



TALN 2009, Senlis, 24–26 juin 2009

Vers une méthodologie d'annotation des entités nommées en corpus ?

Karën Fort^{1,3} Maud Ehrmann² Adeline Nazarenko³

(1) INIST, 2 allée du Parc de Brabois, 54500 Vandoeuvre-lès-Nancy
karen.fort@inist.fr

(2) XRCE, 6 Chemin de Maupertuis, 38240 Meylan
maud.ehrmann@xrce.xerox.com

(3) LIPN, Université Paris 13 & CNRS, 99 av. J.B. Clément, 93430
Villetaneuse
adeline.nazarenko@lipn.univ-paris13.fr

Cet article a été publié dans les actes de la 16ème Conférence sur le Traitement Automatique des Langues Naturelles 2009. La publication originale est disponible sur le site <http://www-lipn.univ-paris13.fr/taln09/>.

Résumé. La tâche, aujourd'hui considérée comme fondamentale, de reconnaissance d'entités nommées, présente des difficultés spécifiques en matière d'annotation. Nous les précisons ici, en les illustrant par des expériences d'annotation manuelle dans le domaine de la microbiologie. Ces problèmes nous amènent à reposer la question fondamentale de ce que les annotateurs doivent annoter et surtout, pour quoi faire. Nous identifions pour cela les applications nécessitant l'extraction d'entités nommées et, en fonction des besoins de ces applications, nous proposons de définir sémantiquement les éléments à annoter. Nous présentons ensuite un certain nombre de recommandations méthodologiques permettant d'assurer un cadre d'annotation cohérent et évaluable.

Abstract. Today, the named entity recognition task is considered as fundamental, but it involves some specific difficulties in terms of annotation. We list them here, with illustrations taken from manual annotation experiments in microbiology. Those issues lead us to ask the fundamental question of what the annotators should annotate and, even more important, for which purpose. We thus identify the applications using named entity recognition and, according to the real needs of those applications, we propose to semantically define the elements to annotate. Finally, we put forward a number of methodological recommendations to ensure a coherent and reliable annotation scheme.

Mots-clés : annotation, reconnaissance d'entités nommées.

Keywords : annotation, named entities extraction.

1 Introduction

Si l'extraction d'entités nommées (EN), apparue au milieu des années 1990 à la faveur des dernières conférences MUC¹, fait aujourd'hui figure d'incontournable en Traitement Automatique des Langues (TAL), l'annotation de corpus qui la sous-tend est encore peu étudiée en tant que telle. Les enjeux de l'annotation manuelle sont pourtant importants. Qu'il s'agisse de la performance des systèmes mis au point (à partir d'un travail de modélisation ou d'apprentissage automatique sur corpus), de l'évaluation de ces derniers, ou encore de la bonne réponse apportée à des besoins applicatifs, l'annotation de corpus est une composante fondamentale. Au cœur des divers processus constituant la reconnaissance d'entités nommées (REN), nous souhaitons examiner de plus près la problématique de l'annotation manuelle, laquelle conduit à s'interroger sur ce que sont les entités nommées et ce à quoi elles servent.

Nous présentons cette pratique établie qu'est aujourd'hui l'annotation d'EN (2) puis nous en détaillons les principales difficultés (3), aussi bien sur des textes de langue générale que de spécialité. Nous examinons ensuite les applications dans lesquelles les EN sont utilisées et en déduisons les différents types d'annotation (4). Enfin, nous proposons des recommandations méthodologiques permettant d'assurer un cadre d'annotation cohérent et évaluable (5).

2 Annotation d'entités nommées, une pratique établie

La reconnaissance d'entités nommées est une tâche bien connue : initiée il y a une vingtaine d'années à l'occasion des conférences américaines MUC, elle fut rapidement reprise lors d'autres campagnes d'évaluation, suscitant des travaux toujours plus nombreux. Sans revenir ici sur le "succès" de cette tâche (Nadeau & Sekine, 2007), on peut retracer son évolution selon trois directions principales. La première correspond à des travaux dans le domaine "général", avec la poursuite de la tâche définie par MUC pour d'autres langues que l'anglais, selon un jeu de catégories plus ou moins revisité et pour annoter des entités dans des corpus de nature journalistique essentiellement². La seconde direction concerne des travaux dans des domaines dits "de spécialité", avec la reconnaissance d'entités dans les domaines de la médecine, de la chimie ou de la microbiologie. Il fut ainsi proposé de reconnaître des noms de gènes, de protéines, etc. dans de la littérature spécialisée, lors des campagnes JNLPBA (Kim *et al.*, 2004) et BioCreAtIvE (Hirschman *et al.*, 2005). La dernière direction, transversale aux domaines général et de spécialité, correspond à des travaux sur la désambiguïsation : résolution de métonymie des EN dans SemEval2007 (Markert & Nissim, 2007) et désambiguïsation de noms de personnes (Artiles *et al.*, 2007)³.

