Résolution de métonymie des entités nommées : proposition d'une méthode hybride

Caroline Brun* — **Maud Ehrmann*** — **Guillaume Jacquet***

* *Xerox Research Center Europe - XRCE*
6, chemin de Maupertuis
38240 Meylan
{Caroline.Brun, Maud.Ehrmann, Guillaume.Jacquet}@xrce.xerox.com

RÉSUMÉ. *La résolution de métonymie des entités nommées constitue un réel enjeu pour le traitement automatique des langues et bénéficie depuis peu d'un intérêt grandissant. Dans cet article, nous décrivons la méthode que nous avons développée pour la résolution de métonymie des entités nommées dans le cadre de la compétition SemEval 2007. Afin de résoudre les métonymies sur les noms de lieux et noms d'organisations, tel que requis pour cette tâche, nous avons mis au point un système hybride fondé sur l'utilisation d'un analyseur syntaxique robuste combiné avec une méthode d'analyse distributionnelle. Nous décrivons cette méthode ainsi que les résultats obtenus par le système dans le cadre de la compétition SemEval 2007.*

ABSTRACT. *Named Entity metonymy resolution is a challenging natural langage processing task, which has been recently subject to a growing interest. In this paper, we describe the method we have developed in order to solve Named entity metonymy in the framework of the SemEval 2007 competition. In order to perform Named Entity metonymy resolution on location names and company names, as required for this task, we developed a hybrid system based on the use of a robust parser that extracts deep syntactic relations combined with a non supervised distributional approach, also relying on the relations extracted by the parser. We describe this methodology as well as the results obtained at SemEval 2007*

MOTS-CLÉS : *entités nommées, métonymie, méthode hybride, analyse syntaxique robuste, approche distributionnelle.*

KEYWORDS: *Named Entities, metonymy, hybrid method, robust parsing, distributional approach.*

1. Introduction

Le traitement des entités nommées fait aujourd'hui figure d'incontournable en traitement automatique des langues (TAL). Apparue au milieu des années 1990 à la faveur des dernières conférences MUC (*Message Understanding Conferences*, (MUC-6, 1995 ; MUC-7, 1998)), la tâche de reconnaissance et de catégorisation des noms de personnes, de lieux, d'organisations, etc. apparaît en effet comme fondamentale pour diverses applications participant de l'analyse de contenu. Nombreux sont les travaux se consacrant à sa mise en œuvre, obtenant des résultats plus qu'honorables (F-mesure[1] dépassant généralement 0,9), et ce pour diverses langues. Fort de ce succès, le traitement des entités nommées s'oriente désormais vers de nouvelles perspectives avec, entre autres, la catégorisation fine (Ehrmann et Jacquet, 2006), la normalisation et la désambiguïsation. En effet, à l'instar des autres unités lexicales, plusieurs phénomènes de glissement ou de superposition de sens peuvent avoir lieu au regard des entités nommées et il importe de pouvoir les résoudre afin de traiter au mieux ces unités. Nous nous intéressons plus particulièrement à la métonymie des entités nommées et présentons dans cet article une méthode hybride pour la résolution de ce type particulier de polysémie.

Il conviendra tout d'abord de préciser le cadre général de ce travail avec, et ce sera l'objet de la première section, la définition et la caractérisation de la métonymie des entités nommées, la considération des enjeux de son traitement en TAL ainsi qu'un tour d'horizon des travaux existants. La section suivante présentera rapidement la campagne d'évaluation SemEval (*Semantic Evaluation*), à l'occasion de laquelle les travaux ici présentés ont été réalisés et évalués. La troisième section s'attachera à décrire en détail le système mis au point, reposant sur la combinaison d'un composant symbolique et d'un composant distributionnel. Enfin, la dernière section rendra compte des résultats obtenus lors de l'évaluation.

2. La métonymie des entités nommées

Il importe de revenir sur la définition linguistique de la métonymie avant de considérer les enjeux de son traitement en TAL.

2.1. *Caractérisation linguistique*

La métonymie est, aux côtés de la métaphore, une opération linguistique qui autorise l'emploi d'une entité pour en représenter une autre et qui, de ce fait, conduit à l'émergence de phénomènes de polysémies pour les unités lexicales. Elle correspond plus exactement au fait d'employer un mot (par exemple, le mot *récolte*) attaché à

1. Moyenne harmonique de la précision et du rappel : (2 × précision × rappel) / (précision + rappel).

une certaine entité (l'action de récolter) pour en désigner une autre (les produits recueillis), la seconde étant liée à la première par une relation de type partie-tout ou une relation fonctionnelle (ici relation de cause à effet). L'usage de tournures métonymiques peut soit faire état d'une dimension créative en étant lié à un contexte très particulier (*L'omelette au jambon est partie sans payer*, (Fauconnier, 1984)) soit, au contraire, faire état d'un usage considéré comme courant (*Je suis garé au bout de la rue, allons boire un verre*). Le premier cas correspond à de la métonymie « vive », créée sur l'instant par le locuteur en fonction de la situation d'énonciation et le second à de la métonymie « systématique », due à un processus lexical productif généralisé valable pour un certain nombre d'unités lexicales partageant des caractéristiques sémantiques. La métonymie donne ainsi lieu à la création de nombreux sens nouveaux, au départ créations éphémères étroitement dépendantes d'une situation d'énonciation mais dont certaines, bien souvent, deviennent d'usage courant, sans même que l'on se souvienne du glissement de sens dont elles sont issues : de la métonymie discursive à la métonymie lexicalisée, opère ainsi le « passage dans la langue ».

S'il n'est pas possible de prévoir les changements de sens en étroite dépendance avec le contexte, il est en revanche possible de concevoir des règles permettant de rendre compte des possibilités de métonymie « systématique » pour certaines unités lexicales. Parmi les cadres théoriques élaborés pour analyser et expliquer ce type de phénomène, il est possible de citer, entre autres, les travaux de G. Nunberg avec une analyse en termes de changement de référent puis en termes de changement de prédicat (Nunberg, 1995 ; Kleiber, 1999), ceux de J. Pustejovsky avec le *lexique génératif* (Pustejovsky, 1995 ; Kleiber, 1999), ainsi que ceux de G. Kleiber avec ce qu'il appelle le principe de métonymie intégrée (Kleiber, 1999). Ainsi, s'il s'agit bien d'une classe ouverte permettant un nombre indéfini de glissements de sens, il existe des changements de sens réguliers ou systématiques, au regard notamment des entités nommées. Examinons les exemples suivants :

*La politique américaine est plombée par l'**Irak**.*
*La **France** a gagné en demi-finale.*

Dans la première phrase, il n'est bien sûr pas question du pays proprement dit mais de l'événement qui s'y déroule, tout comme dans la seconde où il s'agit non pas de la France en tant que telle mais d'une équipe sportive française. Cet usage des unités *Irak* et *France* en tant qu'événement et équipe sportive respectivement est possible pour d'autres noms de pays dans des situations similaires. Il existe bien d'autres exemples possibles, sur lesquels nous reviendrons, mais il est d'ores et déjà possible de postuler une certaine régularité et productivité des phénomènes de métonymie pour les entités nommées.

Outre ces caractéristiques, des études conduites par K. Markert, U. Hahn et M. Nissim (Markert et Hahn, 2002 ; Markert et Nissim, 2006) ont fait état de la fréquence de ce phénomène, montrant que 17 % de l'ensemble des occurrences dans un corpus de 27 magazines allemands étaient métonymiques, tout comme 20 % des occurrences des noms de pays et 30 % de celles des noms d'organisations sur des extraits signi-

ficatifs du *British National Corpus*. Régulière, productive et fréquente, la métonymie des entités nommées représente ainsi un réel enjeu pour le traitement automatique des langues.

