Traitement Automatique du Discours Rapporté

Pierre-André Buvet[1]

[1]LDI CNRS-Université Paris 13 pabuvet@ldi.univ-paris13

Abstract

We discuss about the main features of the reported speech that must be taken into account to automatically identify citations, then we describe the resources which allow automatic processing of reported speech. The work presented here is about a research with the goal to identify citations and allocated them a weight according to the semantic markers.

Résumé

Nous discutons des principales caractéristiques du discours rapporté qu''il faut prendre en compte pour identifier automatiquement les citations, puis nous décrivons les ressources qui permettent le traitement automatique du discours rapporté. Les travaux présentés portent sur une recherche en cours dont l'objectif général est d'identifier automatiquement les citations, et de leur allouer un poids selon les marqueurs sémantiques spécifiés.

0. Introduction

La linguistique informatique a comme finalité, d'une part, d'analyser et d'expliquer les faits de langue et, d'autre part, de représenter les résultats obtenus d'une façon formelle, c'est-à-dire d'une façon totalement explicite et reproductible, de telle sorte qu'ils fassent l'objet d'exploitations informatiques dédiées au traitement de l'information textuelle. Les faits de langue étudiés ici concernent le discours rapporté. L'identification automatique des citations dans les textes ne va pas de soi car le repérage des marques du discours rapporté, notamment lorsqu'il est du type indirect, est fondé sur la combinatoire d'unités linguistique hétérogènes. L'analyse du discours rapporté n'en est pas moins fondamentale pour interpréter un texte car les informations n'ont pas le même poids sémantique selon qu'elles sont directement prises en charge par le locuteur ou présentées comme des citations.

Nous discutons des principales particularités du discours rapporté dont il faut tenir compte pour identifier automatiquement les citations, nous présentons ensuite les ressources qui permettent le traitement automatique du discours rapporté, puis nous précisons comment s'effectue ce traitement. Les travaux exposés ici portent sur une recherche en cours de réalisation dont l'objectif général est d'identifier automatiquement les citations, d'indiquer à qui elles doivent être attribuées et de leur allouer un poids sémantique en fonction des particularités énonciatives des marqueurs de discours rapporté. L'évaluation des résultats obtenus jusqu'à présent montre la faisabilité de l'étiquetage effectué.

1. Analyse des données linguistiques

Le discours rapporté implique «un dédoublement de l'énonciation : le discours tenu par un locuteur de base contient un discours attribué à un autre locuteur, qui est rapporté par le locuteur premier» [RIEGEL & Alii 1994]. Après avoir défini la parole citée et mentionné ses principales propriétés linguistiques, nous exposons une typologie des formes du discours rapporté, puis nous discutons de la dimension modale du discours rapporté.

1.1. La parole citée

Parmi les nombreux travaux portant sur le discours rapporté dans le champ disciplinaire de la linguistique, il y en a peu qui s'inscrivent dans la perspective du traitement automatique des langues[1]. Trois aspects du discours rapporté sont étudiés ici : (i) la façon dont procède le locuteur pour stipuler qu'il insère le discours de son interlocuteur dans son propre discours ; (ii) la forme que prend le discours de l'interlocuteur dans celui du locuteur ; (iii) la manière dont le discours de l'interlocuteur est interprété par le locuteur.

L'étude du discours rapporté consiste à identifier les procédés linguistiques permettant de faire une citation puis à expliquer leur mode de fonctionnement. Le discours rapporté impliquant une double énonciation, il s'agit aussi de comprendre comment les deux énonciations interagissent lorsque le locuteur prend en charge le discours d'un autre, [KERBRAT-ORECCHIONI 1980] et [TODOROV 1970]. L'analyse de cette prise en charge dépend de trois paramètres.

Le premier paramètre concerne les indices de discours rapporté ; leur forme peut être de nature diverse : *vous dites* dans *Vous dites souhaiter un arrangement à l'amiable*, *ayant pris connaissance de* dans *Ayant pris connaissance de votre demande d'un arrangement à l'amiable*, *vous souhaitez* dans *comme vous souhaitez un arrangement à l'amiable*, *le souhait* dans *le souhait d'un arrangement à l'amiable*, etc. [ROSIER 2008].

Le deuxième paramètre porte sur le caractère plus ou moins explicite des indices. Le même propos est exprimé dans les quatre énoncés ci-dessus mais l'indication d'un discours rapporté est plus marquée dans les deux premiers que dans les deux derniers. Dans le quatrième énoncé c'est seulement la situation d'énonciation (il s'agit d'un dialogue) qui permet cette interprétation.