A l'occasion de chacune de ces campagnes, des corpus furent constitués et annotés manuellement. De manière générale, ces corpus annotés servent à la mise au point d'outils d'annotation automatique. "Mise au point" est à entendre ici au sens large : il s'agit de décrire le plus précisément possible ce que les systèmes doivent faire pour guider le travail d'écriture des règles sur lesquelles ils reposent, pour apprendre automatiquement ces règles de fonctionnement ou des critères de décision et, enfin, pour évaluer les résultats obtenus en les confrontant à une analyse de référence. Le processus d'annotation met en jeu deux acteurs, un annotateur et un texte, et

¹Message Understanding Conferences, (MUC, 1995), (MUC, 1998)

²Voir les campagnes d'évaluation MET, IREX, CoNLL, ACE, ESTER et HAREM (Ehrmann, 2008, pp. 19-21).

³La campagne *Web People Search* fut par la suite rééditée, voir <http://nlp.uned.es/weps/>.

aboutit à une annotation du texte dont la qualité doit répondre à une *méthodologie* et supporter un *protocole d'évaluation*, et le contenu suivre un *guide d'annotation*.

Pour le domaine général, les campagnes MUC, CoNLL et ACE ont travaillé avec des corpus issus de la presse. Ces campagnes semblent avoir porté attention au processus d'annotation manuelle des EN, avec la rédaction de guides d'annotation et des calculs d'accord inter-annotateurs (mais non intra-annotateur), procédant par allers-retours entre le corpus à annoter et le guide d'annotation à réviser, mais des points d'hésitations ont perduré quant à l'annotation, dus principalement à “*different interpretations of vague portions of the guidelines*” (Sundheim, 1995) où à des phénomènes de superposition de sens (Doddington *et al.*, 2004). Dans les domaines de la biologie et de la biomedecine, des textes des bases de données de publications scientifiques (PubMed et MedLine⁴) ont été annotés : on observe que les guides d'annotation – quand ils existent⁵ – laissent des zones d'ombre quant à la manière d'annoter les entités et que peu d'études ont porté sur la qualité de l'annotation. Que ce soit pour les corpus GENIA (Kim *et al.*, 2003), PennBioIE (Kulick *et al.*, 2004) ou GENETAG (Tanabe *et al.*, 2005), aucun accord inter ou intra-annotateur n'est rapporté. A la fin de leur expérience d'annotation, les auteurs de (Tanabe *et al.*, 2005) constatent que “*a more detailed definition of a gene/protein name, as well as additional annotation rules, could improve interannotator agreement and help solve some of the tagging inconsistencies*”.

On observe par ailleurs des pratiques d'annotation manuelle et une réflexion méthodologique intéressantes dans d'autres domaine du TAL, notamment dans les travaux issus de la communauté parole (Gut & Bayerl, 2004). La campagne EVALDA MEDIA (Bonneau-Maynard *et al.*, 2005), par exemple, a mis en oeuvre une annotation manuelle de corpus en deux passes, la première sur un échantillon pour faire le point sur les éventuels désaccords entre annotateurs, et une seconde grandeur nature, après réajustement des guides d'annotation.

3 Difficultés rencontrées dans l'annotation

Pour les corpus de langue générale, un certain nombre de difficultés sont identifiées (Ehrmann, 2008). La première concerne le choix des catégories et la détermination de ce qu'elles recouvrent. En effet, au-delà de la triade “universelle” définie par MUC (PERSONNE, LIEU et ORGANISATION), l'inventaire des catégories à annoter est difficile à stabiliser et à définir. Prenons l'exemple de la catégorie PERSONNE : s'il est évident qu'un nom d'individu tel que *Lionel Jospin* est à annoter à l'aide de cette catégorie, que faut-il faire des *Kennedys*, de *Zorro*, des *Démocrates* ou de *St. Nicolas* ? Pour les autres catégories, il est également difficile de choisir la granularité des catégories et de déterminer ce qu'elles recouvrent vraiment. Un autre type de difficulté concerne la sélection des mentions à annoter ainsi que la délimitation des frontières des EN. A titre d'exemple, considérons l'EN “Barack Obama” et les diverses unités lexicales suivantes permettant d'y faire référence : *Barack Obama*, *Monsieur Obama*, *le Président des Etats-Unis*, *le nouveau président*, *il*. Faut-il annoter les noms propres uniquement, ou peut-on également considérer les descriptions définies permettant d'identifier cette personne, voire les pronoms qui, contextuellement, y renvoient ? Et que fait-on des différents attributs accompagnant l'EN (“monsieur” et “président”) ? De nombreux autres cas, plus ou moins compliqués (*la Maire du 7e arrondissement de Paris* et *Garde des Sceaux Rachida Dati*) se rencontrent en corpus. Dans

