Vers une double annotation des Entités Nommées

Maud Ehrmann— Guillaume Jacquet

Centre de Recherche Xerox de Grenoble
6, chemin de Maupertuis
38 240 Meylan
France

{Maud.Ehrmann, Guillaume.Jacquet}@xrce.xerox.com

RÉSUMÉ. *Ayant acquis cette dernière décennie une maturité certaine, la tâche de reconnaissance et catégorisation d'« Entités Nommées » (EN) s'oriente aujourd'hui vers de nouveaux défis : le typage plus fin et la désambiguïsation des entités. Cet article présente une méthode de double annotation d'entités nommées, combinant l'information issue d'une ressource lexico-sémantique construite automatiquement et celle provenant d'un système symbolique « classique » de reconnaissance d'EN. Permettant de spécifier davantage les référents des entités nommées par une annotation fine, cette approche constitue également une première étape de désambiguïsation de ces unités.*

ABSTRACT. *The Named Entity Recognition task has reached, this last decade, an undeniable maturity. Research on Named Entity (NE) is now taking up new challenges with fine-grained annotation and disambiguation of named entities. In this article we present a method for named entity double annotation, combining information from an automatically constructed (semantic) lexical resource that provides semantic label(s) for proper names, and information from a symbolic NE recognition system. This approach provides fine grained annotation of NE and is also a first step of named entity disambiguation.*

MOTS-CLÉS : *entités nommées, annotation sémantique fine, méthode hybride, extraction d'information, désambiguïsation.*

KEYWORDS: *named entities, fine grained semantic annotation, hybrid method, information extraction, disambiguation*

1. Introduction

Un des principaux enjeux actuels du Traitement Automatique du Langage Naturel (TALN) est de capter l'information portée par un texte et d'accéder à son sens. Ceci passe, entre autres, par le traitement des unités linguistiques à forte valeur informative (et/ou référentielle). Le cours de l'histoire, ou plutôt de la recherche, a voulu que l'on désigne un certain nombre d'entre elles sous le nom d' « entités nommées » (EN). Ces dernières correspondent traditionnellement à l'ensemble des noms propres présents dans un texte, qu'il s'agisse de noms de personnes, de lieux ou d'organisations, ensemble auquel sont souvent ajoutées d'autres expressions comme les dates, les unités monétaires, les pourcentages et autres. La reconnaissance et la catégorisation de ces entités apparaissent ainsi comme une tâche fondamentale pour diverses applications de TALN participant de l'analyse de contenu, à l'instar de la recherche et l'extraction d'information, le *question-answering* ou encore l'intelligence économique. Considérée à tort comme triviale, elle a fait cette dernière décennie l'objet d'une attention plus soutenue, consécutivement à la récente multiplication des demandes issues du secteur industriel, et à la définition de tâches spécifiques lors des dernières compétitions MUC (*Message Understanding Conference*) et ACE (*Automatic Content Extraction)*[1]. Les recherches menées jusqu'à ce jour ont ainsi permis l'élaboration de systèmes de reconnaissance d'EN relativement performants, permettant d'identifier dans des textes de nature journalistique des noms relevant de types généraux tels que « personne », « lieu » et « organisation », et ce avec un taux de F-mesure[2] dépassant généralement les 0,90.

Dans la droite ligne de ces réalisations, de nouvelles perspectives de recherche sont actuellement mises en œuvre, selon deux orientations. D'un point de vue théorique tout d'abord, une réflexion semble s'amorcer autour de la notion d'entités nommées, avec la question de leur définition. Répondant expressément à des besoins de TALN, force est de constater que ces unités linguistiques ne bénéficient, à l'heure actuelle, d'aucune véritable assise théorique dans la littérature. Il semblerait qu'il s'agisse d'un des premiers « retours à l'envoyeur » du TALN vis-à-vis de la théorie linguistique, amenée aujourd'hui à considérer un objet qu'elle n'avait nullement défini auparavant[3]. Compte tenu de la variété et de l'étendue (linguistiquement parlant) des unités pouvant faire partie de cet ensemble, un exercice de définition des EN n'est bien sûr pas des plus faciles. Considérer d'un point de vue « linguistique théorique » des critères et mécanismes distinctifs des EN et, partant, s'interroger sur leur statut référentiel (voir les

[1] MUC : http://www-nlpir.nist.gov/related_projects/muc/index.html
ACE : http://projects.ldc.upenn.edu/ace/.
[2] Taux combiné de Précision et de Rappel : (2*Précision*Rappel)/(précision + Rappel).
[3] Du moins en tant que tel, les noms propres étant un sujet largement traité par les linguistes.

travaux de T. Poibeau), constituent ainsi une perspective de recherche significative autour de la problématique des EN. La seconde orientation est quant à elle plus « pratique », touchant aux objectifs et méthodes de traitement automatique des EN. Il existe en effet une volonté d'améliorer et d'enrichir leur annotation avec, d'une part, l'annotation de nouveaux types d'entités (tel les noms de produits, voir (Nilsson *et al.*, 2005)) et, d'autre part, une annotation plus fine allant au-delà des catégories générales désormais aisément reconnues, c'est-à-dire une annotation indiquant précisément, du point de vue de la référence, les caractéristiques distinctives de l'entité considérée et permettant (si besoin) sa désambiguïsation (Sekine, 2004). Ces champs d'investigation autour des EN apparaissent bien sûr comme complémentaires, la définition précise de l'objet d'étude aidant à son traitement. Conscients des enjeux et intérêts d'une appréhension plus poussée de la notion d'entité nommée d'un point de vue purement théorique, notre propos ne traitera cependant pas de cette question : les travaux présentés ci-après s'inscrivent en effet dans la perspective dite « pratique » de traitement plus fin des EN.

Aussi, cherchant à aller au-delà des catégories générales d'annotation des EN et à améliorer leur traitement, nous présentons une méthode ayant pour objet l'annotation fine d'EN ainsi que la désambiguïsation de certaines de ces unités. L'approche proposée repose sur, d'une part, la construction dynamique d'une ressource lexico-sémantique dédiée aux EN, proposant pour ces dernières des étiquettes sémantiques fines et, d'autre part, la combinaison de cette ressource (permettant une annotation fine) avec un système standard de reconnaissance d'EN (proposant des étiquettes générales classiques) afin d'obtenir une annotation enrichie, ou double annotation, des EN. L'originalité de cette approche réside dans l'exploitation de relations syntaxiques profondes lors de la construction de la ressource d'EN, dans l'annotation à l'aide de groupes d'étiquettes plutôt qu'à l'aide d'étiquettes et dans la mise en œuvre d'une double annotation des EN offrant des informations de niveau tant général que particulier. Nos travaux ont été guidés par la volonté de prendre en compte l'extrême mouvance ou instabilité référentielle des EN ainsi que leur caractère polysémique.

Dans la partie liminaire, nous nous attacherons à illustrer les nouveaux besoins (et problèmes) en matière d'annotation des EN et à donner un aperçu général de l'approche. Nous évoquerons ensuite les travaux connexes à nos travaux, nous poursuivrons par une description plus précise de la méthode de construction de la ressource, pour ensuite présenter la combinaison permettant la double annotation. Enfin, nous terminerons par une évaluation de la ressource et de la méthode.

2. Traitement des entités nommées : une annotation enrichie pour répondre à de nouveaux besoins

2.1. *Des unités plus complexes qu'il n'y paraît...*

Au regard du traitement des EN, l'enjeu principal a été jusqu'à présent de les reconnaître et de les catégoriser selon des ensembles « conceptuels » relativement larges. Cette catégorisation sémantique grossière, si elle correspond bien entendu à une première étape indispensable dans des travaux de recherche, rend compte également d'une appréhension particulière de ces unités. Comprenant essentiellement des noms propres, les EN ont en effet été perçues, et ce à juste titre, comme des unités exclusivement référentielles et de ce fait se sont vu attribuer les mêmes caractéristiques que le nom propre. Souvent décrit en termes de « désignateur direct et rigide », ce dernier renvoie alors à un particulier sans l'intermédiaire d'un sens lexical codifié et désigne le même particulier dans tous les « mondes possibles »[4]. Ces caractéristiques semblent avoir encouragé l'assignation d'un type sémantique unique pour rendre compte du référent d'une entité nommée. Satisfaisante dans la plupart des cas et pour de nombreuses applications, cette approche traditionnelle peut cependant être enrichie : si les EN sont d'intelligence avec le nom propre dans le sens où elles sont éminemment référentielles, elles flirtent également avec les unités lexicales plus classiques du point de vue de leur comportement en langue, révélant une complexité jusqu'alors mise de côté lors de leur traitement automatique.