Le troisième paramètre se réfère à la position plus ou moins distanciée du locuteur par rapport au discours rapporté ; il s'agit de la modalité [CHARAUDEAU 1992]. Elle peut correspondre à une modalité élocutive, c'est-à-dire centrée exclusivement sur le locuteur (par exemple *Nous sommes ravis* dans *Nous sommes ravis d'apprendre votre souhait d'un arrangement à l'amiable*) ou à une modalité allocutive, c'est-à-dire elle implique également l'interlocuteur (*Nous approuvons pleinement* dans *Nous approuvons pleinement votre demande d'un arrangement à l'amiable*). La modalité en rapport avec le discours rapporté, qu'elle soit élocutive ou allocutive, peut porter sur l'énoncé, comme dans les exemples précédents, ou bien sur l'énonciation, c'est-à-dire la position plus ou moins distanciée de l'interlocuteur par rapport à son propre énoncé (*Votre déception est tout à fait légitime*).

L'étude des indicateurs du discours rapporté est un préalable à celle des formes du discours cité. On distingue traditionnellement les discours rapportés directement des discours rapportés

1 On peut citer néanmoins des travaux en linguistique formelle, par exemple SAGO & DANLOS 2010, FAIRON 2000, GIGUET et LUCAS. 2004 et MOESCHLER J. 1989.

indirectement. Dans le premier cas de figure, l'analyse ne présenterait aucune difficulté particulière dans la mesure où le contenu attribué à l'interlocuteur est accompagné d'indices typographiques spécifiques, typiquement les guillemets, et il est censé avoir strictement la même forme que dans l'énoncé cité. Dans le second cas de figure, l'analyse est plus complexe car des variations de forme distinguent le discours cité du discours citant ; la question centrale est alors celle de la paraphrase. Il s'agit ici d'une reformulation dont l'identification est d'autant plus nécessaire que l'indice est peu explicite voire implicite. De nouveaux paramètres d'analyse entrent en ligne de compte ; nous en avons retenu trois.

Le premier paramètre concerne les variations qui affectent l'agencement de l'énoncé cité, c'est-à-dire la reformulation est constituée des mêmes mots, ou d'une partie d'entre eux, mais disposés autrement (source dans l'énoncé cité : *Je n'ai toujours pas réceptionné ma commande*/ reprise dans l'énoncé citant : *Votre commande n'a toujours pas été réceptionnée*). Nous appelons 'restructurations' de tels phénomènes. Il s'agit des différentes instances des structures prédicat-argument dans les énoncés qui ne correspondent pas à des phrases canoniques. Les restructurations sont les phrases passives (comme celle de l'exemple ci-dessus), les phrases clivés (*Ma commande, je ne l'ai toujours pas réceptionnée*), les phrases nominalisées (*La réception de ma commande n'a pas eu lieu*), les infinitives (*Ne pas avoir réceptionné votre commande*) etc. En rapportant un ensemble de restructurations à une même phrase canonique, on fait état d'un même contenu propositionnel quelles que soient les formes qu'il recouvre.

Le deuxième paramètre a trait aux variations qui affectent le lexique de l'énoncé cité, c'est-à-dire la reformulation est constituée de mots différents, en partie ou en totalité, mais conserve la même signification (source dans l'énoncé cité : *Je n'ai pas reçu mon chèque*/reprise dans l'énoncé citant : *Vous n'avez pas touché votre argent*). Il s'agit de paraphrases fondées sur les relations sémantiques, du type synonymique ou hyperonymique, au sein du lexique [BUVET 2001].

Le troisième paramètre porte sur les variations qui affectent les prédications de l'énoncé cité parce qu'il est rapporté d'une façon synthétique. Typiquement, la reformulation est un groupe nominal résumant deux contenus propositionnels attribués à un autre locuteur (sources dans l'énoncé cité : *J'ai une carte fidélité que je voudrais activer via internet mais je n'y arrive pas*/reprise dans l'énoncé citant : *Les difficultés rencontrées pour activer votre carte de fidélité*). La difficulté ici est d'établir que la construction paraphrastique a comme source deux structures prédicat-argument différentes. La reformulation peut correspondre également à un phrase complexe (sources dans l'énoncé cité : *Je m'inquiète car cela fait plusieurs fois que vous essayé de m'envoyer un mandat et que je ne reçois rien*/reprise dans l'énoncé citant : *Vous vous inquiétez de ne pas avoir reçu vos mandats*). Dans de tels cas, les stratégies qui envisagent d'exploiter les restructurations (premier paramètre) ou les relations sémantiques au sein du lexique (deuxième paramètre) sont nécessaires mais non suffisantes pour relier la reprise à sa source car elles impliquent généralement des phénomènes d'ellipse qui rendent difficile l'identification des structures prédicat-argument dans l'énoncé citant.