2.2. *Enjeux et moyens pour le TAL*

2.2.1. *TAL et métonymie*

Le traitement de la métonymie lexicale, tout comme celui de la métonymie des entités nommées, peut améliorer nombre de traitements, parmi lesquels les tâches d'extraction d'information, de résolution de coréférence et de questions-réponses. S'agissant des entités nommées, il peut être utile, par exemple, de distinguer l'entité *France* en tant que « unité géographique » de celle en tant que « équipe sportive » ou encore de celle en tant que « gouvernement » dans un texte ou un ensemble de textes, afin d'obtenir des résultats plus fins en cas de recherche d'information sur l'un ou l'autre de ces référents. Le repérage de glissements métonymiques peut également aider à la résolution de coréférence, comme dans l'exemple suivant emprunté à (Markert et Nissim, 2006) :

China *has agreed to let a United Nations [...] But it was unclear whether* ***Beijing*** *would meet past UN demands for unrestricted acess to [...]*

où le fait de savoir que *China* et *Beijing* renvoient tous deux au gouvernement chinois permet d'établir une coréférence entre ces deux unités. Enfin, la résolution de métonymie peut améliorer les systèmes de questions-réponses, comme le montre D. Stallard (Stallard, 1993), cité également par (Markert et Nissim, 2006), rapportant une amélioration de 27 % des résultats grâce à la prise en compte de la métonymie pour un système ayant à répondre à des questions du type « *Which airlines fly from Boston to Denver ?* ».

Si la résolution de métonymie constitue ainsi un réel enjeu pour le TAL, l'analyse du phénomène a cependant, jusqu'à ces dernières années, manqué d'« envergure », de même que les moyens disponibles pour mettre en œuvre son traitement ont fait défaut. C'est en effet le constat opéré par (Markert et Nissim, 2006) qui, sans remettre en cause les nombreux travaux linguistiques sur le sujet, font toutefois remarquer que les exemples de métonymies sur lesquels ils se fondent sont généralement imaginés pour l'occasion et avancés en dehors de tout contexte, sans laisser de place à l'hésitation quant à leur interprétation. Cependant, ces exemples ne reflètent pas la réalité du phénomène de métonymie tel qu'il existe dans les textes, avec ses nombreux cas de figure et leur distribution (en terme de fréquence), réalité dont le TAL doit pourtant être informé, compte tenu des traitements sur le langage naturel, bien souvent à grande échelle, qu'il est amené à implémenter. Au-delà du manque d'une caractérisation à visée applicative indispensable au TAL, les auteurs soulignent également l'insuffisance de ressources dédiées à ce phénomène. Noms propres et sens métonymiques sont en effet bien souvent absents ou peu nombreux dans la plupart des ressources lexicales actuellement disponibles en TAL, ressources à partir desquelles est par ailleurs ef-

fectuée l'annotation sémantique de corpus de référence (WordNet pour SenSeval-II, SenSeval-III), ces derniers étant par conséquent inexploitables pour le traitement de la métonymie[2]. K. Markert et M. Nissim imputent, à juste titre nous semble-t-il, à ce manque de caratérisation « à l'échelle » et de ressources l'absence de véritables travaux et évaluations de traitement automatique de la langue portant sur la résolution de métonymie.

2.2.2. *Les recherches de K. Markert et M. Nissim*

Ce constat établi, (Markert et Nissim, 2006) se sont appliquées à conduire des études en corpus sur la métonymie de certaines catégories sémantiques afin de mieux les caractériser et de procéder à leur annotation dédiée en corpus. Il convient de souligner, à l'instar de (Poibeau, 2006), l'importance de ces travaux pour la compréhension et le traitement de la métonymie des entités nommées en TAL.

K. Markert et M. Nissim ont tout d'abord commencé par définir une série de principes pour la construction de schémas d'annotation de métonymies. Parmi ces derniers, figurent des recommandations à caractère technique (encoder le texte en *XML* pour assurer l'indépendance de la plate-forme) tout comme d'autres plus linguistiques, telles que, entre autres, la prise en compte de textes relevant de domaines et de genre différents (pour assurer la couverture de divers types de métonymies), l'utilisation de classes sémantiques et de patrons métonymiques[3] pour définir les catégories d'annotation ou encore la prévision d'une catégorie d'annotation pour les cas de métonymies non conventionnelles. Se conformant à ce cadre de travail, les auteurs ont ensuite déterminé un schéma d'annotation pour les métonymies en général, avant d'en élaborer deux plus précis, pour les classes sémantiques LOCATION et ORGANISATION. Le cadre général prévoit de faire la distinction entre trois types d'interprétations : interprétation littérale (*literal reading*), interprétation métonymique (*metonymic reading*), pour les cas réguliers comme pour les cas non réguliers de métonymies, et interprétation mixte (*mixed reading*) pour les cas faisant état de deux lectures métonymiques différentes ou conjointement d'une lecture métonymique et littérale. Des schémas plus précis d'annotation (Markert et Nissim, 2005) ont ensuite été réalisés à l'aide d'indications figurant dans la littérature linguistique sur ce sujet ainsi qu'à l'aide d'études en corpus. Nous reviendrons plus précisément sur ces schémas par la suite. Une fois établies ces instructions, les auteurs ont annoté des corpus, pour la classe LOCATION (Markert et Nissim, 2002b) et la classe ORGANISATION (Markert et Nissim, 2006).

Ces travaux ont permis de dégager certains points essentiels à la compréhension et au traitement de la métonymie des entités nommées en TAL. L'étude en corpus a tout d'abord révélé l'existence d'autres patrons métonymiques que ceux traditionnel-

2. Il est à noter que le recensement de glissements métonymiques réguliers ne fait pas partie de l'objectif initial de ce genre de ressources ; notre propos ici est de souligner l'absence de ressources dédiées à ce phénomène.

3. Un patron métonymique permet de rendre compte du glissement de sens opéré entre le référent premier et le référent cible ; ces patrons sont de type `org-for-product` ou encore `place-for-people` et peuvent servir pour l'annotation.

lement reconnus, tel celui `org-for-index` pour les occurrences d'organisations en tant qu'indice boursier, ou encore `org-for-event` (*cf.* section 3 pour des exemples). Le travail d'annotation a ensuite permis de mieux caractériser la distribution des cas métonymiques pour les différentes classes sémantiques. Les taux satisfaisants d'accord interannotateurs ont également démontré la fiabilité d'annotation pour certains cas, comme la difficulté pour d'autres (interprétation mixte notamment). Enfin, deux corpus, dont les noms de pays et d'organisations employés métonymiquement ont été annotés, sont désormais disponibles. La stabilité des patrons métonymique dépend probablement du type et du domaine des corpus. Cependant, sur des corpus de type journalistique, les patrons identifiés par K. Markert et M. Nissim se révèlent parfaitement stables. Fortes de ces travaux et recherches, K. Markert et M. Nissim ont proposé une tâche d'évaluation sur la résolution de métonymie lors de l'édition 2007 de la campagne SemEval. Avant de présenter cette dernière, nous souhaitons rendre compte rapidement des travaux existants sur la résolution de métonymie des entités nommées.

2.3. *Travaux existants*

Comme nous venons de le voir, la tâche de résolution de métonymie des entités nommées émerge depuis peu à la faveur des recherches de M. Nissim et K. Markert. Il n'existe, par conséquent, que quelques travaux s'intéressant au traitement de ce phénomène, parmi lesquels, naturellement, ceux de M. Nissim et K. Markert. Ces dernières, après avoir défini un cadre de travail (Markert et Nissim, 2002b ; Markert et Nissim, 2006) ont en effet pu se pencher sur le problème de la résolution effective des glissements de sens pour les entités nommées, commençant tout d'abord par rapprocher cette tâche de celle de désambiguïsation lexicale (*word sense disambiguation*) (Markert et Nissim, 2002a). Tout comme il importe de choisir, pour un mot donné, un sens précis parmi un ensemble de sens possibles (désambiguïsation lexicale), il importe de choisir, pour un mot appartenant à une classe sémantique, un changement de sens métonymique (ou patron métonymique) précis parmi un ensemble de partons métonymiques possibles ; l'objet de la désambiguïsation n'est plus un mot mais une classe sémantique. Ayant établi cela, M. Nissim et K. Markert ont dès lors expérimenté des méthodes de désambiguïsation lexicale pour la métonymie des entités nommées, utilisant des algorithmes d'apprentissage supervisés avec des traits dont elles avaient auparavant étudié la pertinence (Markert, 2000), pour les noms de lieux (Nissim et Markert, 2003) puis pour les noms d'entreprises (Markert et Nissim, 2005), obtenant à chaque fois des résultats prometteurs. Ces travaux ont permis de mettre en valeur l'importance du contexte avec le rôle des traits grammaticaux et de montrer leur possible généralisation au travers de contextes similaires (exploitation d'un thésaurus construit à partir de mesures de similarité entre mots). Les autres travaux existants sur le sujet sont également à base d'apprentissage. Dans la lignée des recherches de M. Nissim et K. Markert, Y. Peirsman a testé, pour les noms de lieux, des algorithmes supervisés et non supervisés (*memory-based learning*), examinant la convenance de différents traits (Peirsman, 2006). Enfin, T. Poibeau s'est lui aussi attelé à la résolution de métonymie des entités nommées (Poibeau, 2006), dans un cadre d'annotation différent et pour le

français, s'appuyant sur des calculs de probabilités évaluant le pouvoir discriminant de tel ou tel trait pour les noms de lieux. À la suite de ce rapide état de l'art, il est temps de présenter la campagne SemEval.