Les phénomènes de pluralité interprétative sont légion en langue naturelle, concernant de nombreux types d'unités linguistiques (noms, verbes, etc.) et jouant à tous les niveaux (morphosyntaxique, lexical, phrastique). Largement décrits et étudiés pour les unités lexicales classiques, les changements, transferts ou superpositions de sens le sont en revanche très peu pour les unités de type EN. Or celles-ci semblent au contraire être régies par les mêmes phénomènes. Considérons les énoncés suivants :

(1) *Orange a invité M. Dupont.*

(2) *Leclerc a fermé ses magasins en Rhône-Alpes.*

(3) *La France a signé le traité de Kyoto.*

Que décider quant à la catégorie des ces entités ? Est-il question de la ville d'*Orange* ou bien de la société de téléphonie ? De la personne *Michel Edouard Leclerc* ou de la chaîne de supermarché ? Faut-il préférer une annotation de *France* en tant qu' « organisation » ou « gouvernement » ou en tant que « lieu » ou « pays » ? De

[4] Jonasson expliquant Kleiber et Kripke, (Jonasson, 1994).

l'homonymie (1) à la polysémie (2), en passant par la métonymie (3), force est donc de constater que les EN n'échappent pas aux phénomènes d'ambiguïté et sont, à l'instar des autres unités lexicales, polyréférentielles. Parallèlement à ces exemples « linguistiques », la réalité du traitement automatique de l'information révèle peu ou prou la même chose. Une interrogation du moteur de recherche *Google* pour les entités *Orange* et *Leclerc* propose (sur les dix premières réponses) pour l'une, des renvois aux référents « opérateur » et « ville » et, pour l'autre, des renvois aux référents « centres d'achat », « général », « char », et « homme d'affaires ». Cette réalité des choses conduit à penser l'annotation des entités nommées en des termes différents et à mettre en œuvre de nouveaux traitements.

2.1. *... encourageant à de nouveaux traitements.*

La polysémie avérée des EN appelle en effet à abandonner l'approche catégorisante unilatérale pour préférer une annotation à caractère modulaire permettant une caractérisation plus fine et plus complète du référent dénoté par l'entité en contexte. Deux aspects semblent se dégager en considération de ces objectifs : il importe, d'une part, de proposer une annotation plus précise des EN et, d'autre part, de désambiguïser certaines entités ambiguës. Une annotation plus précise des EN consiste à leur associer une information sémantique fine, permettant de mieux circonscrire leur référent en contexte. Dans une phrase telle que :

> *Arnold Schwarzenegger s'est rallié le soutien de la majorité démocrate pour s'assurer un vote favorable*[5].

Il s'agit alors d'indiquer, au-delà de l'étiquette « personne », qu'il est question ici du « gouverneur républicain de Californie » et non du « bodybuildeur » ou de l' « acteur ». Dans le même esprit, il importe de distinguer, dans un texte ou un ensemble de textes, les occurrences de l'entité *Jacques Chirac* renvoyant au « président de la République » de celles renvoyant au « maire de Paris ». Cette spécification précise du référent auquel renvoie l'entité est à mettre en relation avec les notions de « facettes sémantiques » (Croft *et al.*, 2004) et de signification en contexte : de même que le sens d'une unité lexicale est activé différemment suivant son contexte d'apparition, le référent d'une entité nommée revêt ou non certaines caractéristiques. D'un point de vue pratique, cette annotation fine s'avère quelque peu différente de la catégorisation réalisée jusqu'à maintenant. S'il est relativement facile de penser une catégorisation sémantique large préalablement à l'annotation des EN, il semble plus périlleux de le faire pour des informations plus précises : en effet, les étiquettes mentionnées ci-avant, « *gouverneur*

[5] *Le Figaro,* 1er septembre 2006.

républicain de Californie, *bodybuildeur*, *acteur, président de la République* et *maire de Paris* » n'ont de toute évidence rien de prévisible[6]. Ce type d'information est en outre de nature essentiellement encyclopédique, ce qui pose le problème de son acquisition. De ce fait, certains « mécanismes » de l'approche traditionnelle que sont la catégorisation *a priori* et le recours à des lexiques exogènes, apparaissent comme compromis dans la mise en œuvre d'une annotation fine.

Le deuxième aspect mentionné au regard d'un traitement plus fin des EN est la désambiguïsation. L'énoncé (1) avec *Orange*, ville ou société, illustre bien le type de problème à résoudre, et il serait possible de multiplier à l'envi les exemples d'homonymie-polysémie d'EN (*Washington*-ville ou *Washington*-personne, *Vienne* en France ou *Vienne* en Autriche, *Bush*-père ou *Bush*-fils, etc.). Cet aspect est bien sûr très proche de la spécification du référent dont il a été question ci-avant et ces problématiques sont à penser comme complémentaires : pour désambiguïser une entité, encore faut-il savoir qu'elle est ambiguë, et donc avoir des informations précises sur son référent et les diverses facettes qu'il peut revêtir. Nous posons donc le problème du traitement des EN en ces termes : il importe de caractériser plus précisément le référent d'une EN en contexte afin, d'une part, d'apporter davantage d'information et, d'autre part, d'être en mesure de mettre en œuvre un processus de désambiguïsation si l'information précise mise à jour révèle une ambiguïté. Notons que certains cas d'ambiguïté peuvent s'avérer difficiles à résoudre si le contexte ne donne aucun indice, comme cela est le cas pour l'énoncé (1) ; le modèle de traitement fin des EN doit pouvoir rendre compte de ces cas d'ambiguïté. Il importe à cet endroit de pointer une spécificité de l'ambiguïté des EN : cette dernière est en effet « évolutive » dans le temps dans la mesure où une entité monoréférentielle aujourd'hui peut devenir polyréférentielle demain. Pour reprendre un exemple désormais bien connu, *Orange* est une société de téléphonie mobile depuis 1994 seulement, et les marques de produits, sociétés et autres célébrités ne manquent pas d'apparaître, de disparaître et de changer chaque jour. Aussi, l'espace des possibles au regard des sens, ou plutôt des référents, attachés à une entité semble relativement ouvert et peu contrôlable (ensemble fort productif du point du vue référentiel, mais sans règles de production du point de vue linguistique). En considération de ces nouveaux besoins en matière d'annotation des EN, nous proposons une méthode d'annotation fine d'EN. Un aperçu général est présenté ci-après, davantage de précisions seront données dans les parties suivantes.

---

[6] Ce caractère imprévisible des facettes des référents des EN fait écho à celui déjà pointé par (Croft *et al.*, 2004) à propos des facettes de sens des unités lexicales: « While only a few facets may be strongly favored by conventional constraints, in principle there is no reason to believe that the facets of a sense constitute a determinate set. » (p 117).

2.2. *Aperçu général de l'approche.*

La méthode que nous proposons repose sur la construction d'une ressource lexico-sémantique dédiée aux EN et sur la combinaison de cette dernière avec un système classique de reconnaissance d'EN. L'objectif est d'associer une information sémantique fine aux EN (construction de la ressource) tout en conservant le fruit des travaux réalisés jusqu'à aujourd'hui (combinaison avec un système classique) afin d'offrir une double annotation des entités nommées et de pouvoir désambiguïser certaines de ces unités.

La première étape correspond à la construction de la ressource d'EN[7]. Cette ressource permet d'associer, pour chaque entité, une ou plusieurs étiquette(s) sémantique(s) fine(s) rendant compte de certaines caractéristiques du ou des référents possibles de l'entité. En d'autres termes, elle permet, idéalement, d'associer à l'entité *Leclerc* les étiquettes « société, homme d'affaires, char, général ». Par son mode de construction, cette ressource pallie certains écueils de la recherche actuelle sur les EN tout en prenant en compte les particularités de ces unités. En effet, constituée automatiquement à partir de corpus, cette ressource peut tout d'abord être construite à moindre frais (or chacun connaît le coût de construction de lexiques spécialisés ou le coût de l'annotation manuelle d'un corpus d'apprentissage). Ensuite, son élaboration est réalisée de manière non supervisée : aucun type d'entité n'est visé particulièrement (or les lexiques concernant les EN ont été jusqu'à maintenant réservés à certains types, comme les lieux) et aucune catégorie n'est définie préalablement, ce qui confère un caractère « inventif » à la ressource (nous avons souligné ci-dessus le fait que le type d'information recherché aujourd'hui correspond à des étiquettes difficilement prévisibles). Elle permet ensuite de « suivre l'actualité » des EN, ces dernières pouvant être caractérisées de différentes manières en fonction de la période et du domaine du corpus. La construction de la ressource se fait pour chaque nouveau corpus, elle est ainsi exclusivement dépendante du corpus et c'est ce qui lui confère une caractéristique dynamique. Un processus incrémental de construction de la ressource serait bien sûr envisageable mais serait alors perdue l'idée d'adaptation en fonction du corpus, caractère important à nos yeux car offrant la possibilité de rendre compte avec exactitude des référents des entités telles qu'elles apparaissent dans un corpus[8].