⁴Respectivement : <http://www.ncbi.nlm.nih.gov/pubmed/> et <http://medline.cos.com/>

⁵Certaines campagnes comme JNLPBA et BioCreAtIvE I n'en ont pas fourni.

le même ordre d'idée, des phénomènes de coordination et d'imbrication peuvent poser problème aux annotateurs (une ou plusieurs entités pour *Bill et Hillary Clinton* et *l'Université de Corte* ?). Enfin, une dernière difficulté résulte de phénomènes de pluralité référentielle, avec des EN homonymes (*Orange* ville et *Orange* compagnie) et des glissements métonymiques, parfois difficiles à distinguer (*France* en tant que lieu géographique, gouvernement ou équipe sportive). Même si elles sont surmontables grâce aux guides d'annotation, ces difficultés entraînent un coût supplémentaire et une baisse de qualité de l'annotation.

Notre expérience en microbiologie montre que ces difficultés sont plus aiguës encore dans les langues de spécialité. Une expérience d'annotation a été réalisée à l'INIST pour le programme Quaero, en collaboration avec l'équipe MIG. Un corpus anglais de 499 notices PubMed (titres et résumés, soit environ 110 000 "mots"), pré-annotées par application d'un dictionnaire et d'un anti-dictionnaire, a été fourni à deux experts de l'INIST, dont le travail a consisté en une révision de ces pré-annotations. La principale difficulté rencontrée a concerné la distinction qui était demandée entre noms propres et noms communs, la limite morphologique entre les deux étant peu marquée dans ces domaines où les noms communs sont souvent reclassés comme "noms propres", comme en atteste la présence de ces noms dans des nomenclatures ("small, acid-soluble spore protein A" est ici un cas extrême) ou les phénomènes d'acronymisation (on trouve par exemple "across the outer membrane (OM)"). Dans ces cas, la consigne donnée aux annotateurs était de se référer à des listes d'autorité, telle que Swiss-Prot⁶, ce qui entraîne une perte de temps conséquente. La délimitation des frontières des éléments à annoter a elle-aussi soulevé de nombreux questionnements, les annotateurs se demandant ce qu'il fallait inclure ou non dans le segment annoté. On peut ainsi choisir d'annoter "nifh messenger RNA" si on considère que la mention de l'état "messenger RNA" entre dans la détermination de la référence, ou seulement "nifh", si on considère que le nom propre suffit à construire la référence. Le typage sémantique choisi a aussi posé problème aux annotateurs, notamment pour les éléments génétiques mobiles, comme les plasmides ou les transposons. En effet, ceux-ci devaient être annotés dans les *taxons* et non dans les *genes* alors que ce sont des fragments d'ADN, donc des parties de génome. Une directive particulièrement perturbante pour les annotateurs a été d'annoter l'acronyme "KGFR" comme nom propre et sa forme développée "keratinocyte growth factor receptor" comme nom commun. Ce type de consigne, préconisée au départ pour entraîner plus efficacement les outils de REN (Nédellec *et al.*, 2006), est difficile à appréhender et aurait dû être mieux documentée.

Ces problèmes se traduisent par un coût d'annotation élevé, des guides d'annotation de taille trop grande par rapport au corpus et trop d'hésitations de la part des annotateurs, ce qui induit des incohérences et une qualité moindre de l'annotation. Cette expérience nous a ainsi permis de reposer la question de ce que les annotateurs doivent annoter et surtout, pour quoi faire.

4 Annoter quoi ?

Pour mieux comprendre "quoi annoter dans quel texte", nous revenons sur les critères linguistiques qui permettent de définir la notion d'EN : l'importance de ces critères varie d'une application à l'autre et les annotations résultantes en dépendent.