3. Résolution de métonymie pour la campagne SemEval

La campagne SemEval 2007 proposait 18 tâches d'évaluation autour de problèmes sémantiques tels que la désambiguïsation de prépositions, l'annotation d'expressions et de relations temporelles (*TempEval*) ou encore la désambiguïsation des noms de personnes sur le Web (*Web People Search*). La tâche proposée par K. Markert et M. Nissim[4] est une tâche lexicale sur l'anglais, portant plus précisément sur la résolution de métonymie pour deux classes sémantiques, la classe LOCATION avec des noms de pays et la classe ORGANISATION avec des noms d'entreprises. L'objectif pour les participants est de classer automatiquement des occurences présélectionnées et en contexte de noms de pays et d'entreprises, et ce en fonction de leur interprétation littérale ou non littérale. Cette première alternative correspond à l'annotation « gros grain » ou *coarse-grained annotation*. Deux autres niveaux d'annotation sont possibles : un niveau « moyen » (*medium*) pour lequel il faut faire la distinction entre des interprétations littérales, métonymiques et mixtes et, enfin, un niveau « fin » (*fine*), pour lequel il importe de préciser, en cas d'interprétation métonymique, le patron métonymique dont il est question. À titre d'illustration, nous pouvons reprendre les exemples donnés par les organisatrices dans le document décrivant la tâche (Markert et Nissim, 2007) :

*At the time of **Vietnam**, increased spending led to inflation.*
***BMW** slipped 4p to 31p.*
*The **BMW** slowed down.*

Pour la première phrase, le nom *Vietnam* ne renvoie pas au pays mais à la guerre qui s'y est déroulée, il convient donc d'annoter cette entité comme `non-literal` (niveau 1), comme `metonymic` (niveau 2) ou comme `place-for-event` (niveau 3). De même, les occurrences de *BMW* dans les exemples suivants ne renvoient pas à l'entreprise mais à l'action de l'entreprise pour la première (`org-for-index`) et au produit de cette entreprise pour la seconde (`org-for-product`). Les trois niveaux d'annotation pour les noms d'entreprises et pour les noms de pays sont représentés ci-après dans le tableau 1[5] .

Voici une liste d'exemples illustrant ces différents patrons métonymiques, fournis dans le guide d'annotation de la compétition.

4. http ://www.comp.leeds.ac.uk/markert/MetoSemeval2007.html
5. Davantage de précisions sur les catégories d'annotation (ou patrons métonymiques) sont disponibles dans (Markert et Nissim, 2007).

Catégorie : ORGANISATION	*coarse*	*medium*	*fine*
	`literal`	`literal`	`literal`
	`non-literal`	`mixed`	`mixed`
		`metonymic`	`othermet`
			`object-for-name`
			`object-for-representation`
			`organisation-for-members`
			`organisation-for-event`
			`organisation-for-product`
			`organisation-for-facility`
			`organisation-for-index`
Catégorie : LOCATION	*coarse*	*medium*	*fine*
	`literal`	`literal`	`literal`
	`non-literal`	`mixed`	`mixed`
		`metonymic`	`othermet`
			`object-for-name`
			`object-for-representation`
			`place-for-people`
			`place-for-event`
			`place-for-product`

Tableau 1. *Niveaux de granularité et catégories d'annotation pour les classes* ORGANISATION *et* LOCATION

Pour les noms de lieux :

1) Annotation « `Literal` », désignant une interprétation purement locative (a), ou géopolitique (b) :

- (a) *Coral Coast of* ***Papua New Guinea***;

- (b) ***Britain****'s current account deficit.*

2) Annotation « `Metonymic` » :

- « `place-for-people` », lorsqu'un nom de lieu désigne une personne ou une organisation qui lui est associée : ***England*** *lost the semi-final ;*

- « `place-for-event` », lorsqu'un nom de lieu désigne un evènement qui s'y est déroulé : ***Vietnam*** *haunted him ;*

- « `place-for-product` », lorsqu'un nom de lieu désigne le produit qui y est développé : *A smooth* ***Bordeaux*** *that was gusty enough ;*

- « `object-for-name` », lorsqu'un nom de lieu est en usage autonyme : *Guyana (formerly British* ***Guyana****) ;*

- « `object-for-representation` », lorsqu'un nom réfère à une représentation (image, photo, etc.) : *This is* ***Malta****;*

- « `othermet` », pour tous les cas non déjà répertoriés :*The bottom end is very* ***New York****/New Jersey and the top is very melodic.*

3) « `Mixed` », dans les cas où le contexte déclenche des interprétations différentes pour une même occurence : *They arrived in* ***Nigeria*** *, hitherto a leading critic of [...].* (`literal` et `place-for-people`).

Pour les noms d'organisations :

1) Annotation « `Literal` », faisant référence à l'organisation comme entité légale en général :

- ***NATO*** *countries ;*

- ***Sun*** *acquired that part of East-man Co's subsidiary.*

2) Annotation « `Metonymic` » :

- « `organisation-for-members` », lorsqu'un nom d'organisation réfère aux membres qui la compose : *Last February,* ***IBM*** *announced [...] ;*

- `organization-for-event`, lorsqu'un nom d'organisation réfère à un événement associé :*[ ] in the aftermath of* ***Westland****;*

- « `organisation-for-product` », lorsqu'un nom d'organisation réfère à ses produits : *A red light was hung on the* ***Ford****'s tail gate ;*

- « `organization-for-facility` », le nom de l'organisation désigne le bâtiment qui l'héberge : *The opening of a* ***MacDonald****'s is a major event ;*

- « `organisation-for-index` », le nom est utilisé pour désigner l'indice boursier correspondant : ***BMW*** *slipped 4p to 3p ;*

- « `object-for-name` », le nom est en usage autonyme : ***Chevrolet*** *is feminine ;*

- « `objet-for-representation` », le nom réfère à une représentation, par exemple un logo : *Graphically, it lacked what kings call the word class of* ***Shell****;*

- « `othermet` », pour tous les cas non déjà répertoriés : *Funds had been paid into* ***Barclays Bank****.*

3) « `Mixed` », dans les cas où le contexte déclenche des interprétations différentes pour une même occurence : ***Barclays****, slipped 4p to 351p after confirming 3,000 job losses.* (`org-for-index` et `org-for-members`).

La répartition des différents phénomènes illustrés ci-dessus sur le corpus d'entraînement[6] est présentée dans le tableau 2. Les cas litéraux sont largement représentés (63,3 % pour les organisations et 79,7 % pour les lieux[7]), les cas de métonymies les plus fréquents sont « `organization-for-members` », (20,2 %), et

6. Les participants à la campagne ont disposé d'un corpus d'entraînement puis d'un corpus de test, tous deux issus du *British National Corpus* ; ces corpus comportaient respectivement 925 et 928 noms de pays et 1 090 et 842 noms d'organisations.

7. Cette proportion correspond à la « baseline », c'est-à-dire au taux de réussite minimal atteignable grâce à l'assignation de l'annotation `literal` à toutes les entités à annoter.

« `place-for-people` », (17,4 %). On peut remarquer que la classe ORGANISATION est plus riche « métonymiquement » que la classe LOCATION ; elle comprend en effet plus de patrons spécifiques et, si la fréquence d'apparition de certains est parfois relativement faible, ils sont tout de même présents. C'est en fonction de l'ensemble de ces catégories que nous avons tenté de mettre au point un système automatique de résolution de métonymie.