---

[7] Cette ressource comprend majoritairement des noms de personnes, lieux ou organisation, c'est-à-dire des *noms propres*. Nous nous autorisons néanmoins à dire qu'il s'agit d'une ressource d'*entités nommées* dans la mesure où elle comprend des noms de produits et de marques, même si les dates et autres expressions de quantité ne sont pas présentes.

[8] Une alternative bienvenue serait d'associer aux étiquettes « récoltées » dans un corpus donné une information temporelle : l'évolutivité des référents des EN auraient alors un véritable ancrage (non plus palpable par le seul biais du corpus) et une construction incrémentale de la ressource serait

Au final, cette ressource fournit une représentation fidèle des EN présentes dans un corpus et a pour principal intérêt, au travers d'étiquettes sémantiques fines, de mettre en valeur les facettes sémantiques des entités et de *révéler* leur éventuelle polysémie. Il importe ici de spécifier la nature de l'information obtenue par le biais de ces étiquettes : si nous avons distingué ci-avant deux types d'amélioration pour le traitement des EN touchant, pour l'un, à davantage de précision quant au référent et, pour l'autre, à une désambiguïsation, notre approche n'est cependant pas en mesure de dire à quel niveau jouent les étiquettes de la ressource ni de caractériser le type d'ambiguïté révélée (homonymie, polysémie, métonymie), s'il est question d'ambiguïté. Autrement dit, par cette ressource nous ne visons pas un phénomène de sens lié aux EN en particulier : qu'il s'agisse d'homonymie, de polysémie ou de métonymie, notre objectif est de les révéler, sans nécessairement les caractériser.

La seconde étape revient à combiner l'annotation fine permise par la ressource avec l'annotation plus classique d'un système à base de règles, donnant ainsi lieu, lorsque cela est possible, à une double annotation des entités. Afin d'illustrer le gain de cette dernière, imaginons une application (relevant de l'extraction d'information) pour laquelle il importerait d'extraire d'un corpus toutes les personnes appartenant à l'armée. Un système classique de reconnaissance d'entités (qu'il soit symbolique ou à base d'apprentissage) serait capable de repérer dans le texte toutes les « personnes », parmi lesquelles des journalistes ou autres, non pertinentes du point de vue de l'application ; couplé à notre ressource, le système serait en mesure de pointer, par le biais de la double annotation, les « personnes » ayant aussi une étiquette « général, lieutenant, etc. », satisfaisant par là aux exigences de l'application. Cette combinaison de systèmes à grains différents a en outre permis une évaluation des résultats de la ressource.

L'objectif de cette approche est donc de fournir une double annotation des EN, de « poser » ou révéler la polysémie de certaines entités, et de faire quelques pas sur le chemin de la désambiguïsation. Avant de détailler plus avant la réalisation de cette double annotation, il convient d'évoquer les travaux relatifs à ce champ de recherche.

3. Travaux connexes

Les travaux relatifs à l'identification et l'annotation des entités nommées sont très nombreux et sont, pour l'essentiel fondés sur des méthodes par apprentissage exploitant des propriétés de type patrons morphosyntaxiques de surface. L'approche ici présentée différant quelque peu dans ses objectifs des systèmes de reconnaissance d'EN

---

alors envisageable. Ceci implique cependant une maîtrise fine des marqueurs temporels, des travaux dans ce sens sont actuellement en cours.

classiques[9], nous proposons de centrer ce tour d'horizon sur des travaux se situant dans la même veine que ce que nous proposons selon deux points de vue : l'annotation fine et la désambiguïsation des EN d'une part, et la construction de lexiques sémantiques d'autre part.

Bien que relativement récente, la tendance à considérer de plus près les EN a déjà suscité des travaux intéressants. Il est possible de distinguer trois types de travaux : ceux dont le but est de désambiguïser les EN, ceux cherchant à construire une ressource spécifique pour le traitement des EN, et enfin ceux combinant les deux précédents, c'est à dire cherchant à faire de la désambiguïsation tout en exploitant une ressource spécifique. Commençons par le premier type de travaux. La plupart d'entre eux reposent sur des méthodes d'apprentissage : à partir d'un corpus dans lequel les EN visées ont été annotées, l'objectif est d'« apprendre » des traits (ou caractéristiques) de nature linguistique et statistique caractérisant ces entités afin d'en déduire des modèles probabilistes de reconnaissance et de typage. Des lexiques (simple mise en correspondance, une entité = un type) peuvent aussi être utilisés par ces algorithmes. Les recherches exploitant ces méthodes se sont d'abord intéressées aux noms de lieu, avec notamment les travaux de (Fleischman M., 2001), (Lee S. *et al.*, 2004) et (Li H. *et al.*, 2006). En usant de méthodes similaires (Fleischman M., *et al.*, 2002) ainsi que (Mann G. *et al.*, 2003) ont porté leur attention sur la sous-catégorisation de noms de personne. Dans la même veine, il est possible de citer les travaux de T. Poibeau, en pointant néanmoins quelques différences : s'il exploite lui aussi des méthodes d'apprentissage, il le fait cependant suivant un objectif légèrement différent, et s'attache à poser le problème en des termes tant linguistiques qu'applicatifs. En effet, dans (Poibeau T., 2006), l'auteur rend compte d'une expérience de repérage d'usages métonymiques d'EN, et ce pour tous les types d'EN (au moins au départ, l'expérience se focalisant plus sur les EN de type « organisation » par la suite). Sa pierre angulaire n'est pas tant l'obtention de résultats fameux, mais plutôt l'établissement d'un cadre de « représentation des connaissances » attachées aux EN. Ce dernier doit permettre de rendre compte d'une analyse de la saillance de tel ou tel trait d'une entité dans un contexte donné. Ceci fait bien sûr écho à (Poibeau T., 2005), où une réflexion sur le statut référentiel des EN conduit l'auteur à plaider pour une « analyse dynamique par profilage de sens en fonction du contexte ». L'approche proposée ici s'inscrit sans conteste dans cette perspective. Le deuxième type de travaux s'intéresse à la construction de ressource spécifique pour les EN. Peu de choses (à notre connaissance) existent à ce jour : (Mann G., 2002) construit automatiquement une ontologie de noms propres à l'aide de patrons de co-occurrence, avec pour objectif une intégration des

[9] Pour un état de l'art sur les systèmes de reconnaissance des EN, nous renvoyons à (Poibeau T., 2001)

données obtenues à WordNet et (Maurel D. *et al.*, 2006) mènent des travaux similaires au sein du projet *Prolex*, réunissant (« manuellement», à partir de ressources existantes) des informations sur des noms propres, avec une dimension multilingue[10]. Ayant pour idée sous-jacente la notion d'ontologie, ces travaux cherchent à caractériser les noms propres de manière statique, tandis que la ressource présentée ici s'attache à refléter une information contextuelle, en prenant en compte tous ses changements. Enfin, certains tentent de traiter finement les EN en s'appuyant sur des ressources spécifiques. (Bunescu R. *et al.* 2006) présentent une approche fort intéressante de désambiguïsation d'EN qui exploite la ressource encyclopédique Wikipédia. Le travail présenté dans (Paşca M., 2004) semble être celui qui se rapproche le plus de ce que nous présentons : elle construit une ressource à partir de corpus pour annoter finement les EN. Sa méthode de construction de ressource diffère cependant de celle présentée ci-après dans la mesure où elle n'utilise pas d'analyse syntaxique. Signalons pour finir (Nissim M. *et al.*, 2003), dont les travaux sur la métonymie participent d'un domaine de recherche similaire.

Il convient de dire un mot sur l'acquisition de lexiques sémantiques, dans la mesure où nous nous inspirons de ces travaux sur les unités lexicales pour traiter les EN. Cette tâche a connu un essor important depuis l'apparition de corpus de grande taille. Jusque récemment, l'acquisition, à partir de textes, de labels sémantiques associés à des unités lexicales était basée sur des patrons lexico-syntaxiques développés à l'aide d'algorithmes d'apprentissage ; seule Paşca a proposé d'appliquer ce processus aux EN. L'exploitation de corpus analysés syntaxiquement (avec des relations de dépendances) est plus récente, les travaux de (Phillips W. *et al.,* 2002) vont dans cette direction.

Enfin, il importe de replacer notre approche dans la droite ligne des nombreux travaux dit de « linguistique de corpus » (Habert *et al.*, 1997), et notamment ceux de Didier Bourigault (http://w3.univ-tlse2.fr/erss/voisinsdelemonde/) qui propose une méthode de rapprochement sémantique entre mots ou groupes de mots en fonction de leur distribution syntaxique dans un corpus donné.