1.2. Typologie des formes du discours rapporté

Il est tenu compte dans cette étude uniquement des discours directs et indirects non libres. Ils sont l'un et l'autre structurés en deux composantes. Dans la première composante, le locuteur mentionne un autre discours que le sien ; ce qui implique l'identification du locuteur de l'énoncé cité et des marqueurs de discours rapporté. Seuls les marqueurs correspondant à des verbes sont

considérés ici, mais rappelons qu'il en existe de toutes sortes, par exemple *selon* dans *Selon Paul, tout va bien* [CHAROLLES 1997]. La citation proprement dite figure dans la seconde composante, son identification découle de la première.

La forme canonique de la première composante est du type :

GN_HUMAIN V_PAROLE

L'élément GN_HUMAIN équivaut soit à une entité nommée (*Léonard raconte qu'il a passé son enfance au* Québec ; Léonard *Cohen raconte qu'il a passé son enfance au Québec* ; *Monsieur Cohen raconte qu'il a passé ses vacances au Québec*) soit à un *pronom* (*Il raconte qu'il a passé ses vacances au Québec*) soit un groupe nominal formé à partir d'un nom d'HUMAIN *(Un chanteur très connu raconte qu'il a passé son enfance au Québec* ; *Le chanteur raconte qu'il a passé son enfance au Québec*). Lorsque l'élément est un pronom ou un groupe nominal défini, il faut établir en outre à qui il est fait référence soit du point de vue de l'anaphore soit du point de vue de la deixis [KLEIBER 1990].

Le symbole V_PAROLE signifie qu'il s'agit d'un verbe de la classe sémantique PAROLE, [ESHKOL I. 2002]. Les verbes de cette classe ne sont homogènes ni sur le plan syntaxique ni sur le plan sémantique. Tous ne sont pas sont des marqueurs de discours rapportés (*Il papote avec elle* ; **Il papote avec elle de ses vacances* ; *Il jacasse avec elle* ; **Il jacasse avec elle de ses vacances* ; *Il parle à Léa* ; *Il parle à Léa de ses vacances*) et lorsqu'ils le sont, ils n'ont pas touts la même distribution droite (*Il raconte qu'il a passé ses vacances en Irlande* ; *Il raconte ses vacances en Irlande* ; *Il raconte avoir passé ses vacances en Irlande* ; *Il dit qu'il a passé ses vacances en Irlande* ; **Il dit ses vacances en Irlande* ; *Il dit avoir passé ses vacances en Irlande* ; **Il bavarde qu'il a passé ses vacances en Irlande ; Il bavarde de ses vacances en Irlande* ; *Il bavarde avoir passé ses vacances en Irlande*). Sur le plan sémantique, on observe de fortes disparités entre ces verbes. La signification de verbes comme *dire* se limite à l'introduction d'un discours cité (*Albert Camus a dit que la tâche de sa génération consiste à empêcher que le monde se défasse*) alors que d'autres ont des significations supplémentaires (*Il a crié que ça suffisait* ; *Il m'a chuchoté que mes invités venaient d'arriver* ; *Il a exigé que nous partions* ; *Il a répondu que ce n'était pas à l'ordre du jour*).

Les formes de la citation dépendent du type de discours rapporté. Pour ce qui est du type direct, la parole de l'autre est présentée comme telle et il incombe aux guillemets de spécifier que la citation est reproduite d'une manière fidèle. Il s'ensuit que tout contenu propositionnel ainsi structuré et précédé d'une mention à un humain combinée à un marqueur de discours rapporté est identifiable comme un discours cité (*Il a dit : « Je suis d'accord et je ferai selon vos instructions* »). Il s'agit d'une construction canonique. Les constructions qui ne le sont pas mettent en jeu des incises soit à l'intérieur du discours rapporté (*« Je suis d'accord, a-t-il dit, et je ferai selon vos instructions* ») soit en postposition (*« Je suis d'accord, a-t-il dit, et je ferai selon vos instructions* ». Tous les discours rapportés du type direct ne sont pas construits de la façon indiquée. Des indices graphiques suppléent parfois aux guillemets, par exemple, lorsque le discours cité est écrit en gras ((*Il a dit : Je suis d'accord et je ferai selon vos instructions*). Pour l'instant, une telle situation n'est pas repérée. Par contre, il est tenu compte des formes mixtes telles que *Il dit « nulle situation n'est perdue »* qui sont abondamment représentées dans la presse.