LOCATION			ORGANISATION		
Reading	N	%	Reading	N	%
`literal`	737	79,7	`literal`	690	63,3
`place-for-people`	161	17,4	`org-for-members`	220	20,2
`place-for-event`	3	0,3	`org-for-event`	2	0,2
`place-for-product`	0	0,0	`org-for-product`	74	6,8
			`org-for-facility`	15	1,4
			`org-for-index`	7	0,6
`object-for-name`	0	0,0	`object-for-name`	8	0,7
`object-for-representation`	0	0,0	`object-for-representation`	1	0,1
`othermet`	9	1,0	`othermet`	14	1,3
`mixed`	15	1,6	`mixed`	59	5,4
total	925	100,0	total	1 090	100,0

Tableau 2. *Distribution des cas de métonymies pour les classes* LOCATION *et* ORGANISATION

4. Un système de résolution de métonymie pour les EN

Notre participation à la tâche de résolution de métonymie pour les noms de pays et d'entreprises (Brun *et al.*, 2007) a consisté en l'élaboration d'un système automatique hybride reposant sur la combinaison d'un composant symbolique et d'un composant distributionnel. Nous présenterons successivement ces deux composants.

4.1. *Composant symbolique*

4.1.1. *Analyse syntaxique robuste avec XIP*

L'élément fondamental sur lequel repose notre approche est l'analyseur syntaxique robuste XIP (Aït-Mokhtar et Chanod, 1997 ; Aït-Mokhtar *et al.*, 2002). *Xerox Incremental Parser* prend en entrée du texte tout-venant ou des documents au format XML et produit en sortie de façon robuste une analyse syntaxique profonde. À partir d'un ensemble de règles, l'analyseur est capable de faire de la désambiguïsation de catégories, de construire des syntagmes noyaux et d'extraire des relations de dépendances syntaxiques entre unités lexicales simples et/ou complexes. En plus de l'analyse des relations syntaxiques de surface, une couche supplémentaire de développement a été

rajoutée à la grammaire de base, réalisant une analyse syntaxique dite « profonde » ou « normalisée », l'objectif applicatif visé étant l'extraction d'information. Ces développements permettent d'avoir, après l'analyse, une représentation commune pour des suites de signifiants qui ne sont pas identiques mais qui véhiculent une information similaire. À l'heure actuelle, ce travail de normalisation s'effectue selon trois axes :

– l'exploitation des relations syntaxiques mises en évidence lors de l'analyse générale : l'analyse par la grammaire générale est tout d'abord raffinée afin de considérer les sujets et objets de verbes non finis et les antécédents des relatives dans le calcul du sujet et de l'objet, de normaliser la forme passive en forme active et de typer certains compléments. Ensuite, certaines alternances verbales telles qu'elles sont définies dans (Levin, 1993) sont exploitées, ainsi que quelques éléments de Framenet (Ruppenhofer *et al.*, 2006) ;

– la hiérarchisation des propositions dans une phrase : la grammaire normalisée permet de reconnaître les degrés d'enchâssement des verbes par rapport au verbe principal ;

– l'exploitation d'information de morphologie dérivationnelle : cette information permet d'exprimer des équivalences entre verbe-compléments et nom-arguments.

Cet analyseur intègre également un module de reconnaissance des entités nommées (Brun et Hagège, 2004), prenant en compte les types les plus « classiques » des entités nommées, à savoir les expressions numériques, les monnaies, les dates, ainsi que les noms de lieux, de personnes, et d'organisations. Il s'agit d'un module à base de règles, consistant en un ensemble de règles locales ordonnées utilisant des informations lexicales et des informations contextuelles concernant les parties du discours, les formes lemmatisées et un ensemble de traits lexico-sémantiques. Ces règles détectent la séquence de mots représentant l'entité nommée, et assignent un trait sémantique (date, organisation, lieu, personne, etc.) au nœud parent de cette séquence, généralement un nom. Dans la chaîne de traitements, ces règles sont appliquées en amont du composant syntaxique, avant l'extraction des syntagmes noyaux (textitchunks) et l'extraction des dépendances. Le système a été évalué en interne pour l'anglais, sur un corpus de dépêches d'environ 87 000 mots, et montre une f-mesure de 0,9 tous types d'entités confondus (Ehrmann, 2004).

Ci-dessous un exemple de sortie XIP (avec les EN, l'arbre des textitchunks et les relations de dépendance, fournis par la grammaire de normalisation) :

Iran wanted the two centers to generate part of its electricity needs.

```
TOP { SC{ NP{Iran} FV{wanted}} NP{the two centers} IV{to generate} NP{part}
PP{of NP{its electricity needs}} .}

MOD_PRE (needs,electricity)      SUBJ-N (generate,centers)
LOCATION (Iran)                  EXPERIENCER-PRE(wanted,Iran)
CONTENT(wanted,centers)          OBJ-N(generate,needs)
```

Le module de reconnaissance des entités nommées intégré à XIP ne permet pas le

traitement de la métonymie. Il a, par conséquent, fallu procéder à son adaptation dont la description est l'objet du paragraphe suivant.

4.1.2. *Adaptation à la tâche de résolution de métonymie*

Nous avons étudié les corpus d'entraînement fournis par les organisatrices pour cette tâche à l'aide de XIP, en tenant compte des directives d'annotation. Pour les noms de pays et d'entreprises, notre attention s'est focalisée sur les types de relations grammaticales dans lesquelles étaient impliquées les unités à étudier et sur les informations lexicales ou sémantiques attachées aux arguments de ces relations. Pour chaque patron métonymique, nous avons donc analysé l'ensemble de ses occurrences (attachées à une unité lexicale) dans le corpus d'entraînement pour dégager des configurations grammaticales et lexicales jouant le rôle d'« amorces » d'interprétations métonymiques.

Au terme de cette étude de corpus, il fut possible de prendre en compte l'ensemble des indices récoltés et d'élaborer un module de résolution de métonymie dans XIP. Intervenant à la fin de la chaîne de traitements de l'analyseur, cette adaptation a consisté en l'écriture de règles traduisant les configurations discriminantes relevées pour tel ou tel patron d'une part, et en l'ajout de lexique d'autre part. À titre d'exemple, prenons l'hypothèse suivante : « *Si un nom de pays est sujet d'un verbe renvoyant à une action économique, alors le patron* `place-for-people` *doit s'appliquer.* » ; cette hypothèse se retrouve dans XIP sous la forme de règle :

```
if (LOCATION(#1) & SUBJ-N (#2[v_econ],#1))
PLACE-FOR-PEOPLE (#1)
```

Cette règle se lit ainsi : si le parseur a détecté un nom de pays (#1) qui est le sujet d'un verbe (#2) portant le trait « `v_econ` », alors il crée un relation unaire PLACE-FOR-PEOPLE pour le nom de pays.

En sus des indices récoltés lors de l'étude de corpus, nous avons exploité des informations lexicales déjà codées dans XIP, comme par exemple les traits attachés aux verbes de communication (*say, deny, comment)* et les catégories relevant du cadre « *experiencer* » de Framenet, à savoir des verbes tels que *feel, sense, see*, les noms *despair, compassion, adoration* ou les adjectifs *synpathetic, sad,* etc. En effet, eu égard au fait que ce cadre « *experiencer* » renvoie à des personnes ou à des groupes de personnes, dès lors qu'un nom de pays ou d'entreprise a ce rôle, il peut être annoté en tant que `place-for-people` ou `organisation-for-members`. Ci-dessous deux exemples de sorties de l'analyseur intégrant les nouveaux développements :

Malta *endorsed a serie of proposals.*

```
PLACE-FOR-PEOPLE(Malta)        PREP_OF(serie,proposals)
SUBJ-N_PRE(endorsed,Malta)     OBJ-N(endorsed,serie)
DETD(serie,a)                  LOCORG(Malta)
```

Dans cet exemple, la relation sujet entre le nom de pays et le verbe *to endorse* permet

d'orienter l'analyse vers le patron PLACE-FOR-PEOPLE, l'action d'*appuyer, donner son aval à* renvoyant à des personnes associées à Malte.