4. Méthode

4.1. *Construction d'une ressource d'entités nommées*

Ce que nous appelons ressource d'EN est une liste d'EN avec pour chacune d'elles une liste d'étiquettes sémantiques fines potentielles (par exemple les étiquettes « porte-

[10] Contrairement à notre approche dépendante du corpus, le projet *Prolex* s'inscrit nettement dans une perspective ontologique ; néanmoins, les informations associées aux noms propres dans cette base sous le chapeau « expansion classifiante » relèvent du même esprit que nos étiquettes.

avions », « maréchal », « avenue », « hôpital » pour l'entité nommée *Foch*) provenant d'un corpus. Le principe général de construction de cette ressource est l'identification dans le corpus de mots ou groupes de mots étant en relation avec les EN et pouvant servir d'étiquettes sémantiques. Afin de repérer et d'associer pertinemment entités et étiquettes, nous proposons d'exploiter un analyseur syntaxique robuste. Le processus de construction de la ressource se déroule en trois étapes : identification des relations syntaxiques pertinentes permettant d'associer des entités avec des étiquettes, construction « effective » de la ressource et gestion des étiquettes par le calcul de cliques. Il convient de détailler chacune de ces étapes ; nous précisons auparavant les données et les outils utilisés.

4.1.1. *Analyseur syntaxique et corpus utilisés*

Pour les différentes expérimentations effectuées nous avons utilisé deux corpus en français : un corpus contenant l'ensemble des articles du journal *Le Monde* de 1992 à 1996 (2 830 180 phrases) et un corpus contenant articles et dépêches provenant de différentes sources et traitant tous de la crise en Côte d'Ivoire entre 2002 et 2003[11] (331 433 phrases). Dorénavant, nous les nommerons respectivement corpus LM92-96 et corpus CI02-03. Ces corpus ont été traités à l'aide de l'analyseur syntaxique robuste XIP (*Xerox Incremental Parser* (Aït *et al.,* 1997) et (Aït *et al.,* 2002)).

4.1.2. *Identification des relations syntaxiques pertinentes*

Nous disposons de corpus, d'un analyseur syntaxique et avons pour objectif d'associer à des EN des étiquettes sémantiques précisant leur référent. Les objets manipulés lors de la construction de la ressource sont donc des *entités*, des *étiquettes* et des *relations syntaxiques* : il importe dans une première étape de déterminer précisément les caractéristiques de ces objets. Pour cela, il est possible de se baser sur des critères linguistiques d'une part, et de prendre en compte les conclusions d'une observation empirique des résultats de l'analyse syntaxique d'autre part. Ainsi nous cherchons tout d'abord des *entités*, que nous choisissons de définir de la manière suivante : est une *entité nommée potentielle* tout nom ou groupe nominal dont la tête commence par une majuscule. Nous cherchons ensuite des mots pouvant servir d'étiquettes à ces entités ; pour ce faire, nous nous intéressons davantage à des syntagmes modifieurs entretenant avec l'entité potentielle un rapport déterminatif et non explicatif ou descriptif. Est ainsi une *étiquette potentielle* tout nom ou syntagme nominal dont la tête nominale commence par une minuscule. Nous excluons donc les adjectifs (plus qualifiants que classifiants) ainsi que les expressions temporelles (nom de mois, années, etc.) et numériques. Nous

[11] Ce corpus a été constitué dans le but d'expérimentations sur les entités nommées dans le cadre du projet Infom@gic.

observons ensuite toutes les relations syntaxiques établissant un lien entre ces deux types d'objets, étiquette potentielle et entité potentielle. Ainsi, pour chaque entité potentielle, on identifie une liste d'étiquettes potentielles rattachées à cette entité par un certain nombre de relations. Nous appellerons dorénavant chaque combinaison [étiquette potentielle-relation syntaxique], un *contexte syntaxique*.

Le tableau 1 illustre les contextes syntaxiques les plus fréquents pour trois EN potentielles (*Chirac*, *Foch*, *PC*) dans le corpus LM92-96 à partir d'une analyse XIP :

NOUN;Chirac	NOUN;Foch	NOUN;PC
1 NOUN;président.NMOD	1 NOUN;porte-avions.NMOD	1 NOUN;PS.COORDITEMS
1 NOUN;candidat.NMOD	1 NOUN;avenue.NMOD	1 NOUN;secrétaire général.NMOD_DE
1 NOUN;gouvernement.NMOD	1 NOUN;hôpital.NMOD	1 NOUN;congrès.NMOD_DE
1 NOUN;école.NMOD	1 NOUN;maréchal.NMOD	2 NOUN;Macintosh.COORDITEMS
1 NOUN;Balladur.COORDITEMS	2 NOUN;Clémenceau.COORDITEMS	1 NOUN;secrétaire.NMOD_DE
1 NOUN;monsieur.NMOD	1 NOUN;successeur.NMOD_A	1 NOUN;membre.NMOD_DE
2 NOUN;Jospin.COORDITEMS	1 NOUN;bord.NMOD_DE	1 NOUN;dirigeant.NMOD_DE
1 NOUN;époux.NMOD	2 NOUN;premier.NMOD_POUR	1 NOUN;comité.NMOD_DE
1 NOUN;élection.NMOD_DE	1 NOUN;service.NMOD_DE	1 NOUN;ordinateur.NMOD
…	…	…

Tableau 1. *Contextes syntaxiques les plus fréquents pour les entités* **Chirac**, **Foch** *et* **PC**

Le contexte syntaxique *1.NOUN ;président.NMOD* (contexte syntaxique le plus fréquent pour l'entité *Chirac*) se lit de la manière suivante : « *NOUN ;président* » décrit l'étiquette potentielle impliquée dans la relation syntaxique à l'aide de son type (*NOUN*) et de son lemme (*président*) ; « 1 » signifie que cette étiquette potentielle est le recteur de la relation syntaxique (« 2 » : étiquette en position régie) ; « NMOD » décrit le type de la relation syntaxique (NMOD : modifieur de nom ; NMOD_DE : modifieur de nom impliquant la préposition *de* ; COORDITEMS : relation de coordination ; etc.)

C'est à partir de ces listes de contextes que nous déterminons les relations syntaxiques pertinentes pour établir un lien entre une entité nommée et ses étiquettes potentielles. Nous avons identifié trois relations pertinentes (elles sont surlignées en gris dans le tableau 1) :

- La relation modifieur de nom sans préposition (NMOD). Dans le tableau 1, elle permet de faire le lien entre *Chirac* et « président, candidat, gouvernement ».
- La relation attribut (ATTRIB). Dans le tableau 1, elle permet de faire le lien entre *Chirac* et « candidat » et le lien entre *Foch* et « place ».
- la coordination (COORDITEMS) Dans le tableau 1, elle permet de faire le lien entre deux entités du même type (*Chirac* avec *Balladur*, *Jospin* et *Juppé*)[12].

[12] La relation COORDITEMS fait le lien entre deux entités potentielles. C'est pourquoi les noms employés avec cette relation (Balladur, Jospin, Juppé) possèdent les caractéristiques des entités potentielles. Nous n'exploiterons pas cette relation dans cette étude.

Le choix manuel de ces relations syntaxiques est la seule étape supervisée de notre méthode. Ce choix dépend de l'analyseur syntaxique utilisé et de la langue du texte. Si ces choix sont faits empiriquement, il n'en reste pas moins que les relations identifiées correspondent à celles habituellement reconnues au sein du groupe nominal étendu et qu'elles sont constantes au sein d'une langue.

L'avantage d'une approche syntaxique par rapport à une approche de simple *pattern matching* peut être illustré par deux types d'énoncés : les énoncés du type *Le très vieux Foch* sont retenus à tort avec une approche a base de patrons puisque *vieux* peut être un nom mais pas avec l'approche syntaxique puisque celle-ci permet d'attribuer à *vieux* le type « adjectif » ce qui l'exclut des étiquettes potentielles. De même, des énoncés plus complexes peuvent être exploités. Par exemple, pour l'énoncé *Bush est, pour le moment, un très (bon/mauvais) président* l'analyse syntaxique profonde permet d'établir une relation « attribut » entre *Bush* et *président*, et ainsi de saisir une étiquette à distance.