En ce qui concerne le type indirect, la parole de l'autre est présentée sous la forme d'une complétive (*Il a avoué qu'il avait commis le crime*), d'une infinitive (*Il a avoué avoir commis le crime*) ou d'un groupe nominal (*Il a avoué le crime*). La principale difficulté pour identifier correctement la citation est de repérer comment elle s'achève. Dans les corpus pris en compte, des marques de ponctuation sont exploités pour clore la citation (*Il ajoute que le prix actuel de l'action est intéressant, avec la possibilité d'un retour dans la partie supérieure de la fourchette des 29/34 euros au cours des mois à venir*.). Signalons également le cas où la citation est un pronom anaphorique ou déictique (*Il a dit cela*)

1.3. Modalité et discours rapporté

L'analyse syntactico-sémantique du lexique permet d'interpréter les contenus propositionnels constitutifs d'un discours en les rapportant à autant de structures prédicat-argument. Il en a résulté, dans la perspective du Traitement Automatique des Langues, la création de dictionnaires électroniques et de grammaires locales, *cf*. *infra*. L'interprétation d'un discours n'est pas uniquement fondée sur les contenus propositionnels qu'il incorpore, elle repose également sur des facteurs énonciatifs.

La conception du traitement sémantique des textes diffère selon qu'il porte sur des unités lexicales ou sur des unités énonciatives. Les deux types de traitement sémantique sont présentés comme antagonistes dans la mesure où le premier procéderait de la linguistique de la phrase, le second de la linguistique de l'énoncé [REBOUL & MOESCHLER 1998]. Nous rendons caduque ce clivage en tenant compte de tous les facteurs sémantiques, c'est-à-dire aussi bien d'ordre lexico-syntaxique que d'ordre énonciatif, et en établissant en quoi leur combinaison contribue à l'interprétation des textes. Nous appelons 'linguistique de discours' cette approche intégrative [BUVET 2011]

L'analyse sémantique trois niveaux discursifs : le niveau logico-sémantique ; le niveau énonciatif ; le niveau interprétatif. Le premier niveau concerne la langue ; il est postulé qu'elle est structurée en termes de prédicat et d'argument. Le second niveau est centré sur la position du locuteur par rapport à ce qu'il énonce. Le troisième niveau a trait à la compréhension du discours, c'est-à-dire à la représentation qu'en a l'interlocuteur. Les niveaux logico-sémantique et interprétatif ne sont pas directement observables, seul le niveau énonciatif l'est car il implique la production d'un discours.

L'analyse des faits de langue est fondée sur les trois fonctions primaires : la fonction prédicative, la fonction argumentale et la fonction actualisatrice [MEJRI 2009]. Elles permettent de catégoriser les unités linguistiques sur le plan syntactico-sémantique et d'expliquer leur rôle dans la construction d'un énoncé. Les deux premières fonctions rendent compte notamment du contenu d'un énoncé ; la fonction prédicative stipule quel est l'élément structurant et la fonction argumentale, relativement à l'élément structurant, quels éléments sont mis en relation ou quel élément est qualifié. La fonction actualisatrice fait état des autres éléments, c'est-à-dire ceux dont dépend l'instanciation de la structure prédicat-argument dans l'énoncé. La fonction prédicative et la fonction argumentale concernent également la structure des textes ; la première caractérise les connecteurs logiques, explicites ou implicites, la seconde les énoncés qui sont dans la portée des connecteurs [GROSS et PRANDI 2004].

La modalité est définie comme la prise en charge par le locuteur du contenu de son discours. Elle est traitée en termes d'actualisation, d'une part, de catégorisation lexico-énonciative, d'autre part. Les deux types de traitement participent à l'analyse sémantique du discours rapporté. Par manque de place, nous discutons uniquement du second type dans la mesure où il est moins étudié que le premier. La catégorisation lexico-énonciative résulte des propriétés sémantiques des prédicats car elles les impliquent dans un type de modalité. Par exemple, l'adjectif *triste* peut être constitutif d'une assertion, *On pleure quand on est triste*, ou, en tant que description subjective, contribuer à l'expression d'une modalité élocutive, *C'est triste de faire cela*. De même, le verbe *applaudir* peut être constitutif d'une assertion, *La presse a applaudi sa performance*, ou, en tant que description interindivuelle, peut contribuer à l'expression d'une modalité allocutive, *Je vous applaudis des deux mains*, *cf. infra*.