⁶<http://www.expasy.org/sprot/>

4.1 Différents critères définitoires

(Ehrmann, 2008) propose une analyse linguistique de la notion d'EN qu'elle présente comme une "création" du TAL. Dans ce qui suit, nous reprenons la distinction introduite dans (LDC, 2004) : les EN sont des "mentions" qui renvoient à des "entités" du domaine, ces mentions pouvant relever de différentes catégories linguistiques : des noms propres ("Rabelais"), mais aussi les pronoms ("il"), et plus largement des descriptions définies ("le père de Gargantua"). On peut identifier plusieurs critères définitoires des EN.

Unicité référentielle L'une des caractéristiques principales des noms propres est leur fonctionnement référentiel : un nom propre renvoie à une entité référentielle unique, même si cette unicité est contextuelle. A la différence de (Poibeau, 2005), nous considérons que cette propriété est essentielle dans l'utilisation que fait le TAL des EN.

Autonomie référentielle Les EN sont de surcroît autonomes du point de vue référentiel. C'est évident dans le cas du nom propre qui permet à lui seul l'identification du référent, tout au moins dans une situation de communication donnée (*Eurotunnel*). Le cas des descriptions définies (*l'opérateur du tunnel sous la Manche*) est un peu différent : si elles suffisent à identifier le référent, c'est par le truchement de connaissances extérieures.

Stabilité dénominative Les noms propres sont également des dénominations stables. Même s'il y a des variations (*Angela Merkel/Mme Merkel/A. Merkel*), elles sont plus régulières et moins nombreuses que pour les autres syntagmes nominaux⁷.

Relativité référentielle L'interprétation se fait toujours relativement à un modèle du domaine, qui peut être implicite dans les cas simples (on suppose une connaissance partagée de ce qu'est une personne ou un pays) mais qui doit être explicité dès que la diversité des entités à prendre en compte s'accroît (il faut au moins une typologie pour les catégoriser).

4.2 Différentes visées applicatives

La REN voit le jour dans un cadre d'extraction d'information où il s'agit d'identifier les actants de certaines situations (des attentats aux interactions géniques), cette situation étant décrite par un formulaire à instancier avec les informations extraites du texte. La sémantique sous-jacente est clairement référentielle : on identifie dans le texte les "mentions" des "entités" qui jouent un rôle dans les situations considérées (LDC, 2004). Dans les systèmes de questions/réponses, les EN jouent le même rôle mais on a souvent recours à des typologies plus fines⁸.

Les EN sont également utilisées comme "descripteurs" dans de nombreuses applications d'indexation. On a des index de noms propres dans certains ouvrages et moteurs de recherche et on sait qu'une importante proportion des requêtes adressées aux moteurs de recherche sont des EN. On surligne les EN pour aider la lecture ou la navigation dans de gros volumes documentaires. On utilise enfin la REN pour construire et mettre à jour des nomenclatures (Tran & Maurel, 2006) utilisées pour l'indexation, la traduction, etc. Dans ce deuxième type d'application aussi, la sémantique sous-jacente est référentielle : si les EN sont de bons descripteurs, c'est qu'elles fonctionnent comme des ancrs référentielles qui permettent de situer ce à quoi le texte fait

⁷Ceci explique *a contrario* l'importance du repérage de la synonymie dans les domaines où les dénominations sont peu stables (la génomique, par exemple).

⁸Ce qui pose des problèmes de désambiguïsation et de résolution des métonymies.

référence⁹. En revanche, on s'intéresse exclusivement aux mentions de type nom propre.

Troisième type d'applications, les EN sont utilisées pour l'intégration de données. Cela concerne à la fois l'analyse des bases documentaires (suivi de thèmes, découverte de communautés de pratiques (Li & Liu, 2005) et l'articulation des documents avec d'autres sources de connaissances (bases de données, bases d'images, etc.) pour interroger les unes et les autres de manière homogène et naviguer facilement de l'une à l'autre (Dragos & Nazarenko, 2009). On s'appuie alors sur les EN référencées dans les bases de données pour établir des liens entre les différentes sources. Pour ces tâches d'intégration, on se contente des EN qui figurent dans les nomenclatures, l'objectif étant d'articuler les sources entre elles et pas d'en décrire le contenu : il suffit de savoir qu'une dépêche parle de *Nelson Mandela*, inutile de trouver toutes les mentions qui en sont faites dans le texte.