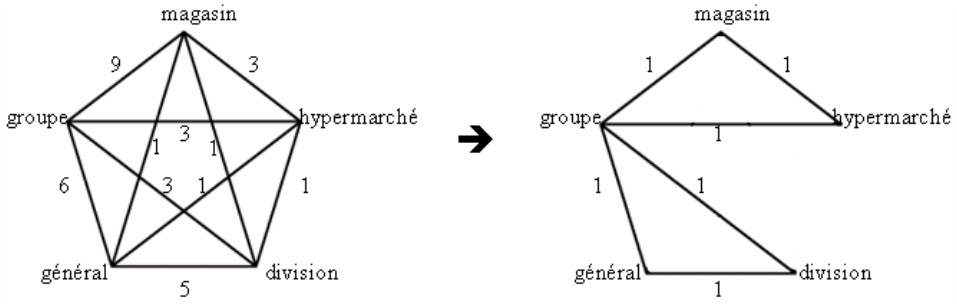
4.1.3. *Construction effective de la ressource*

La seconde étape correspond à la construction de la ressource à proprement parler. À partir du corpus, nous extrayons toutes les occurrences des relations syntaxiques pertinentes identifiées lors de l'étape précédente. Dorénavant, nous appellerons R cette liste de relations pertinentes. Pour chacune de ces relations syntaxiques, le recteur est introduit dans la ressource en tant qu'*étiquette potentielle* et le régit en tant qu'*entité nommée potentielle.* De cette manière, nous pouvons construire une matrice M (*étiquettes* x *entités*) où chaque ligne est une *étiquette potentielle* et chaque colonne est une *entité nommée potentielle.* Dans cette matrice, la valeur de (ligne = *candidat ;* colonne = *Chirac*) est égale à la somme des fréquences de chaque relation syntaxique de R entre « candidat » et *Chirac* dans le corpus. Nous verrons dans le paragraphe 3.1.4 qu'il existe différentes manières de filtrer les informations contenues dans cette ressource. Pour nos expérimentations, nous avons choisi d'exclure simplement toutes les valeurs (ligne, colonne) égales à 1, autrement dit toutes les relations étiquette-entité présentes une seule fois dans le corpus.

Les étiquettes obtenues sont la plupart du temps pertinentes mais il peut y avoir un effet de surproduction de ces étiquettes. *Leclerc* peut être un « char », un « supermarché », un « maréchal », mais les étiquettes obtenues sont les suivantes : « groupe, magasin, char, division, général, centre, programme, combat, colonne, hypermarché, maréchal, supermarché, bataille ». « Supermarché » et « magasin » semblent renvoyer à la même annotation fine alors que « groupe » reste une étiquette ambiguë (un groupe peut renvoyer à un *parti politique*, une *entreprise*, un *rassemblement d'individus*, etc.). Nous proposons de substituer à cette liste d'étiquettes une liste de groupes d'étiquettes, c'est-à-dire un ensemble d'étiquettes pouvant renvoyer à une même annotation fine.

La méthode de construction de ces ensembles d'étiquettes est la suivante : à partir de la matrice M(*étiquettes* x *entités*) citée ci-dessus, nous construisons un graphe valué G d'étiquettes. Dans ce graphe, chaque sommet est une étiquette et le nombre d'arêtes reliant deux sommets correspond au nombre d'entités ayant ces deux étiquettes en commun. À partir de ce graphe nous effectuons des regroupements d'étiquettes par le calcul de cliques. Une clique est un sous-graphe complet maximal, soit un ensemble le plus grand possible d'étiquettes toutes reliées deux à deux. Les cliques ont déjà été exploitées pour la représentation d'espaces sémantiques d'unités lexicales à partir de dictionnaires de synonymes (Ploux et Victorri, 1998 ; Victorri, 2002), l'idée étant alors de considérer que chaque clique de synonymes correspond à un sens très précis de l'unité lexicale étudiée. De la même manière, dans notre cas, une clique d'étiquettes correspond à une annotation d'entité très fine. Si l'étiquette « groupe » est ambiguë, la clique « groupe, société, firme » permet de contraindre l'emploi de « groupe » en tant qu'*entreprise*.

Pour une entité nommée *E* possédant une liste d'étiquettes *L* nous extrayons de ce graphe G le sous-graphe G(L,U), c'est-à-dire le sous-graphe composé des sommets (étiquettes) *L* et des arêtes U reliant ces sommets. Nous construisons alors un graphe simple G'(L,U') dans lequel deux sommets sont reliés par une arête si dans le sous-graphe G(L,U) le nombre d'arêtes entre ces deux mêmes sommets est supérieur à *k*. C'est dans ce graphe simple G'(L,U') que nous calculons les cliques d'étiquettes.



Extrait du sous-graphe G(L,U)
pour l'entité Leclerc

Extrait du sous-graphe G'(L,U')
pour l'entité Leclerc avec k=2

Si $k = 0$, l'entité *E* possédant toutes les étiquettes du sous graphe G(L,U), *E* permet de faire le lien entre toutes ces étiquettes et le calcul donne une et une seule clique composée des étiquettes *L*, ce qui n'est pas notre but ici. Nous proposons d'illustrer les cliques obtenues avec $k = 1$ et $k = 2$ à l'aide de l'entité *Leclerc* (cf. ci-dessous). Si l'étiquette « groupe » est ambiguë, les cliques dont elle fait partie ne le sont plus : pour $k = 2$, les deux premières cliques « groupe ; magasin ; hypermarché » et « groupe ;

division ; général » correspondent bien à deux annotations fines et non ambiguës pour l'entité *Leclerc*.

cliques obtenues pour l'entité *Leclerc* avec $k = 1$	cliques obtenues pour l'entité *Leclerc* avec $k = 2$
NOUN:groupe ;NOUN:magasin ;NOUN:hypermarché	NOUN:groupe ;NOUN:magasin ;NOUN:hypermarché
NOUN:groupe ;NOUN:magasin ;NOUN:supermarché	NOUN:groupe ;NOUN:division ;NOUN:général
NOUN:groupe ;NOUN:division ;NOUN:général	NOUN:groupe ;NOUN:programme
NOUN:groupe ;NOUN:division ;NOUN:programme	NOUN:centre
NOUN:général ;NOUN:maréchal	NOUN:général ;NOUN:maréchal
NOUN:char	NOUN:char
NOUN:centre	NOUN:programme ;NOUN:combat
NOUN:programme ;NOUN:combat	NOUN:colonne
NOUN:colonne	NOUN:supermarché
NOUN:bataille	NOUN:bataille

Tableau 1. *Cliques obtenues pour l'entité* Leclerc

Si nous insistons sur cette variable *k*, c'est qu'elle permet de jouer sur le nombre de cliques calculées, autrement dit sur la finesse d'annotation des entités. $k = 0$ revient à considérer que toutes les entités sont non ambiguës puisqu'elles possèdent toutes une seule clique d'étiquettes et donc une seule annotation possible. Au plus on augmente *k*, au plus on se rapproche de la liste classique d'étiquettes telle celle présentée pour l'entité *Leclerc (« groupe, magasin, char, etc.)*. Pour les expérimentations qui suivent nous avons utilisé $k = 2$ (ce choix correspond à un choix empirique).

En utilisant cette méthode, nous avons construit une ressource à partir du corpus LM92-96. Pour cette construction, l'ensemble de relations R est composé des relations « modifieur de nom sans préposition » (NMOD) et « attribut » (ATTRIB) et sont exclues toutes les relations étiquette-entité présentes une seule fois dans le corpus. La ressource obtenue contient 15 040 entités différentes et 2 547 étiquettes différentes. Le temps de construction de cette ressource correspond au temps d'analyse du parseur[13].

Entité : Chirac (appartient à 25 cliques d'étiquettes)	**Entité : Foch 4cliques d'étiquettes**	**Entité : PC (appartient à 14 cliques d'étiquettes)**
NOUN:président ;NOUN:famille ;NOUN:ère ; NOUN:gouvernement	NOUN ;porte-avions	NOUN:groupe ; NOUN:maire ;NOUN:liste ; NOUN:candidat ; NOUN:maison ;…
NOUN:candidat ;NOUN:ère ; NOUN:plan ;NOUN:effet ;NOUN:vote	NOUN ;avenue	NOUN:groupe ;NOUN:machine ; NOUN:micro-ordinateur ;NOUN:ordinateur
NOUN:famille ;NOUN:époux ; NOUN:couple	NOUN ;hôpital	NOUN:microprocesseur ;NOUN:machine ; NOUN:puce ;NOUN:processeur
…	NOUN ;maréchal	…

Tableau 2. *Entités* Chirac, Foch *et* PC *avec un extrait de leurs cliques.*

[13] Le parseur que nous avons utilisé, XIP, traite 1 800 mots par seconde. Il faut 2 h 15 pour construire la ressource à partir du corpus CI02-03.

Les tableaux 2 et 3 présentent quelques entités et étiquettes extraites de cette ressource. Le tableau 2 présente trois entités (*Chirac, Foch* et *PC*) avec pour chacune d'elles un extrait de leurs cliques d'étiquettes. Les cliques de l'entité *Chirac* rendent compte de différentes facettes d'une même entité, celles de l'entité *Foch* décrivent différents emplois polysémiques, enfin celles de l'entité *PC* révèlent un cas d'homonymie (*ordinateur* vs. *Parti communiste*). Le tableau 3 décrit quelques étiquettes de la ressource avec les entités les plus fréquentes pour chacune de ces étiquettes (les entités *Foch, Chirac* et *PC* sont surlignées en gris).