Appliquée aux marqueurs du discours rapporté, la catégorisation lexico-énonciative explique les deux modalités que l'on observe dans les discours rapportés. La première est la modalité élocutive ; elle concerne aussi bien le locuteur du discours cité (*En laçant ses chaussures, mon fils bougonne que ce n'est pas évident*) que le locuteur du discours citant (*Mon voisin prétend que ce n'est pas de sa faute*). La seconde est la modalité allocutive ; elle concerne uniquement le locuteur du discours cité (*La police nous a interdit de quitter la ville*). Les verbes *bougonner*, *prétendre* et *interdire* correspondent respectivement à des prédicats caractérisés par les classes sémantiques PAROLE_DESAGREMENT, PAROLE_ARROGANCE et PAROLE_INJONCTION. Du point de vue énonciatif, les deux premières classes sont catégorisées comme des descriptions subjectives et la troisième comme une description interindividuelle. La première composante du discours rapporté implique toujours la mention explicite du locuteur du discours cité et la mention implicite du locuteur du discours citant, *cf. supra*. Cette composante implique également une occurrence d'un marqueur du discours rapport. Lorsqu'il s'agit d'un verbe, il est associé à l'une des deux étiquettes suivantes : DESCRIPTION_SUBJECTIVE, DESCRIPTION_INTERINDIVIDUELLE. On déduit de la présence de ces deux étiquettes et de la mention d'un locuteur qu'il s'agit d'une modalité élocutive dans les deux premiers énoncés, d'une modalité allocutive dans le dernier énoncé. La prise en compte de ces modalités dans l'analyse sémantique conduit à pondérer les informations reconnues qui sont dans leur portée.

2. Ressources linguistiques

Nous présentons les trois sortes de ressources linguistiques qui ont permis l'étiquetage automatique des citations dans des textes journalistiques.

2.1. Les corpus

Les corpus exploités en Traitement Automatique des Langues sont généralement des textes bruts ou des textes bruts annotés [MITKOV 2003]. On peut les distinguer selon leur finalité : corpus de travail, corpus d'apprentissage, corpus d'expérimentation, corpus d'évaluation, etc. Chaque type de corpus correspond à une tâche spécifique. Le corpus de travail est dédié aux recherches qui fournissent des outils linguistiques, le corpus d'apprentissage aux recherches qui impliquent une analyse statistique des textes, le corpus d'expérimentation aux recherches menées dans un cadre expérimental et le corpus d'évaluation a pour fonction de vérifier la qualité des résultats d'une analyse automatique.

L'exploitation d'un corpus de travail nécessite qu'il soit bien profilé. Par exemple, dans le cadre de l'acquisition automatique de vocabulaire, il est impératif d'analyser préalablement les sources de l'information traitée afin de vérifier que les documents textuels qui constitueront le corpus comportent un lexique suffisamment riche en ce qui concerne les thématiques étudiées [TROMEUR 2011]. De ce point de vue, le recours à des mots-clefs et l'analyse des méta-informations contenues dans les balises qui structurent les textes s'avèrent très insuffisants car, s'ils permettent d'identifier la thématique du document, ils ne garantissent pas d'accéder à un vocabulaire abondant et diversifié en rapport avec le domaine étudié. Typiquement, une majorité d'articles de journaux récupérés à partir des termes 'automobile' et 'industrie de l'automobile' ne sont pas constitués en rapport avec les voitures (*moteur*, *roue, portière, arbre à came*, etc.). Il est donc impératif d'effectuer en amont de la tâche de constitution de corpus, un travail documentaire et linguistique pour s'assurer que les textes traités conviennent au traitement envisagé.

L'élaboration d'outils performants pour effectuer les analyses linguistiques est fondée sur l'exploitation de trois types de corpus de travail: (i) le corpus d'investigation ; (ii) le corpus de test ; (iii) le corpus de validation. Les trois types portent sur des contenus spécifiques pour éviter les solutions ad hoc.

Le corpus d'investigation permet d'identifier les phénomènes linguistiques qui seront pris en compte et traités automatiquement. Compte tenu de ces phénomènes, il doit indiquer en premier lieu quelles ressources lexicographiques sont nécessaires au bon fonctionnement des outils d'analyse. Il doit permettre en second lieu de calibrer les transducteurs à états finis existants et, le cas échéant, de concevoir de nouveaux transducteurs.

La fonction du corpus de test est d'expérimenter les outils calibrés ou nouvellement développés afin de les corriger au fur et à mesure lorsqu'ils donnent lieu à du bruit ou à du silence. Il doit permettre également de pointer d'éventuelles défaillances lexicographiques. L'exploitation du corpus de test doit accompagner au plus près la mise en place des outils afin d'anticiper les difficultés que peut entrainer leur utilisation après leur intégration dans une plateforme.

Le corpus de validation fournit des résultats qui permettent de vérifier la qualité de l'ensemble des outils créés ou paramétrés. En cas d'invalidation, d'autres corpus d'investigation et de test sont utilisés pour améliorer les outils.

