expression locuteur et le discours repris. La seconde consiste en la mise en juxtaposition de l'expression locuteur et du discours repris, ce dernier pouvant s'étendre sur plusieurs phrases dans le cas du style indirect.

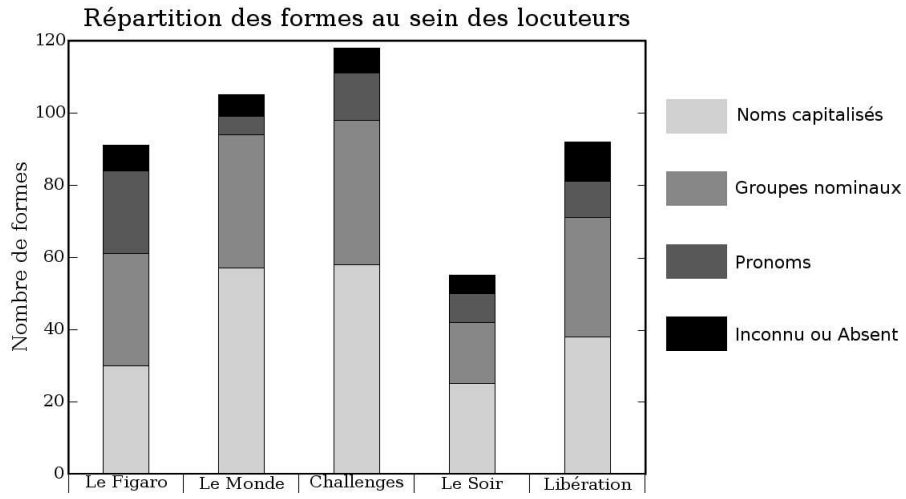

FIG. 3 – Répartition des formes au sein des locuteurs

4.3 Caractérisation des expressions locuteur

La recherche de l'expression locuteur est primordiale, notamment dans le cadre de la détection de plagiat. Nous présentons ci-dessous les caractéristiques des expressions locuteur de notre corpus.

L'observation de (Giguet & Lucas, 2004) selon laquelle les "réductions lexicales et les anaphores auront tendance à apparaître après des formes très déterminées" se confirme au sein de notre corpus (*cf. exemple 15*). La figure 3 décrit les éléments linguistiques composant les différentes expressions locuteurs. Les groupes nominaux et les noms capitalisés y sont les plus utilisés. Les formes réduites faisant référence aux sources correspondent majoritairement à des pronoms ou des adjectifs possessifs.

(15) **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".

Nous avons également considéré les cas où il n'y avait pas de locuteur dans le contexte phrastique du discours repris. Deux solutions apparaissent alors : le plus fréquemment, le locuteur a été précédemment introduit et est suffisamment saillant pour permettre au lecteur de faire le lien (*cf. exemple 16*) ; plus rarement, le contexte permet d'identifier la source (*cf. exemple 17*).

(16) 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."

(17) "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.

En résumé, les formes employées dans les expressions locuteurs varient selon qu'il s'agisse d'une première introduction du dit locuteur (noms capitalisés, groupes nominaux, ...) ou bien d'une référence à un locuteur auparavant introduit (pronoms, adjectifs possessifs, ...). Le suivi de la saillance des différentes expressions locuteur introduites devrait permettre quant à elle de résoudre le problème d'absence de locuteur dans le contexte phrastique.

4.4 Caractérisation des relateurs

L'objet citationnel *relateur* introduit par (Giguet & Lucas, 2004) est le plus complexe à caractériser. Nous classons les relateurs en deux catégories : les prépositions (*selon*, *pour*, *d'après*, ...) et les regroupements de formes verbales et de compléments circonstanciels ou d'adverbes (*cf. exemples 18*). Ces derniers sont les plus utilisés et les plus précis. Les satellites du verbe renseignent en effet sur le contexte d'énonciation.

(18) Le premier ministre, Tony Blair, **a annoncé dimanche à la BBC** vouloir durcir la loi sur les armes à feu afin de lutter contre leur circulation parmi les jeunes.

Nous choisissons de restreindre les relateurs aux formes verbales ou prépositionnelles, car nous considérons les autres formes trop hétérogènes et non spécifiques aux citations. De plus, cette restriction nous permet de profiter des ressources issues de travaux antérieurs, comme les verbes d'énonciation de (Mourad & Minel, 2000).

5 Proposition d'un schéma d'annotation des citations

Plusieurs schémas d'annotation (Prasad *et al.*, 2007; Jackiewicz, 2007) ont été décrits pour l'annotation des opinions et de leurs porteurs. Nous proposons un schéma comparable en certains points, mais largement simplifié pour l'annotation de citations dans l'objectif de mettre au point des méthodes de repérage automatique

Nous avons dans un premier temps considéré l'annotation des relateurs, cependant le degré d'accord[6] trop faible entre trois annotateurs sur un sous-ensemble du corpus nous a poussé à nous limiter finalement aux expressions locuteur et aux discours repris.