NOUN;acteur (contient 11 entités)		NOUN;assurance (contient 19 entités)		NOUN;porte-avions (contient 11 entités)	
NOUN; Marina Vlady	(6)	NOUN;AXA	(36)	NOUN ;Charles-de-Gaule	(34)
NOUN; Sharon Stone	(4)	NOUN;GAN	(16)	NOUN;Foch	(24)
NOUN; Jean-Paul Belmondo	(3)	NOUN;Allianz	(12)	NOUN;Clémenceau	(17)
NOUN; Steve Martin	(3)	NOUN;Generali	(10)	NOUN;Independence	(7)
…		…		…	
NOUN;gouvernement (contient 93 entités)		NOUN;ordinateur (contient 20 entités)		NOUN;avenue (contient 50 entités)	
NOUN;Balladur	(422)	NOUN;Apple	(14)	NOUN;Jean - Jaurès	(85)
NOUN;Juppé	(223)	NOUN;PC	(14)	NOUN;Montaigne	(70)
NOUN;Rocard	(82)	NOUN;Macintoch	(12)	NOUN;Victor – Hugo	(23)
NOUN;Chirac	(45)	NOUN;IBM	(12)	NOUN;Daumesnil	(22)
…		…		…	

Tableau 3. *Étiquettes de la ressource avec leurs entités les plus fréquentes*

4.1.4. *Étape facultative*

Une troisième étape facultative consistant en un filtrage de la ressource peut être effectuée. Ce filtrage peut correspondre à l'adaptation de la ressource pour une application donnée. Il est possible de supprimer certaines annotations non pertinentes comme d'en regrouper d'autres, ou encore d'ajouter certaines annotations ou EN provenant d'une autre ressource. Si nous insistons sur cette étape facultative alors que nous ne l'avons pas appliquée dans nos expérimentations, c'est que même si notre ressource peut être exploitée telle quelle et déjà améliorer les résultats d'annotation d'EN dans un texte, elle n'est pas une boîte noire et peut être manipulée en toute transparence en fonction d'une application donnée.

4.2. *Annotation fine ou première désambiguïsation*

Dans cette partie nous présentons l'intérêt d'une telle ressource pour l'annotation et la désambiguïsation d'EN. En effet, sans faire appel à une méthode de désambiguïsation classique nécessitant de faire un choix entre plusieurs annotations possibles, il est envisageable, à l'aide de la ressource seule, d'annoter de manière fine et non ambiguë des EN. Pour illustrer cela, nous détaillons les différents contextes d'emploi des occurrences d'EN dans un corpus donné en fonction de l'information présente dans notre

ressource. Nous appuyons cette étude sur le corpus CI02-03 (331 433 phrases) et toutes les fréquences d'occurrences que nous citons dans cette partie proviennent de ce corpus. La ressource a été construite à partir d'une liste de relations R, composée des relations « modifieur de nom sans préposition » (NMOD) et « attribut » (ATTRIB), et en excluant toutes les relations étiquettes-EN présentes une seule fois dans le corpus. La ressource construite contient ainsi 2 360 entités nommées différentes et 967 étiquettes différentes. Ces 2 360 entités représentent 284 702 occurrences d'entités, soit plus d'une entité par phrase. Nous ne connaissons pas le nombre réel d'EN présentes dans ce corpus, cependant nous verrons dans le paragraphe 4.4 que sur un extrait contenant 855 occurrences d'EN, 70 % ont été annotées à l'aide de notre ressource.

Il existe différents cas de figure pour l'annotation des occurrences d'entité nommée en fonction de l'information présente dans la ressource et dans le contexte :

1. L'entité nommée n'a qu'une seule clique d'étiquettes dans la ressource. Dans ce cas l'annotation fine est faite directement. C'est le cas par exemple de l'entité *Daouda Konate* qui ne possède dans la ressource que la clique « sergent ; sergent chef ». Dans la ressource construite à partir du corpus CI92-93, 84 % des entités de la ressource ne possèdent qu'une clique (87 % dans le corpus LM92-96). Ces entités représentent 60 195 occurrences dans le corpus, soit 21,1 % des occurrences d'entités annotées du corpus.
2. L'entité nommée a plusieurs cliques d'étiquettes dans la ressource. Trois types d'occurrences sont alors possibles :
    a. Les occurrences ayant une relation syntaxique appartenant à R avec une étiquette appartenant à *une seule* clique parmi les cliques d'étiquettes possibles pour cette entité dans la ressource. Par exemple, l'énoncé « *Le président de la République, son excellence Laurent Gbagbo, déclarait que…* » permet d'établir une relation entre l'entité *Laurent Gbagbo* et l'étiquette « président de la République». L'entité *Laurent Gbagbo* possède plusieurs cliques d'étiquettes, mais une seule (« président, candidat, chef d'État, président de la République ») contient l'étiquette « président de la République». Dans ce cas, l'annotation fine est faite directement, ceci intervient pour 3 % des occurrences d'EN annotées du corpus CI92-93.
    b. Les occurrences ayant une relation syntaxique appartenant à R avec une étiquette appartenant *à plusieurs* cliques parmi les cliques d'étiquettes possibles pour cette entité dans la ressource. Par exemple, l'énoncé « *Le général Gueï prend la tête du conseil…* » permet d'établir une relation entre l'entité *Gueï* et l'étiquette « général ». L'entité *Gueï* possède plusieurs cliques d'étiquettes dont deux contenant l'étiquette « général » (les cliques « président, général » et « général, division »). Dans ce cas, l'entité reste

ambiguë ; cependant son ambiguïté est réduite puisque dans notre exemple, il n'y a plus que deux cliques possibles parmi les quatre que possède l'entité. Ce cas correspond à 5,6 % des occurrences d'EN de CI92-93.

c. Les occurrences qui n'ont pas de relation syntaxique appartenant à R avec l'une des étiquettes potentielles de l'entité et dont l'entité nommée possède plusieurs cliques dans la ressource (70,3 % des occurrences d'EN de CI92-93). Dans ce cas l'entité reste ambiguë : nous gardons toutes les cliques d'étiquettes de l'entité, et une méthode de désambiguïsation pourra exploiter cette liste de cliques afin de déterminer la ou les cliques d'étiquettes adaptée(s) à l'énoncé.

Cette ressource construite à partir d'un corpus n'est donc pas seulement une ressource de référence permettant de connaître toutes les annotations fines possibles d'une entité en vue d'une désambiguïsation ultérieure, elle permet déjà à elle seule d'annoter de manière fine et non ambiguë un nombre important (24,1 %) d'entités dans le texte. Si l'on considère que l'annotation par une seule étiquette est suffisante, ce taux passe même à 29,7 % puisque l'on peut ajouter le cas de figure 2.b. Ainsi, à qualité de désambiguïsation égale, l'exploitation de cliques d'étiquettes apparaît comme bénéfique (26,3 % d'entités désambiguïsées à l'aide d'étiquettes non regroupées). Cette ressource met aussi en avant le fait que 84 % des entités sont non ambiguës dans ce corpus. Avant d'évaluer la qualité de ces annotations fines, nous proposons d'introduire ce que nous appelons la double annotation.

4.3. *Double annotation*

L'objectif est de combiner l'information contenue dans la ressource que nous venons de présenter avec un système classique de traitement d'EN afin de permettre une annotation selon deux niveaux, général et particulier, avec l'idée sous-jacente d'exploiter chaque méthode pour la tâche pour laquelle elle est le mieux adaptée. Il existe un nombre important de systèmes, qu'ils soient fondés sur une approche symbolique ou par apprentissage, qui sont capables d'annoter des EN selon des catégories générales, et ce avec des taux de précision supérieurs à 90 %. Il importe cependant pour ces systèmes que les textes à annoter soient du même genre et/ou domaine que ceux ayant guidé leur conception, voire qu'ils soient issus d'une même source. Ainsi, l'utilisation de systèmes classiques pour d'autres domaines que celui pour lequel ils ont été conçu amoindrit généralement les performances, notamment en faisant chuter le rappel (ce ne sont plus les mêmes entités que l'on retrouve ni les même types d'entités : certains types d'entités comme *noms de produits* peuvent être présents dans certains domaines et pas dans d'autres). De manière identique, dès que l'on cherche à affiner les annotations d'EN en

passant de « organisation » à « entreprise », « association » ou « parti politique », les taux de précision chutent.

Notre approche d'annotation fine à partir d'une ressource d'EN dépendante d'un corpus, dorénavant approche « ressource », a les qualités inverses : étant dépendante du corpus, elle permet de proposer des annotations précises et fines pour les entités correspondant au corpus puisqu'elles en ont été extraites. La limite de cette approche est qu'elle est difficilement rattachable à une catégorisation générale préalable de type « personne, organisation, lieu, etc. ». Elle permet de dire que *Coca-Cola* est une « société », mais pas que « société » est rattachée à la catégorie « organisation », du moins pas sans faire appel à une ressource de nature ontologique externe de type Wordnet.