L'anonymisation des documents est un quatrième champ d'application (Plamondon *et al.*, 2004). On veut éviter qu'on puisse identifier des entités (des personnes, notamment) à partir des mentions qui en sont faites dans le texte. Il faut repérer toutes les formes de mentions (*la ville qui a reçu la première bombe atomique*) et pas seulement les noms propres.

Dernière famille d'applications, la REN est utilisée pour "peupler" des ontologies. Il s'agit alors de modéliser un domaine, et le modèle formel des ontologies impose de distinguer les entités du domaine, qui sont représentées comme des instances, des concepts ou classes auxquelles ces instances se rattachent. La REN est alors utilisée pour enrichir la structure conceptuelle avec des instances de concepts ou de rôles (relations entre instances) (Amardeilh *et al.*, 2005). Selon les cas, on privilégie les entités nommées ou on s'intéresse à toutes les mentions d'entités. Dans le premier cas, l'ontologie résultante peut être vue comme une nomenclature hiérarchisée ou comme un thesaurus formalisé. Dans le second, on cherche à modéliser un domaine en identifiant les entités qui le composent et les relations qu'elles entretiennent. Dans les deux cas, le typage des EN est essentiel parce qu'on doit relier l'instance nommée à un concept de l'ontologie.

4.3 Des perspectives d'annotation différentes

Les critères définitoires cités en 4.1 ne jouent pas tous le même rôle pour toutes les applications. Dans certains cas (indexation et intégration de connaissances), on s'intéresse à des entités référentielles qui sont désignées par des descripteurs stables et non ambigus. Ce sont donc les EN de type noms propres ou "catalogables" qui sont à retenir et il est important de les normaliser pour s'affranchir des variations qui peuvent apparaître malgré leur stabilité référentielle. Pour ce type d'application, l'essentiel n'est pas de repérer toutes les mentions de telle entité dans un document mais de repérer que ce document mentionne telle entité. Il faut donc privilégier la précision sur le rappel de l'annotation.

A l'autre extrême, on trouve les tâches d'extraction d'information et de modélisation du domaine où toutes les mentions sont importantes à repérer, y compris celles qui sont des descriptions définies (il faut d'ailleurs repérer des relations de coréférence entre les mentions qui ne sont pas suffisamment autonomes référentiellement). La figure 1 montre l'impact de ces deux perspectives d'annotation sur les résultats de l'annotation d'un tout petit exemple¹⁰.

⁹Deux noms propres (*AZF* et *septembre 2001*) suffisent souvent à faire comprendre de quoi parle une dépêche !

¹⁰<http://www.ncbi.nlm.nih.gov/pubmed/1331532>

<p>ANNOTATION D'INDEXATION ; types <i>gene</i> et <i>protein</i> We conclude that <gene>3CDproM</gene> can process both structural and nonstructural precursors of the <EukVirus uncertainty-type="too-generic">poliovirus polyprotein</EukVirus> and that it is active against a synthetic peptide substrate.</p> <p>ANNOTATION DE MODÉLISATION ; types <i>taxon</i>, <i>gene</i> et <i>protein</i> We conclude that <EukVirus>3CDproM</EukVirus> can process both structural and nonstructural precursors of the <EukVirus uncertainty-type="too-generic"><taxon>poliovirus</taxon> polyprotein</EukVirus> and that <EukVirus>it</EukVirus> is active against a synthetic peptide substrate.</p>
--

FIG. 1 – Exemple d'annotation en biologie. La première annotation est moins riche que la seconde qui considère plus de types sémantiques (*taxon*) avec une granularité plus fine (*EukVirus* is a subtype of *gene*), ce qui introduit des enclassements de balises. Par ailleurs, toutes les mentions sont annotées dans la seconde annotation alors que seuls les assimilés noms propres le sont dans la première.

Comme il est impossible de repérer les mentions de toutes les entités référentielles, le modèle du domaine détermine quelles sont les entités "d'intérêt" et la limite entre ce qui doit ou non être annoté. Illustrons ce point sur un exemple. Quand un directeur des ressources humaines s'intéresse aux grilles salariales de son organisation, il raisonne sur des catégories de personnels et pas sur les personnes physiques que sont les employés. Cela se reflète dans le modèle du domaine : les différentes catégories de personnes (techniciens, ingénieurs, etc.) sont modélisées comme des instances rattachées au concept CAT-DE-PERSONNEL et les personnes physiques ne sont pas représentées. A l'inverse, quand il s'occupe des fiches de paye et de la progression des employés, il s'intéresse aux individus. Dans ce cas, le modèle doit considérer les personnes comme instances et les catégories de personnels comme des concepts. Le même problème se rencontre en biologie où la mention du gène G peut renvoyer au gène G de l'espèce E, au gène G d'un individu de l'espèce E ou à un gène particulier parmi les gènes G de cet individu.