L'annotation que nous proposons est fondée sur l'emploi de deux balises aux noms explicites : la balise <*expressionlocuteur*> et la balise <*discoursrepris*>. La deuxième balise fait référence à la première à l'aide d'un attribut *source* prenant pour valeur l'identifiant de la balise <*expressionlocuteur*> auquel le discours repris balisé se rapporte. La reconstitution des citations s'effectue en compilant les balises <*discoursrepris*> référant à une même balise <*expressionlocuteur*>. Le regroupement de ces balises avec la balise <*expressionlocuteur*> correspondante constitue une forme approximative de la citation : le *segment citationnel*[7]. Cette forme résoud partiellement la problématique de l'unité citationnelle sur le plan linguistique.

Ce schéma d'annotation nous a servi pour annoter notre corpus. Il subsiste toujours des incertitudes quant à la délimitation des discours repris, mais nous les avons contourné en sélectionnant au plus large. 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.

6 Conclusion

Le corpus francophone que nous avons constitué nous a permis d'observer plusieurs centaines de citations dans leur contexte. Nous avons ainsi été en mesure de faire émerger des caracté-

[6] Le degré d'accord entre les annotateurs n'a pas été calculé automatiquement, mais évalué par comparaison directe des annotations

[7] cf. Section 4.1.2

ristiques concernant les différents objets citationnels constituant les dites citations. Finalement, nous avons proposé un schéma d'annotation permettant de baliser en XML les citations du corpus en contournant la problématique d'unité citationnelle grâce à l'introduction du *segment citationnel*.

La constitution du corpus et l'analyse des citations qu'il regroupe ouvre des perspectives pour le repérage automatique des citations. Nous cherchons à mettre en place des méthodes de repérage utilisant uniquement des marques de surface. Ainsi, nous envisageons l'utilisation d'automates pour l'identification d'expressions locuteur et des techniques statistiques pour le repérage des discours repris. Le corpus servirait alors de base d'apprentissage. L'article à paraître (Poulard *et al.*, 2008) discute certaines de ces méthodes de repérage automatique à partir du corpus et des particularités précédemment présentés.

Références

CHOI Y., CARDIE C., RILOFF E. & PATWARDHAN S. (2005). Identifying sources of opinions with conditional random fields and extraction patterns. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, p. 355–362, Vancouver, Canada : Association for Computational Linguistics.

GIGUET E. & LUCAS N. (2004). *La détection automatique des citations et des locuteurs dans les textes informatifs*, In J. M. L. MUÑOZ, S. MARNETTE & L. ROSIER, Eds., *Le discours rapporté dans tous ses états : Question de frontières*, p. 410–418. L'Harmattan.

HABERT B. (2000). Des corpus représentatifs : de quoi, pour quoi, comment. In M. BILGER, Ed., *Cahiers de l'Université de Perpignan*, volume 31, p. 11–58. Presses Universitaires de Perpignan.

JACKIEWICZ A. (2007). Relations intersubjectives dans les discours rapportés. In M.-P. PÉRY-WOODLEY & D. SCOTT, Eds., *Journal TAL : Discours et document : traitements automatiques*, volume 47, p. 65–87. ATALA.

KIM S. & HOVY E. (2006). Extracting Opinions, Opinion Holders, and Topics Expressed in Online News Media Text. In *Proceedings of ACL/COLING Workshop on Sentiment and Subjectivity in Text*, Sydney, Australia.

MOURAD G. & DESCLÉS J. (2002). Citation textuelle : identification automatique par exploration contextuelle. In L. DANON-BOILEAU & M.-A. MOREL, Eds., *Faits de langues*, number 19, p. 179–188. Ophrys.

MOURAD G. & MINEL J.-L. (2000). Filtrage sémantique du texte, le cas de la citation. In G. M. & T. E., Eds., *3e Colloque International sur le Document Électronique*, p. 41–56 : Lavoisier.

POULARD F., WASZAK T., HERNANDEZ N. & BELLOT P. (2008). Repérage de citations, classification des styles de discours rapporté et identification des constituants citationnels en écrits journalistiques. In *À paraître TALN 2008*, Avignon, France.

PRASAD R., DINESH N., LEE A., JOSHI A. & WEBBER B. (2007). Attribution and its annotation in the penn discourse treebank. In M.-P. PÉRY-WOODLEY & D. SCOTT, Eds., *Journal TAL : Discours et document : traitements automatiques*, volume 47. ATALA.

TEUFEL S., SIDDHARTHAN A. & TIDHAR D. (2006). Automatic classification of citation function. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, volume 6, p. 103–110.