L'idée est alors, non d'essayer de fusionner ces deux approches, mais plutôt de les rendre complémentaires. Il nous semble important de maintenir la combinaison « approche classique/approche ressource » afin d'exploiter au mieux les qualités de chacune (ceci n'excluant pas de les faire interagir) : la première permet d'annoter des entités avec des catégories connues alors que la seconde permet de proposer de nouvelles (sous-)catégories. Autrement dit, nous proposons une double annotation, l'une générale obtenue à l'aide d'une approche symbolique classique et l'autre fine obtenue à l'aide de l'approche « ressource ». Ci-dessous, nous illustrons avec l'entité *Leclerc* le type de résultats que nous cherchons à obtenir.

(1) *Ce supermarché Leclerc vient d'être inauguré*

- Annotation générale = ORGANISATION
- Annotation fine = supermarché

Pour cet énoncé (1) l'approche classique est en mesure de proposer, grâce à la structure (supermarché + nom commençant par une majuscule), une annotation générale de type « organisation ». L'approche « ressource » est quant à elle en mesure de proposer une annotation plus fine : malgré l'ambiguïté de l'entité *Leclerc* (cf. les cliques de cette entité en § 3.1.3), La clique « supermarché » peut lui être attribuée directement grâce à l'identification de la relation « modifieur » avec le nom *supermarché* dans l'énoncé et à la présence de la clique « *supermarché* » dans la ressource (cas 2.a. dans le § 4.2). Ainsi, les deux méthodes se révèlent complémentaires, l'une venant enrichir l'annotation générale de l'autre par une étiquette sémantique de grain plus fin. L'objectif est, idéalement, d'avoir une double annotation pour toutes les occurrences d'entités. Cependant, certains énoncés ne permettent qu'une annotation partielle. Ainsi, si elle n'est pas possible dans tous les cas, la double annotation permet d'enrichir l'annotation de certaines EN d'une part (c'est le cas de l'énoncé (1)), et d'annoter, au moins

partiellement, un nombre plus important d'occurrences d'entités dans le texte d'autre part[14].

4.4. *Évaluation de l'annotation fine*

Dans cette partie nous présentons une évaluation de la qualité des annotations fines obtenues à l'aide de l'approche « ressource ». Toute la difficulté est de savoir comment juger de la pertinence ou justesse d'une annotation fine : dire que *Coca-Cola* est une « société » est-il plus approprié que dire que c'est une « firme » ? Est-ce que dire que *Bush* est un « président » permet de dire que c'est un « homme politique » ?

Nous revendiquons un système d'annotations doubles (classique + fin), mais afin de comparer les approches classique et « ressource » sur des annotations identiques, nous avons choisi d'établir une correspondance entre nos étiquettes fines et les catégories générales « personne », « lieu » et « organisation ». Nous insistons sur le fait que cette mise en correspondance n'est faite que dans le but de l'évaluation et non dans le but de construire une ressource hiérarchique à deux niveaux interdépendants.

Pour l'approche classique, différents systèmes d'annotation peuvent être utilisés, qu'ils soient de type symboliques ou basés sur de l'apprentissage. Nous avons choisi d'utiliser le système de reconnaissance d'EN fondé sur une approche symbolique intégrée au sein de l'analyseur syntaxique XIP (Brun *et al.*, 2004). Ce choix se justifie par le fait que XIP nous donne la possibilité d'effectuer les deux types d'annotation, classique et fin, au sein d'un même processus. Outre le gain en rapidité, cela permet de maintenir une certaine cohésion entre les deux approches, les annotations étant faites sur des unités de texte provenant de la même analyse. Le système symbolique classique que nous avons utilisé a fait l'objet d'une évaluation pour le français (Rebotier, 2006). Précisons tout d'abord que l'élaboration de ce système (constitution de lexiques et écritures de règles) a été guidée par l'étude d'un corpus composé d'articles du journal *Libération* (Lib1). L'évaluation a ensuite été menée à partir de deux corpus : un corpus similaire, composé d'autres articles de *Libération* (Lib2), et un corpus plus éloigné thématiquement afin d'évaluer la transposabilité inter-domaine du système, composé d'extraits du corpus CI92-93 (cf. § 3.1.1.) La F-mesure est de 0,90 pour l'extrait Lib2 et de 0,80 pour l'extrait CI. Nous avons choisi l'extrait CI pour notre évaluation dans le but de montrer que la double annotation, outre sa finesse, permet de combler la perte de qualité d'un système classique lorsque l'on change de domaine. L'extrait CI

[14] Le temps de calcul de cette double annotation est du même ordre que celui de l'annotation classique soit 1 800 mots par seconde.

contient 765 entités annotées manuellement par une catégorie (« lieu », « personne », « organisation »). Sur cet extrait, la précision obtenue par l'approche symbolique classique est de 86,58 %, et le rappel de 74,25 %.

L'évaluation consiste à tester si notre approche « ressource » permet d'améliorer les résultats d'annotation de l'approche classique. C'est pourquoi nous ne nous intéressons dans cette évaluation qu'à la qualité des résultats obtenus sur les entités « oubliées » par l'approche classique. Sur les 765 entités du corpus de référence, 149 ont été « oubliées » par cette dernière. La plupart du temps, ces oublis sont dus à des contextes d'emplois difficiles à traiter pour l'approche classique (ex. : *Mais Hanny Tchelley ne comptait pas s'arrêter là.*). Nous avons donc évalué les résultats de l'approche « ressource » sur ces 149 cas.

La correspondance entre étiquettes provenant du corpus et étiquettes classiques a été faite par quatre juges qui ont dû classer les premières en fonction des secondes, ou dans une catégorie « autre » (par exemple l'étiquette *firme* a été apparentée à ORGANISATION par les quatre juges, *administration* à PERSONNE ou ORGANISATION et *voiture* à AUTRE). L'accord inter-annotateur est de 92 %. Nous avons intégré ces jugements dans chaque clique d'étiquettes (cf. colonnes 3 et 4 du tableau 4). Nous présentons ci-dessous un extrait des 37 entités qui ont été correctement annotées par l'approche « ressource » ainsi que la seule erreur relevée, suivie du type d'énoncé qui explique cette erreur. Une entité correctement annotée correspond à une entité annotée par une seule clique (cas 1 ou 2.a. du § 4.2). En terme d'extraction d'information, le tableau 4 illustre l'intérêt de cette annotation fine. Si l'on recherche toutes les personnes qui on eu la fonction de ministre, les entités *Alassane Ouattara* et *Niamien Messou* sont sélectionnées. Si, en revanche, on ne s'intéresse qu'aux personnes qui ont un grade militaire, seul *Alassane Ouattara* est sélectionné. Dans ces deux requêtes fines l'entité *Hanny Tchelley* n'est pas pertinente et peut facilement être exclue.

L'erreur (cf. tableau 5) vient du fait que l'entité *Lama Bamba* possède dans notre ressource l'étiquette « presse ». Cette erreur provient d'énoncés tels que celui-ci : « *Pour l'instant, je n'en dis pas plus, a sobrement déclaré à la* ***presse Lama Bamba ».*** Dans ce cas, le lien établi entre *presse* et *Lama Bamba* est dû à une erreur de l'analyse syntaxique.

Occurrences d'entités	**Annotations de réf.**	**cliques obtenues**	**Correspondance**
Hanny Tchelley	PERSONNE	animateur	PERSONNE (4/4)
M. Alassane Ouattara	PERSONNE	commandant ;sergent ; professeur ;ministre	PERSONNE (4/4) ; PERSONNE(4/4) PERSONNE (4/4) ; PERSONNE(4/4)
Niamien Messou	PERSONNE	ministre ;délégué	PERSONNE (4/4) ;PERSONNE (4/4)
Le Patriote	ORG	quotidien	ORGANISATION (4/4)
Pdci	ORG	Régime	ORGANISATION (3/4)
Man	LIEU	capitale	LIEU(4/4)

Tableau 4. *Extrait des entités correctement annotées*

Entités	**Annotations de réf.**	**cliques obtenues**	**Correspondance**
Lama Bamba	PERSONNE	presse	ORGANISATION (4/4)

Tableau 5. *Entité mal annotée*

Le bilan global est le suivant : sur les 149 entités oubliées par l'approche symbolique classique, 37 ont été correctement annotées par l'approche « ressource» et une seule a été mal annotée. 111 entités n'ont pas été annotées : elles correspondent soit à des entités qui ne font pas partie de la ressource et donc ne sont pas annotées, soit à des entités qui font partie de la ressource, mais ne sont pas désambiguïsées (cas 2.b. ou 2.c. du § 4.2), soit à des entités qui font partie de la ressource, mais sont annotées avec l'étiquette « autre » par les quatre annotateurs. La précision de ce sous-ensemble de 149 annotations est de 97,4 % et le rappel est de 24,8 %. Rapporté à l'ensemble des entités du texte (soit 765 entités), cela permet une augmentation du taux de précision, passant de 86,58 % à 87,18 %, et surtout une augmentation significative du taux de rappel, qui passe lui de 74,25 % à 79,08 % (cf. tableau récapitulatif 6).