Les corpus de test et de validation sont des éléments essentiels à l'élaboration des dictionnaires et les grammaires locales. Ils doivent permettre d'établir puis de mesurer l'adéquation entre ces ressources et les phénomènes traités aussi bien au niveau intra-phrastique qu'au niveau inter-phrastique.

Les corpus exploités pour la présente étude sont des corpus journalistiques portant sur des sujets politiques, économiques ou sociétaux. Ils ont été obtenus à partir de l'outil « google news ».

2.2. Les dictionnaires électroniques

Les dictionnaires électroniques que nous utilisons sont de deux sortes :

(i) des dictionnaires morphosyntaxiques au format MORFETIK [MATHIEU-COLAS 2009] et [MATHIEU-COLAS et Alii 2009] ;

(ii) des dictionnaires syntactico-sémantiques au format PRED-DIC, ARGU-DIC, ACTU-DIC et ETHU-DIC [BUVET 2009a]

Le format PRED-DIC concerne les prédicats, le format ARGU-DIC les arguments, le format ACTU-DIC les actualisateurs et le format ETHU-DIC les êtres humains. Nous présentons rapidement le premier et le dernier format.

La macrostructure des dictionnaires du type PRED-DIC est constituée des différents emplois prédicatifs. Les emplois prédicatifs sont les différentes instances des prédicats dans les phrases. Le niveau d'analyse des emplois prédicatifs diffère de celui des prédicats [BUVET 2009b].

La microstructure des dictionnaires du type PRED-DIC est constituée d'une entrée et de deux catégories de descripteurs formels : les descripteurs de définition ; les descripteurs de conditions. La fonction des descripteurs de définition est d'indiquer les propriétés sémantiques des emplois, celle des descripteurs de conditions est de vérifier que les emplois prédicatifs ont les propriétés remarquables qui les caractérisent.

Les descripteurs de définition donnent lieu à quatre sortes de spécification : la racine prédicative, la classe sémantique, le type sémantique et l'aspect inhérent. Les spécifications de racine et de classe permettent de préciser la nature de l'articulation entre les emplois prédicatifs et les prédicats. Les interprétations des emplois prédicatifs résultent des quatre spécifications.

entrée (emploi)		Définition			
emploi	indice	racine prédicative	Classe	type	aspect inhérent
bougonner	1	bougonn-	PAROLE_DESAGREMENT	action	duratif perfectif
prétendre	1	préten-	PAROLE_ARROGANCE	action	duratif perfectif
Interdire	1	Interdid-	PAROLE_INJONCTION	action	duratif perfectif

Les descripteurs de conditions donnent lieu à trois sortes de spécifications : la construction ; la distribution morphosyntaxique ; la distribution syntactico-sémantique. Les spécifications de construction permettent de faire état du mode de structuration des arguments par rapport à l'emploi prédicatif, du nombre d'arguments spécifiés et de leur caractère obligatoire ou facultatif en position complément. Les spécifications de distribution morphosyntaxique portent sur les particularités formelles des arguments. Les descripteurs consistent à préciser quelles constructions occupent les positions argumentales (groupes nominaux, complétives ou infinitives). Les spécifications de distribution syntactico-sémantique portent sur les particularités sémantiques des arguments. Les descripteurs sont des classes sémantiques.

entrée (emploi)		Conditions		
emploi	indice	Construction	distribution 1	distribution 2
bougonner	1	X0 V X1	X0=GN X1=COMPLETIVE INFINITIVE	X0= HUMAIN X1= NON_RESTREINT
prétendre	1	X0 V X1	X0=GN X1=COMPLETIVE INFINITIVE	X0= HUMAIN X1= NON_RESTREINT
interdire	1	X0 V X1 PREP2 X2	X0=GN X1=GN X2=GN INFINITIVE	X0= HUMAIN X1= HUMAIN X2= NON_RESTREINT

Les descripteurs de conditions ne désambiguïsent pas toujours une forme prédicative donnée. Les propriétés combinatoires des emplois prédicatifs peuvent contribuer à la levée d'ambiguïté. Elles sont décrites dans des grammaires locales.