*It was the largest **Fiat** everyone had ever seen.*

ORG-FOR-PRODUCT (`Fiat`)	MOD_PRE (`seen,ever`)
SUBJ-N_PRE (`was,It`)	ATTRIB(`It,Fiat`)
EXPERIENCER_PRE(`seen,everyone`)	QUALIF(`Fiat,largest`)

Pour ce cas, la présence de la relation QUALIF entre le nom d'entreprise *Fiat* et l'adjectif *largest*, renforcée par la présence d'un article défini, conduit à l'annotation en tant que ORG-FOR-PRODUCT. Ce composant symbolique constitue le premier volet de notre méthode de résolution de métonymie, il est complété par un composant distributionnel, qu'il convient à présent de détailler.

4.2. *Composant distributionnel*

Intervient en « deuxième passe » de notre système un composant distributionnel. L'idée principale de cette combinaison est d'exploiter deux méthodes relevant d'approches différentes mais complémentaires, autrement dit de pallier les manques de l'analyse symbolique, focalisée sur des données précises nécessitant une étude minutieuse, par une analyse distributionnelle, apte à récolter des informations à grande échelle sur d'importantes données textuelles. Le principe est donc, lorsqu'une unité n'a pu être traitée par le composant symbolique, d'essayer de trouver son annotation en exploitant les informations présentes à propos de cette unité ou d'une unité de même type dans un grand corpus. Nous présentons rapidement l'analyse distributionnelle avant de détailler sa mise en œuvre pour la résolution de métonymie.

4.2.1. *L'analyse distributionnelle*

La notion d'analyse distributionnelle a été introduite par Z. S. Harris. En linguistique de corpus, cette méthode est aujourd'hui largement exploitée, notamment dans les travaux de terminologie, de structuration de terminologie et de construction d'ontologies (Faure et Nedellec, 1999 ; Assadi, 1998 ; Bourigault, 2002). L'hypothèse est la suivante : il serait possible, à partir de régularités syntaxiques observées pour un ensemble de mots, de déduire des propriétés sémantiques pour ces mots. Concrètement, il s'agit d'étudier la distribution des mots, c'est-à-dire les contextes lexico-syntaxiques dans lesquels ils apparaissent, pour ensuite tenter de dégager des parentés sémantiques. S'inscrivant dans ce cadre, G. Jacquet et F. Venant ont élaboré une méthode automatique de désambiguïsation du sens d'un mot en contexte, reposant sur la prise en compte de l'influence des éléments syntaxiques et lexicaux présents dans l'énoncé (Jacquet et Venant, 2003). Ainsi, ils ne cherchent plus seulement à créer des classes de mots relevant du même champ sémantique mais des classes de mots dont le comportement sémantique influence de la même façon le sens des autres mots de la phrase. Le composant distributionnel élaboré pour la résolution de métonymie s'inscrit dans la perspective de ces travaux.

4.2.2. *Méthode distributionnelle pour la résolution de métonymie*

L'objectif ici est de rapprocher des contextes et d'exploiter les résultats du composant symbolique. Cette méthode comporte deux processus. Il s'agit d'une part de construire un espace distributionnel pour être en mesure de rapprocher des contextes en fonction d'une entité donnée et, d'autre part, de capitaliser l'information du composant symbolique sous la forme d'une sorte de « base de données » de contextes avec annotation. Nous détaillons ces deux processus successivement.

4.2.3. *Construction de l'espace distributionnel et rapprochement des contextes*

Ce premier processus comporte cinq étapes. L'objectif est d'être en mesure, pour une entité donnée apparaissant dans un contexte donné, d'établir une liste des contextes les plus proches. Le point de départ est le corpus BNC dans son entier (100 millions de mots), duquel ont été extraits les corpus d'entraînement et de test de la tâche de résolution de métonymie.

La première étape correspond à l'analyse syntaxique de ce corpus (à l'aide de l'analyseur XIP) afin d'en extraire des dépendances syntaxiques. C'est à partir de ces dépendances que sont construits les contextes et les unités lexicales. Prenons un exemple, avec la proposition *provide Albania with food aid*. Les dépendances extraites par XIP sont les suivantes :

IND-OBJ-N[8](VERB :`provide`,NOUN :`Albania`)
PREP_WITH(VERB :`provide`,NOUN :`aid`)
PREP_WITH(VERB :`provide`,NOUN :`food aid`)

On peut voir que les arguments des dépendances sont de simples unités lexicales (*aid*) ou des syntagmes (*food aid*).

La deuxième étape de ce processus est la construction d'un espace distributionnel à partir de ces dépendances. Cet espace distributionnel est l'inverse de celui habituellement construit en analyse distributionnelle : puisque l'objectif est de rapprocher des contextes, chaque point de l'espace est un contexte syntaxique et chaque dimension est une unité lexicale. Cette inversion constitue une première différence avec l'approche de (Jacquet et Venant, 2003). Chaque dépendance obtenue lors de l'étape précédente permet de construire plusieurs contextes simples et/ou composés. Un contexte syntaxique comporte une relation et une unité lexicale. La méthode prévoit également de construire des contextes composés (autre différence par rapport à (Jacquet et Venant, 2003)), autrement dit des contextes combinant plusieurs contextes, dont le premier comporte nécessairement un syntagme verbal et le second a pour recteur ce même syntagme. Pour la phrase analysée ci-dessus, les unités lexicales, les contextes simples (1. pour recteur et 2. pour régi) et les contextes composés sont les suivants :

8. la relation *ind-obj-n* correspond à la relation syntaxique *objet indirect*.

UNITÉS LEXICALES :	CONTEXTES SIMPLES :
VERB :provide	1.VERB :provide.IND-OBJ-N
NOUN :Albania	1.VERB :provide.PREP_WITH
NOUN :aid	2.NOUN :Albania.IND-OBJ-N
NOUN :food aid	2.NOUN :aid.PREP_WITH

CONTEXTES COMPOSÉS :
1.VERB :provide.IND-OBJ-N + 2.NOUN : aid.PREP_WITH
1.VERB :provide.IND-OBJ-N + 2.NOUN :aid.PREP_WITH
1.VERB :provide.IND-OBJ-N + 2.NOUN :food aid.PREP_WITH
1.VERB :provide.PREP_WITH + 2.NOUN :Albania.IND-OBJ-N

Une heuristique permet de filtrer ces données en fonction de leur productivité : chaque unité lexicale doit être présente au moins 100 fois dans le corpus, tout comme les contextes (y compris les contextes composés). Avec le corpus BNC de 100 millions de mots, on obtient au final 60 849 unités lexicales et 140 634 contextes. Il est donc possible de construire un espace distributionnel comportant 140 634 points (les contextes) et 60 849 dimensions (les unités lexicales). Cet espace est le matériau de base à partir duquel les autres traitements viennent s'effectuer.

À partir de **la troisième étape** intervient la prise en compte d'une unité lexicale précise, pour laquelle le composant symbolique n'a pu trouver d'annotation. En fonction de cette unité et de son contexte d'apparition (soit telle qu'elle apparaît dans un extrait de SemEval), il s'agit tout d'abord de construire un sous-espace, de la manière suivante :

Pour un couple donné formé d'un contexte *i* et d'une unité lexicale *j* :

– le sous-espace des contextes équivaut à la liste des contextes occurrant avec l'unité lexicale *j*. S'il y a plus de *k* contextes, alors on ne garde que ces *k* contextes les plus fréquents ;

– le sous-espace des dimensions équivaut à la liste des unités lexicales occurrant avec au moins un des contextes du sous-espace des contextes. S'il y a plus de *n* unités, alors on ne garde que ces *n* unités les plus fréquentes.

La quatrième étape consiste à réduire les dimensions de ce sous-espace, à l'aide d'une Analyse Factorielle des Correspondances (équivaut à une Analyse en Composantes Principales avec la métrique du Chi2).

Enfin, **la cinquième étape** s'attache à rapprocher les contextes restants, c'est-à-dire ceux ayant passé le filtre des deux étapes précédentes. Ce rapprochement est calculé à l'aide de la distance euclidienne. On obtient ainsi, au final, une liste des contextes proches de celui de l'unité considérée. Pour l'unité *Albania* dans le contexte

provide, les contextes les plus proches sont présentés dans la première colonne du tableau 3 (à gauche).