	Méthode classique	Méthode hybride	**Gain**
Précision	86.58 %	87.18 %	**+0.6**
Rappel	74.25 %	79.08 %	**+4.83**
F-mesure	79.9 %	82..9 %	**+3.0**

Tableau 6. *Comparaison des taux de précision et de rappel*

4.5. *Remarque : étude sur les entités ajoutées*

Le corpus de référence que nous avons utilisé a été annoté pour identifier des entités de type « personne », « lieu », « organisation ». Dans ces conditions, toute annotation d'entité ajoutée par l'approche classique par rapport à l'annotation de référence doit être

considérée comme une erreur, ce qui a été fait. En revanche, les annotations provenant de l'approche « ressource » n'ont pas de frontière de catégorie, notre ressource peut attribuer l'étiquette « voiture » à l'entité *BMW* sans se soucier de savoir à quelle catégorie plus large cette étiquette appartient. Dans l'évaluation que nous avons présenté, l'étiquette « voiture » a été classée dans la catégorie « autre » par les quatre annotateurs, autrement dit, l'entité *BMW* a été ignorée. Or, ces entités ajoutées par l'approche « ressource » sont une très bonne illustration des annotations fines que cette approche peut proposer. Nous avons donc décidé de juger, cette fois-ci subjectivement, de la qualité de ces annotations fines. L'approche « ressource » a annoté 44 entités supplémentaires par rapport à celles qui avaient été annotées dans le corpus de référence. Sur ces 44 entités, 26 ont été jugées justes, 2 ont été jugées fausses et 16 ont été jugées neutres. Par neutre, nous entendons soit des entités ayant une étiquette ne portant aucune sémantique telle que l'étiquette « autre », soit des entités possédant plusieurs cliques d'étiquettes ne renvoyant pas à la même catégorie d'entité, autrement dit des entités encore ambiguës (voir le tableau 7 pour quelques exemples de ces jugements).

Occurrences d'entités	**Annotations fines obtenues**	**Jugement**
Ivoiro-Burkinabé	clan	juste
Krou	groupe	juste
Licorne	opération	juste
Messou	professeur	juste
T55	char	juste
Ivoirien	NOUN:ministre ;NOUN:gouvernement ;NOUN:président NOUN:télévision ;…	neutre
ONG	autre	neutre
Républiques	deuxième	faux
Mano	fleuve	faux

Tableau 7. *Exemples de jugements*

Ces jugements ne peuvent pas faire office d'évaluation, néanmoins ces entités « ajoutées » illustrent bien l'intérêt de l'annotation fine à partir du corpus puisqu'elle permet d'identifier des entités et des étiquettes qui n'auraient probablement pas pu être prévues *a priori* comme *opération Licorne*, *clan Ivoiro-Burkinabé* ou encore *char T55.*

5. Conclusion

Nous avons présenté une nouvelle approche pour l'annotation fine des entités nommées. La particularité de cette approche réside dans l'exploitation d'une ressource associant des cliques d'étiquettes sémantiques fines aux EN : construite dynamiquement

à partir de corpus, cette ressource permet de rendre compte des diverses caractéristiques (ou facettes) potentielles du référent d'une entité en contexte. À elle seule, cette ressource rend possible l'annotation fine d'une EN, la mise en valeur de son éventuelle polysémie, et parfois même sa désambiguïsation (désambiguïsation pour 24,1 % des occurrences d'entités dans le corpus CI92-93). Elle présente également l'intérêt de faire émerger de nouveaux types d'entités à partir du texte. Couplée a un système standard de reconnaissance d'EN, cette ressource permet une annotation selon deux niveaux, général et fin. Nous avons parlé de l'intérêt des annotations fines obtenues pour extraire d'un texte des informations précises telles que « personnes ayant un grade militaire » ou « personnes ayant été ministre » (cf. § 4.4). L'intérêt de la double annotation réside dans sa double approche (classique et ressource) et dans la combinaison entre une structure imposée (personne/organisation/lieu) et des informations non-structurées provenant du texte. Malgré la difficulté de l'évaluation de cette double annotation, les résultats obtenus sont prometteurs et semblent attester l'efficience de cette approche.

La poursuite de ce travail est à envisager selon plusieurs perspectives. Il est possible tout d'abord d'affiner la construction de la ressource en considérant d'autres relations syntaxiques pour le repérage d'étiquettes sémantiques, comme par exemple les relations de coordination. Ensuite cette méthode d'annotation fine peut être exploitée pour d'autres langues, l'anglais est en cours de réalisation. Nous envisageons également de réduire les cas d'annotation partielle (annotation seulement générale ou seulement fine) en faisant interagir les deux types d'annotations : savoir qu'une entité possède l'annotation fine « journaliste » doit pouvoir aider à proposer l'annotation générale « personne » pour cette même entité. Enfin, le traitement des entités demeurant ambiguës (16 % des entités dans la ressource construite à partir du corpus CI92-93 représentant 75,9 % des occurrences d'entités dans ce même corpus) constitue une piste de recherche importante, nous développons actuellement un système de désambiguïsation dans cette perspective.

Remerciements

Nous souhaitons remercier l'ensemble des personnes ayant permis la réalisation de ces travaux, et plus particulièrement Caroline Brun, Frédérique Segond et Bernard Victorri pour leurs conseils avisés.

6. Bibliographie

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

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

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

Bunescu R., Paşca M., « Using Encyclopedic knowledge for Named Entity Disambiguation », *Proceedings of the 11th Conference of the European chapter of the ACL*(to appear).

Bourigault D., http://w3.univ-tlse2.fr/erss/voisinsdelemonde

Croft W., Cruse D.A., *Cognitive Linguistics*, Cambridge University Press, 2004.

Fleischman M., « Automated Subcategorization of Named Entities », *Proceedings of ACL Student* Toulouse, France, Association for Computational Linguistics, 2001.

Fleischman M., Hovy E., « Fine grained classification of named entities », *Proceedings of the 19th international conference on Computational linguistics (COLING),* ACL, 2002.

Habert B., Nazarenko A., Salem A., *Les linguistiques de corpus*, Armand Colin, 1997.

Jonasson K., *Le Nom Propre. Constructions et interprétations*, Editions Duculot, collection *Champs Linguistiques*, 1994.

Lee S., Lee G.G., « A Bootstrapping Approach for Geographic Named Entity Annotation, » AIRS, 2004.

Li H., Srihari R., Niu C., Li W., « InfoXtract location normalization: a hybrid approach to geographic references in information extraction », *Proceedings of the HLT-NAACL 2003 workshop on Analysis of geographic references,* ACL, 2003.

Mann G., « Fine-Grained Proper Noun Ontologies for Question Answering », *Proceedings of SemaNet'02,* Building and Using Semantic Networks, 2002.

Mann G., Yarowsky D., « Unsupervised personal name disambiguation », *Proceedings of the seventh conference on Natural language learning at HLT-NAACL*, ACL, 2003.

Maurel D., Tran M., Friburger N. « Projet Technolangue NomsPropres : Constitution et exploitation d'un dictionnaire relationnel multilingue de noms propres », TALN, 2006.

Nilsson K., Malmgren A., « Towards automatic recognition of product names: an exploratory study of brand names in economic texts », *Proceedings of the Nordic Conference of Computational Linguistics (NODALIDA2005),* Finland, 2005.

Nissim N., Market K., «Syntactic features and Word Similarity for supervised metonymy resolution » *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL-03)*, 2003

Pantel P., Ravichandran D., « Automatically Labeling Semantic Classes », *Proceedings of the HLT-NAACL*, 2004.

Paşca M., « Acquisition of categorized named entities for web search », *Proceedings of the thirteenth ACM international conference on Information and knowledge management* (*CIKM '04),* ACM Press, 2004.

Phillips W., Riloff E., « Exploiting strong syntactic heuristics and co-training to learn semantic lexicons », *Proceedings of the ACL-02 conference on Empirical methods in natural language processing*, ACL, 2002.

Ploux S., Victorri B., « Construction d'espaces sémantiques à l'aide de dictionnaires de synonymes », TAL, n°39-1, 1998.

Poibeau T., « *Deconstructing Harry* , une évaluation des systèmes de repérage des entités nommées », *Revue de l'électricité et de l'électronique*, EDP Sciences, 2001.

Poibeau T., « Sur le statut référentiel des Entités Nommées », *Actes de la conférence Traitement Automatique des Langues Naturelles*, Dourdan, France, Atala, 2005.

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

Rebotier A., « Développement d'un module d'extraction d'entités nommées pour le français», rapport de stage, XRCE, Grenoble, 2006.

Sekine S., Named Entity : « History and Future », http://nlp.cs.nyu.edu/sekine/index.html, 2004.

Victorri B., « Espaces sémantiques et représentation du sens », Textualités et nouvelles technologies, éc/artS, 3, 2002.