La macrostructure des dictionnaires du type ETHU_DIC est constituée des noms relatifs à des humains. Il peut s'agir d'individus (*responsable*) ou de collectivités (*équipe*). Les unités monolexicales (*violoniste*) et les unités polylexicales (*chef* d'orchestre) sont décrites dans les mêmes termes). La microstructure comporte la vedette et trois sortes de descripteurs sémantiques: l'hyperclasse (HUMAIN), la classe et la sous-classe.

aaliyah,.N+H_HUMAIN+C_HUMAIN_CELEBRITE+SC_MUSICIEN

ac/dc,.N+H_HUMAIN+C_HUMAIN_CELEBRITE+SC_FORMATION_MUSICALE

adam brody,.N+H_HUMAIN+C_HUMAIN_CELEBRITE+SC_COMEDIEN

adam lambert,.N+H_HUMAIN+C_HUMAIN_CELEBRITE+SC_MUSICIEN+SC_COMEDIEN

adam sandler,.N+H_HUMAIN+C_HUMAIN_CELEBRITE+SC_MUSICIEN+SC_COMEDIEN

2.3. Les grammaires locales

Une grammaire locale décrit le contexte d'une unité lexicale donnée en tant qu'ensemble de configurations de mots, *cf*. GROSS M. 1995, MAUREL 1993.

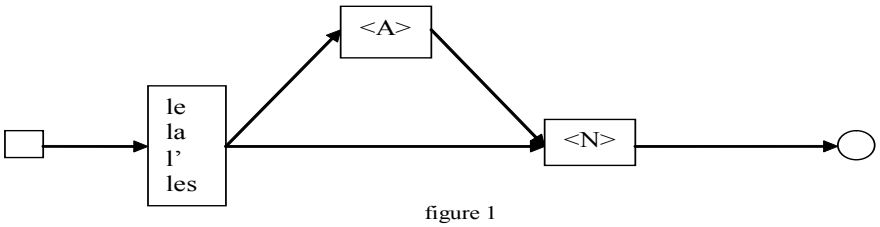

figure 1

La grammaire locale rend compte des groupes nominaux constitués soit d'un article défini suivi d'un nom soit d'un article défini suivi d'un adjectif puis d'un nom. Elle s'applique aux séquences *la voiture* et *la belle voiture* dans la mesure où les séquences correspondent à la configuration de mots qu'elle décrit. En revanche, elle ne s'applique pas aux groupes nominaux *la voiture rouge* et *une belle voiture* dans la mesure où ils ne correspondent pas à cette configuration de mots.

Une grammaire locale est représentée par un graphe comportant : un nœud initial, un nœud final et un ensemble de nœuds intermédiaires ; des arcs qui relient les nœuds en fonction des configurations de mots de la grammaire locale. Les nœuds intermédiaires sont associés à des mots, à des lemmes, à des catégories grammaticales ou à d'autres sortes d'informations métalinguistiques

Un automate à états finis est un outil informatique qui permet de définir une grammaire locale et d'analyser une séquence de mots. Un automate à états finis permet de décider si la séquence de mots analysée correspond à l'une des configurations de la grammaire locale.

Un transducteur à états finis est un automate à états finis qui permet d'associer une nouvelle information à de l'information reconnue. Il permet également de remplacer les items d'une séquence reconnue par d'autres items[2].

Qu'il s'agisse d'automates ou de transducteurs, les informations métalinguistiques enregistrées dans les graphes au niveau des nœuds sont celles qui sont encodées dans les dictionnaires électroniques associés au graphe.

3. Premiers résultats du traitement informatique

Nous terminons en présentant des citations étiquetées dans un corpus journalistique. Il s'agit d'un résultat provisoire. Nous avons encore de nombreux cas de silence à traiter et quelques cas de bruit. Lorsque les résultats seront stabilisés, nous procéderons à une évaluation globale de l'étiquetage. Pour l'instant, les premières évaluations que nous avons effectuées valident

2 La première fonctionnalité est fondamentale pour l'étiquetage. La seconde fonctionnalité permet toutes sortes de manipulations des textes. Par exemple, on peut utiliser des transducteurs pour faire des représentations métalinguistiques, remplacer n'importe quelle phrase par sa construction canonique, etc.

l'intérêt d'exploiter des ressources linguistiques de qualité pour développer un analyseur sémantique performant.

«(Quant à prétendre que l'on aurait, depuis Paris, influencé la police new-yorkaise, c'est du délire !) [CITATION] « a-t-(il) [GN_HUMAIN] ajouté.

(AlphaValue) [ORGANISME] recommande (d'`acheter` concernant Bouygues et fixe l'objectif de cours à 41,) [PROPOS] 3 euros, rappelant que le premier trimestre s'est déroulé conformément aux attentes.

(Le bureau d'études) GN_HUMAIN estime (que l'avenir de Bouygues dans le secteur des télécoms dépendra de sa capacité à conserver une marge supérieure à 20% au cours des années à venir et à limiter ses dépenses d'investissemen) [CITATION], avec l'arrivée de la 4G et d'un nouveau concurrent jugé (jugé) `très novateur'.