Pour pouvoir utiliser cette liste de contextes, encore faut-il savoir à quelles annotations métonymiques ils peuvent correspondre. Intervient alors le second processus.

4.2.4. *Capitalisation de l'information du composant symbolique*

Ce second processus a pour objectif de construire une sorte de base de données de contextes spécifique à chaque annotation. Deux étapes sont nécessaires. Le corpus BNC est analysé avec XIP, augmenté cette fois-ci du module de résolution de métonymie. Cette analyse permet d'obtenir des dépendances syntaxiques mettant en jeu des noms de pays et des noms d'entreprises avec leurs annotations métonymiques (application du module de résolution de métonymie sur les entités nommées reconnues par l'analyseur, indépendamment des données de SemEval). Le résultat de cette première étape est donc une série de contextes impliquant des unités lexicales avec leur annotation. La deuxième étape correspond à la sélection de contextes discriminants au regard des annotations. Par exemple, si le contexte VERB :`allow`.IND-OBJ-N se retrouve pour 10 % des cas avec une unité comportant l'annotation `literal` et pour 90 % des autres avec une unité comportant l'annotation `place-for-people`, alors le contexte est considéré comme discriminant vis-à-vis de cette dernière annotation. En revanche, si un contexte se trouve dans 50 % des cas avec telle annotation et 50 % des cas avec telle autre, alors il n'est pas conservé. Ces contextes discriminants sont le reflet des règles symboliques décrivant les « configurations discriminantes » (*cf.* 4.1.2). Ils peuvent aussi correspondre à des configurations non identifiées par le composant symbolique mais que l'on retrouve de fait dans le corpus. Illustrons ces points par un exemple. Dans la phrase *Sa belle Renault a percuté le mur*, si le composant symbolique reconnaît ici un glissement métonymique grâce à la présence d'un adjectif (*belle*) et qu'il en est de même pour de nombreux autres cas d'emplois métonymiques de noms d'organisations, alors ce contexte devient discriminant pour le patron `org-for-product` et vient augmenter la base de données de contextes spécifiques à chaque annotation. Un autre moyen d'alimenter cette base est de considérer d'autres contextes présents dans la phrase, mais non utilisés par le composant symbolique : dans notre exemple, le contexte composé « sujet de *percuter le mur* » peut également être considéré comme discriminant pour le patron métonymique en question, et peut par conséquent être intégré à la base de contextes discriminants pour cette annotation. À l'issue de cette sélection de contextes discriminants, on dispose alors d'un stock de contextes avec leurs annotations, pouvant être exploité en parallèle avec le processus précédent.

4.2.5. *Annotation d'une unité*

Le premier processus (4.2.3) permet de déterminer une liste de contextes plus ou moins proches du contexte dans lequel apparaît une unité lexicale non annotée. Le second (4.2.4) permet de collecter un certain nombre de contextes avec leur annotation.

Ces deux types de données peuvent dès lors être croisés pour permettre l'annotation d'une unité.

Le tableau 3 présente la liste de contextes proches du contexte VERB :provide.IND-OBJ-N pour la proposition *provide Albania with food aid*, avec l'indication de leur distance et de leur annotation correspondante dans la base de données de contextes.

Contexte	Distance	Annotation
VERB :provide.IND-OBJ-N	0,00	
VERB :allow.OBJ-N	0,76	PLACE-FOR-PEOPLE
VERB :include.OBJ-N	0,96	
ADJ :new.MOD_PRE	1,02	
VERB :be.SUBJ-N	1,43	
VERB :supply.SUBJ-N_PRE	1,47	LITERAL
VERB :become.SUBJ-N_PRE	1,64	
VERB :come.SUBJ-N_PRE	1,69	
VERB :support.SUBJ-N_PRE	1,70	PLACE-FOR-PEOPLE
...	...	...

Tableau 3. *Listes des contextes les plus proches de* VERB :provide.IND-OBJ-N *avec leur annotation*

Annotation	Score
PLACE-FOR-PEOPLE	3,11
LITERAL	1,23
PLACE-FOR-EVENT	0,00
...	...

Tableau 4. *Scores des annotations pour les contextes les plus proches de* VERB :provide.IND-OBJ-N

Pour pouvoir déterminer une annotation pour le contexte provide.IND-OBJ-N, il faut regarder les annotations attribuées à ses contextes les plus proches et examiner si elles sont pertinentes ou non. Pour ce faire, il importe de calculer un score pour chaque annotation, valorisant sa présence dans les contextes de « tête de liste » : le score d'une annotation donnée est égal à l'inverse de la somme des distances des contextes portant cette annotation. Pour notre cas, les scores sont présentés dans le tableau 4 ; il est alors possible, en fonction de ces scores, d'attribuer l'annotation PLACE-FOR-PEOPLE à l'unité *Albania* dans la proposition *provide Albania with food aid*.

Les moyens employés pour ce composant peuvent paraître disproportionnés par rapport à la tâche à accomplir, et ce d'autant plus si l'on replace ce travail dans le cadre applicatif de l'extraction d'information qui nécessite une optimisation du temps de traitement. Nous ne nions pas cette limite, notons que certaines étapes coûteuses en

temps, telles que l'analyse syntaxique et la construction de l'espace distributionnel, peuvent se faire en amont d'une application utilisateur. Si ce travail n' en est qu'au stade des premières expérimentations et qu'aucune optimisation du temps de calcul n'a été effectuée, l'annotation d'une nouvelle entité dans un nouveau contexte prend moins d'une seconde (CPU 3GHz).

5. Évaluation

Dans cette section nous présentons et analysons les résultats obtenus lors de l'évaluation avant d'examiner plus avant la contribution de chacun des composants au sein de notre système.

5.1. *Présentation et analyse des résultats*

Les résultats obtenus avec notre système sur les corpus de test furent relativement encourageants car au-dessus de la baseline. Au milieu de quatre autres systèmes, XRCE-M est en effet parvenu à se placer en deuxième position pour la première tâche (noms de pays) à tous les niveaux de granularité, et en troisième position pour la seconde (noms d'entreprises). De plus, parmis les trois meilleurs systèmes, XRCE-M est le seul à utiliser un analyseur syntaxique plutôt que les traits syntaxiques annotés manuellement par les organisatrices pour chaque entité à annoter. Les taux de précision[9] pour l'ensemble des participants sont présentés ci-après, pour les noms de lieux (tableau 5) et pour les noms d'entreprises (tableau 6).

Systèmes / Tâche	Baseline	up13	FUH	UTD-HLT-CG	XRCE-M	GYDER
LOCATION-coarse	0,794	0,754	0,778	0,841	**0,851**	0,852
LOCATION-medium	0,794	0,750	0,772	0,840	**0,848**	0,848
LOCATION-fine	0,794	0,741	0,759	0,822	**0,841**	0,844

Tableau 5. *Taux de précision pour tous les systèmes pour la classe* LOCATION

Nous revenons tout d'abord sur les résultats de notre système avant de considérer l'ensemble de ceux obtenus pour la tâche, globalement, et par système.

9. Il n'est pas utile de présenter le rappel pour cette évaluation concernant les différents niveaux d'annotation (*coarse*, etc.) : les occurrences à annoter étant indiquées, il est systématiquement égal à 1. Au niveau plus précis de l'évaluation de la reconnaissance de patrons métonymiques (annotation *fine*), il est en revanche intéressant de considérer le rappel et d'observer la couverture des systèmes pour chaque patron. Ces résultats sont présentés dans les tableaux 7 et 8 pour notre système. Pour obtenir plus de détails concernant ce niveau d'évaluation pour les autres participants, nous invitons à consulter les articles suivants : (Poibeau, 2007 ; Leveling, 2007 ; Farkas *et al.*, 2007 ; Nicolae *et al.*, 2007).

Tâche \ Systèmes	Baseline	XRCE-M	UTD-HLT-CG	GYDER
ORGANISATION-coarse	0,618	**0,732**	0,739	0,767
ORGANISATION-medium	0,618	**0,711**	0,711	0,733
ORGANISATION-fine	0,618	**0,700**	0,711	0,728

Tableau 6. *Taux de précision pour tous les systèmes pour la classe* ORGANISATION

Satisfaisants dans l'ensemble, les résultats obtenus par XRCE-M sur les corpus de test sont néanmoins inférieurs à ceux obtenus sur les corpus d'entraînement, montrant par là les possibilités d'amélioration. Comme cela était prévisible, en raison des baselines et de la diversité des patrons métonymiques à prendre en compte pour chacune des classes, les résultats sont meilleurs pour les noms de lieux (0,841 %) que pour les noms d'entreprises (0,700 %). La différence accrue de résultats entre les divers niveaux de granularité pour les noms d'entreprises (perte de 30 centièmes entre *coarse* et *fine*) par rapport aux noms de pays (perte de 10 centièmes seulement) vient par ailleurs confirmer cette caractéristique de plus grande diversité, et donc de plus grande difficulté. Nous donnons dans les tableaux 7 et 8 les résultats détaillés de notre système pour chaque patron dans chacune des classes.

LOCATION-fine \ Scores	Nb occurrences	Précision	Rappel	F-mesure
`literal`	721	0,867	0,960	0,911
`place-for-people`	141	0,651	0,490	0,559
`place-for-event`	10	0,5	0,1	0,166
`place-for-product`	1	-	0	0
`object-for-name`	4	1	0,5	0,666
`object-for-representation`	0	-	-	-
`othermet`	11	-	0	0
`mixed`	20	-	0	0

Tableau 7. XRCE-M *: résultats détaillés pour la classe* LOCATION

Pour la classe LOCATION, on peut remarquer de relativement bonnes précisions pour les patrons `place-for-people`, `place-for-event` mais également `object-for-name`, peu présent dans le corpus d'entraînement mais relativement simple à schématiser à partir de l'information et des exemples disponibles dans les directives d'annotation. Les autres patrons, non couverts par notre méthode car trop peu nombreux, voire absents, dans le corpus d'entraînement et surtout difficiles à configurer (`mixed` et `othermet` entre autres) ont des résultats nuls.

Scores / ORGANISATION-fine	Nb occurrences	Précision	Rappel	F-mesure
literal	520	0,730	0,906	0,808
organisation-for-members	161	0,622	0,522	0,568
organisation-for-event	1	-	0	0
organisation-for-product	67	0,550	0,418	0,475
organisation-for-facility	16	0,5	0,125	0,2
organisation-for-index	3	-	0	0
object-for-name	6	1	0,666	0,8
othermet	8	-	0	0
mixed	60	-	0	0

Tableau 8. XRCE-M *: résultats détaillés pour la classe* ORGANISATION

Les résultats pour la classe ORGANISATION présentent sensiblement la même répartition, entre des patrons bien couverts (`organisation-for-members`, `organisation-for-product` et `object-for-name`) et d'autres non traités.

Certaines annotations erronées sont dues à des erreurs de l'analyseur syntaxique utilisé en amont de notre système. Dans certains cas en effet, la qualité de l'analyse syntaxique fait défaut, comme dans la phrase suivante : « *Many galleries in the States, England and France declined the invitation* », où l'entité *France*, en raison d'une mauvaise analyse de la coordination, est analysée comme étant le sujet du verbe *declined*, et donc comme portant l'annotation `place-for-people`. Malgré ces problèmes liés à l'analyse syntaxique, XRCE-M est relativement bien placé au sein des autres systèmes ayant participé à l'évaluation (*cf.* tableaux 5 et 6). Ainsi, sans utiliser l'annotation syntaxique manuelle, XRCE-M n'a qu'une entité d'écart pour la classe LOCATION par rapport au système le plus performant GYDER.

Cinq systèmes ont participé à la tâche sur les LOCATION et trois à celle sur les ORGANISATION, pour tous les niveaux de granularité à chaque fois. Hormis XRCE-M, les systèmes reposaient sur des méthodes statistiques par apprentissage, exploitant divers algorithmes et un échantillon de traits différents. Deux systèmes ne sont pas parvenus à atteindre la baseline pour les noms de lieux (Poibeau, 2007 ; Leveling, 2007), montrant par là la difficulté de la tâche. Ces derniers n'ont cependant exploité que des traits de surface sans utiliser de ressources externes, contrairement aux deux autres systèmes (Farkas *et al.*, 2007 ; Nicolae *et al.*, 2007) qui, pour leur part, ont utilisé des traits syntaxiques et sémantiques et fait usage de ressouces externes (WordNet) et/ou du Web, avec au final de meilleurs résultats, attestant de la pertinence de ce type de traits pour le traitement de la métonymie. Pour l'ensemble des systèmes, seulement quelques catégories ont pu être annotées avec succès, telles celles `org-for-members`, `org-for-products`, `place-for-people` et `literal`. Les patrons les plus rares n'ont pas été reconnus pour certains, et mal annotés pour d'autres. Le patron `mixed`,

d'emblée reconnu comme délicat et difficile par les organisatrices, ne fut traité avec succès que par le système GYDER et pour quelques cas seulement, ce fait confirmant l'absence de configuration contextuelle stable et productive pour ce type de glissement métonymique. Seul le patron `object-for-name` semble faire exception, avec un traitement relativement réussi par les systèmes UTD-HLT CG et XRCE-M.

À la suite de cette présentation des résultats, nous souhaitons considérer d'un peu plus près notre système, en étudiant l'apport de chacun de ses composants.

5.2. *Examen approfondi du système : contribution de chaque composant*

L'objectif est ici d'évaluer, au sein de notre système, la contribution du module symbolique d'une part, et celle du module distributionnel d'autre part. Pour ce faire, nous avons appliqué séparément chaque module sur le corpus de test de SemEval. Les résultats sont présentés dans le tableau 9, avec « Xip » correspondant au système symbolique, « Distrib » au système distributionnel, et « XipDistrib » au système combinant les deux précédents. Il importe avant toute chose de souligner un point important : ces résultats correspondent à une exécution du système postérieure à notre soumisssion à SemEval. Ils reflètent les récentes évolutions de l'analyseur XIP et ne peuvent pas, par conséquent, être comparés aux résultats des autres participants lors de la campagne.

Systèmes / Tâche	Baseline	Xip	Distrib	XipDistrib
	Taux de précision par niveaux de granularité sur la classe LOCATION			
LOCATION-coarse	0,794	0,863	0,814	0,867
LOCATION-medium	0,794	0,860	0,813	0,863
LOCATION-fine	0,794	0,855	0,812	0,859
	Taux de précision par niveaux de granularité sur la classe ORGANISATION			
ORGANISATION-coarse	0,618	0,710	0,635	0,732
ORGANISATION-medium	0,618	0,699	0,626	0,711
ORGANISATION-fine	0,618	0,690	0,618	0,700

Tableau 9. *Contribution de chaque composant*

Le premier constat que l'on peut faire est que l'apport du composant distributionnel est plutôt faible, avec seulement 0,4 à 1 point de plus par rapport au composant symbolique seul. Même si ce chiffre déçoit nos attentes, il est possible de voir les choses sous un autre angle. Si l'on s'intéresse aux résultats obtenus par le composant distributionnel seul, on observe qu'ils sont légèrement au-dessus de la baseline mais loin derrière le composant symbolique seul. Pour la classe ORGANISATION-fine, le composant distributionnel est même égal à la baseline. Pourtant, pour cette même

classe, il permet d'augmenter de 1 point la précision du composant symbolique. Ces résultats montrent donc que le composant distributionnel ne suffit pas, à lui seul, à résoudre une tâche telle que la résolution de métonymie des entités nommées, mais qu'il peut compléter le module symbolique en traitant des entités non annotées par ce dernier. Ces observations nous encouragent à ne pas se satisfaire de ce trop faible apport du composant distributionnel. Nous restons convaincus que le principe d'hybridation symbolique/distributionnel mis en œuvre est pertinent, c'est pourquoi nous travaillons à l'amélioration du composant distributionnel afin d'améliorer sa contribution. De manière plus précise, nous développons un nouveau système permettant, d'une part, de rendre le composant distributionnel moins dépendant des règles symboliques construites spécifiquement pour la tâche et, d'autre part, d'améliorer le calcul des contextes les plus proches (*cf.* cinquième étape de 4.2.3) en remplaçant la distance euclidienne par l'appartenance ou non à un cluster (Ah-Pine et Jacquet, 2009). Nous travaillons actuellement à l'application de ce système à la résolution de cas métonymiques.