Il existe actuellement «(énormément d'initiatives disparates) [CITATION]», relève-t-(il) GN_ HUMAIN.

(Amazon) [ORGANISME] rappelle (les dates d'anniversaire des proches) [CITATION] et suggère (des idées de cadeaux) [CITATION] suivant leurs centres d'intérêt.

Bibliographie

Buvet P.-A. (2001) « Représentations métalinguistiques de phrases simples à l'aide de transducteurs », pp. 86-99.*Revue Informatique et Statistique dans les Sciences Humaines* 36, Liège :CIPL

Buvet P-A. (2009a) « Quelles procédures d'étiquetage pour la gestion de l'information textuelle électronique ? », *L'information grammaticale*, 122, Louvain, Peeters, pp. 40-48.

Buvet P.-A. (2009b) « Des mots aux emplois : la représentation lexicographique des prédicats », *Le Français Moderne,* 77, 1, Paris, CILF, pp. 83-96.

Buvet P.-A. (2011) « Catégorisation sémantico-énonciative du lexique à partir d'un dictionnaire électronique »,in *Os di.ci.o.nă.rios Fontes, métotodos e novas technologias*, Instituto de Letras da Universidade Federal da Bahia.

Charaudeau P. (1992) *Grammaire du sens et de l'expression*, Hachette

Charolles M.(1997) *L'encadrement du discours : univers, champs, domaines et espaces*, Cahier de Recherche Linguistique, LANDISCO, URA-CNRS 1035 Université Nancy 2, n° 6, 1-73.

Eskhol I. (2002) *Typologie sémantique des prédicats de parole*, Thèse de doctorat en Sciences du Langage, Université Paris 13

Fairon C. (2000) « Structures non-connexes. Grammaire des incises en français: description linguistique et outils informatiques », in *Language Arts & Disciplines*, John Benjamins.

Giguet E. et N. Lucas (2004) « La détection automatique des citations et des locuteurs dans les textes informatifs » in *Le discours rapporté dans tous ses états : Question de frontières,* J. M. López-Muñoz S. Marnette, L. Rosier, (eds.). Paris, l'Harmattan, pp. 410-418.

Gross G. et M. Prandi (2004) *La finalité : fondements conceptuels et genèse linguistique*, Duculot, Louvain-la-Neuve.

Gross M. (1995)1 « Une grammaire locale de l'expression des sentiments », *Langue française*, 105, Larousse, Paris, pp. 70-87.

Kerbrat-Orecchioni, C. (1980) *L'Énonciation*, *De la subjectivité dans le langage* ,Armand Colin

Kleiber G. (1990) *Les démonstratifs de près et de loin. Recueil d'études sur les expressions démonstratives*, Publication du Groupe Anaphore et Deixis, n° 2, Strasbourg, Université des Sciences Humaines (219 p).

Mathieu-Colas M. (2009), « Morfetik : une ressource lexicale pour le TAL », *Cahiers de lexicologie*, n° 94, pp. 137-146.

Mathieu-Colas M et P. Buvet, E. Cartier, F.Issac, Y. Madiouni, S. Mejri (2009), « Morfetik, ressource lexicale pour le TAL » *Actes du colloque TALN'09* (http://www-lipn.univ-paris13.fr/taln09/pdf/TALN_26.pdf).

Maurel D. (1993) « Reconnaissance automatique d'un groupe nominal prépositionnel. Exemple des adverbes de date », *Lexique*, 11, Presses Universitaires de Lille, pp. 147-181.

Mejri S. (2009), « Le mot, problématique théorique » *Le Français Moderne* 77 (1), pp. 68-82.

Mitkov R. ed. (2003) *The Oxford Handbook of Computational Linguistics*, Oxford University Press.

Moeschler J. (1989), « Modélisation du dialogue. Représentation de l'inférence argumentative » Ed. Hermes.

Reboula A. et J Moeschler (1998) *La pragmatique aujourd'hui*, Seuil, Paris.

Riegel M. J.-C. Pellat et R. Rioul 1994, *Grammaire méthodique du français*, Paris, Presses Universitaires de France.

Rosier L. (2008) *Le discours rapporté en français*, Ophrys.

Sagot B.et L. Danlos (2010), « Verbes de citation et Tables du Lexique-Grammaire » in *Proceedings of the 29th International Conference on Lexis and Grammar*, Belgrade, Serbie.

Todorov T. (1970) « Problèmes de l'énonciation », *Langages* 17.

Tromeur L. (2011 *Mise en place d'une interface en langue naturelle pour la plateforme Ontomantics*, Thèse de doctorat en Sciences du Langage, Université Paris 13.