6. Conclusion

Dans cet article, nous avons décrit un système hybride pour la résolution de métonymie. Notre approche, quoique perfectible, a d'ores et déjà montré des résultats encourageants lors de sa participation à la compétition SemEval 2007. Le traitement de la métonymie des entités nommées peut se révéler très intéressant pour des systèmes d'extraction d'information ou de questions-réponses, ainsi que pour améliorer des composants de TAL tels que ceux dédiés à la résolution de coréference. Cette problématique participe d'une réflexion plus vaste sur le statut et le traitement des entités nommées (Ehrmann, 2008).

Les perspectives de développement sont nombreuses pour notre système tout comme pour la tâche de résolution de métonymie en général. Il serait particulièrement intéressant de poursuivre la réflexion amorcée par K. Markert et M. Nissim autour de cette problématique, en particulier en ce qui concerne l'annotation des phénomènes métonymiques, qui, si elle s'avère relativement consensuelle pour les glissements métonymiques les plus fréquents (par exemple `place-for-people` ou `org-for-members`, selon la typologie de SemEval), semble l'être nettement moins dans d'autres cas (nous pensons aux cas « mixed » ou encore « object-for-representation », « object-for-name » qui sont fort peu fréquents dans les corpus et particulièrement difficiles à appréhender, même pour un annotateur humain).

À notre échelle, nous envisageons de poursuivre le travail entrepris afin d'améliorer les performances du système pour les glissements de sens déjà pris en compte et d'étendre ce traitement à d'autres types d'entités nommées et d'autres langues. Du point de vue méthodologique, nous travaillons actuellement sur l'amélioration du composant distributionnel ainsi que sur l'application du principe d'hybridation symbolique/distributionnel à l'annotation classique et à la désambiguïsation des entités nommées (Ah-Pine et Jacquet, 2009).

7. Bibliographie

Ah-Pine J., Jacquet G., « Clique-Based Clustering for improving Named Entity Recognition Systems », *Acte de la conférence EACL 2009*, Athènes, Grèce, 2009.

Assadi H., Construction d'ontologies à partir de textes techniques. Application aux systèmes documentaires., PhD thesis, Université Paris VI, 1998.

Aït-Mokhtar S., Chanod J. P., « Incremental finite-state parsing », *Proceedings of Applied Natural Language Processing*, Washington, DC, 1997.

Aït-Mokhtar S., Chanod J. P., Roux C., « Robustness beyond shallowness : incremental dependency parsing », *NLE Journal*, 2002.

Bourigault D., « Upery : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus », *Actes de TALN 2002*, Nancy, 2002.

Brun C., Ehrmann M., Jacquet G., « XRCE-M : A hybrid system for named entity metonymy resolution », *4th International Workshop on Semantic Evaluations*, ACL-SemEval 2007, Prague, Juin, 2007.

Brun C., Hagège C., « Intertwining deep syntactic processing and named entity detection », *ESTAL*, Alicante, Espagne, 2004.

Ehrmann M., Évaluation d'un système d'extraction d'entités nommées, Technical report, DESS Texte, Nancy, 2004. Rapport de stage.

Ehrmann M., Les entités nommées, de la linguistique au TAL : statut théorique et méthodes de désambiguïsation, PhD thesis, Université Paris VII, 2008.

Ehrmann M., Jacquet G., « Vers une double annotation des entités nommées », *Traitement Automatique des Langues*, 2006. Disponible sur : `http://www.atala.org`.

Farkas R., Simon E., Szarvas G., Varga D., « GYDER : maxent metonymy resolution », *4th International Workshop on Semantic Evaluations*, ACL-SemEval 2007, Prague, Juin, 2007.

Fauconnier G., *Espaces mentaux, Aspects de la construction du sens dans les langues naturelles*, Minuit, 1984.

Faure D., Nedellec C., « Knowledge Acquisition of Predicate Argument Structures from technical texts using Machine Learning : the system Asium », *Poceeedings of EKAW'99*, Juan-les-Pins, France, 1999.

Jacquet G., Venant F., « Construction automatique de classes de sélection distributionnelle », *Actes de TALN 2003*, Dourdan, France, 2003.

Kleiber G., *Problèmes de sémantique. La polysémie en questions*, Presses Universitaires du Septentrion, 1999.

Leveling J., « FUH (FernUniversit¨at in Hagen) : Metonymy Recognition Using Different Kinds of Context for a Memory-Based Learner », *4th International Workshop on Semantic Evaluations*, ACL-SemEval 2007, Prague, Juin, 2007.

Levin B., *English Verb Classes and Alternations - A Preliminary Investigation*, The University of Chicago Press,, 1993.

Markert K., Features integration in metonymy resolution, Technical report, Université d'Edimburgh, 2000.

Markert K., Hahn U., « Understanding metonymies in discourse », *Artificial Intelligence*, vol. 135, n° 1/2, p. 145-198, 2002.

Markert K., Nissim M., « Metonymy Resolution as a Classification Task », *Proc. of the 2002 Conference on Empirical Methods in Natural Language Processing*, Philadelphia, Penn., 2002a.

Markert K., Nissim M., « Toward a corpus annotated for metonymies : the case of location names », *Proc. of the 3rd International Conference on Language Resources and Evaluations*, Las Palmas, Iles Canaries, 2002b.

Markert K., Nissim M., Annotation Scheme for metonymies AS1, Technical report, Université de Leeds, Université d'Edimburgh, 2005. Disponible sur : http ://www.comp.leeds.ac.uk/markert/Papers/index.html.

Markert K., Nissim M., « Metonymic proper names : a corpus based account », *in* A. Stefanowitsch, S. Gries (eds), *Corpus-based approaches to Metaphor and Metonymy*, Mouton de Gruyter, p. 152-174, 2006.

Markert K., Nissim M., « SemEval-2007 Task 08 : Metonymy Resolution at SemEval-2007 », *4th International Workshop on Semantic Evaluations*, ACL-SemEval 2007, Prague, Juin, 2007.

MUC-6, « Proceedings of the Sixth Message Understanding Conference », 1995. Informations disponibles à l'adresse suivante : `http://www.cs.nyu.edu/cs/faculty/grishman/muc6.html`.

MUC-7, « MUC-7. Proceedings of the Seventh Message Understanding Conference », 1998. Informations disponibles à l'adresse suivante : `http://www-nlpir.nist.gov/related\_projects/muc/`.

Nicolae C., Nicolae G., Harabagiu S., « UTD-HLT-CG : Semantic Architecture for Metonymy Resolution and Classification of Nominal Relations », *4th International Workshop on Semantic Evaluations*, ACL-SemEval 2007, Prague, Juin, 2007.

Nissim M., Markert K., « Syntactic features and word similarity for supervised metonymy resolution », *Proceedings of the 41st Annual Meeting of the Association of Computational Linguistics (ACL-03)*, Sapporo, Japon, 2003.

Nunberg G., « Transfers of meanings », *Journal of Semantics*, vol. 12, n° 2, p. 109-132, 1995. Disponible sur : `http://jos.oxfordjournals.org/cgi/content/abstract/12/2/109`.

Peirsman Y., « What's in a name ? Computational approaches to metonymical location names », *Proceedings of the Workshop on Making Sense of Sense : Bringing Psycholinguistics and Computational Linguistics Together*, Trento, Italie, 2006.

Poibeau T., « Dealing with Metonymic Readings of Named Entities », *Proceedings of the 28th Annual Conference of the Cognitive Science Society*, Canada, 2006.

Poibeau T., « up13 : Knowledge-poor Methods (Sometimes) Perform Poorly », *4th International Workshop on Semantic Evaluations*, ACL-SemEval 2007, Prague, Juin, 2007.

Pustejovsky J., *The generative lexicon*, The MIT Press, Cambridge, 1995.

Ruppenhofer J., Ellsworth M., Petruck M., Johnson C., Scheffczyk J., Framenet II : Extended Theory and Practice, Technical report, University of Berkeley, 2006. Disponible sur : `http://framenet.icsi.berkeley.edu/`.

Stallard D., « Two Kinds of Metonymy », *Proc. of the 31st Meeting of the Association for Computational Linguistics*, Colombus, Ohio, June, 1993.