Modéliser suppose de faire des choix explicites là où les textes peuvent être flous et mêler les points de vue. Il est donc impossible d'annoter les EN d'un texte indépendamment d'un modèle de référence. Dans le cas de l'expérience décrite plus haut, le modèle était, comme souvent, simplement décrit par une liste de concepts : il fallait nommer les gènes et protéines, mais aussi leurs familles, compositions, et composants. Par ailleurs, comme la REN doit servir à modéliser le réseau d'interactions entre gènes, il était essentiel de repérer toutes les mentions des entités considérées, chacune pouvant contribuer à enrichir sa modélisation.

5 Recommandations méthodologiques

Annoter est difficile. Il faut donc guider le travail des annotateurs, ce qui passe par des guides et des outils d'annotation, mais aussi par l'évaluation de la qualité des annotations.

Guides d'annotations Comme le type d'annotation à produire dépend de ce qu'on cherche à annoter et de ce à quoi cette annotation doit servir, il est essentiel de fournir aux annotateurs des guides (ou conventions) d'annotation qui indiquent ce qu'il faut annoter plutôt que comment annoter. Trop souvent en effet, les contraintes de faisabilité prennent le pas sur les critères sémantiques¹¹ ce qui brouille l'objectif pour les annotateurs. Par ailleurs, il est important de prendre la mesure de la tâche dans toute sa complexité sans exclure *a priori* ce qui serait douteux ou trop difficile à reproduire automatiquement. C'est même l'intérêt de l'annotation manuelle que de donner une idée de l'ampleur de la tâche d'annotation.

¹¹"In [src homology 2 and 3], it seems excessive to require an NER program to recognize the entire fragment, however, 3 alone is not a valid gene name." (Tanabe *et al.*, 2005).

Il faut donner aux annotateurs une vision claire de l'application visée. Cette vision doit s'appuyer sur un modèle de référence explicite, du type de celui donné dans la campagne GENIA, avec des définitions précises et des explications sur les choix méthodologiques réalisés (catégories, typage sémantique, etc.). Des exemples peuvent être ajoutés à titre d'illustration mais ils ne doivent pas se substituer à la définition des objectifs. Ces informations permettent de limiter les incompréhensions, mais aussi de responsabiliser et de motiver les annotateurs en leur donnant accès à la logique sous-jacente. On entre ainsi dans une démarche didactique, permettant de passer d'un "rapport de père à un rapport de pair" (Akrich & Boullier, 1991), ce qui est d'autant plus nécessaire que l'annotation porte sur des corpus spécialisés, les annotateurs étant des experts à qui on a intérêt à donner la plus grande autonomie possible. Dans certains cas, la tâche d'annotation étant trop complexe, on doit se restreindre à une annotation exploitable pour l'apprentissage automatique – en se limitant par exemple aux noms propres – mais cela doit être fait de manière explicite pour que les annotateurs aient une vue claire des choix à faire.

Outils d'annotation Le travail d'annotation doit être outillé. On utilise des dispositifs techniques pour annoter les textes mais aussi pour préparer le travail d'annotation par une pré-annotation (projection d'un dictionnaire, par exemple). Ces outils sont cependant à manier avec précaution du fait des biais qu'ils introduisent. Dans l'expérience que nous avons menée, le corpus d'annotation avait été pré-annoté par projection d'un dictionnaire de noms de gènes et de protéines, pour alléger le travail des annotateurs, mais cette pré-annotation a influencé l'annotation. Elle a introduit un biais en faveur de la correction des pré-annotations, au détriment de la recherche de nouvelles EN. Une solution consiste à procéder en deux temps, en étiquetant à la main un échantillon du corpus, puis en le pré-annotant pour comparer les résultats et évaluer le biais introduit. On peut ensuite faire une annotation à plus grande échelle, précédée par une pré-annotation dont on connaîtra cette fois les biais.

En ce qui concerne le processus d'annotation lui-même, s'il existe aujourd'hui de nombreux outils d'aide à l'annotation manuelle, peu sont effectivement disponibles, c'est-à-dire gratuits, téléchargeables et utilisables. On peut citer, entre autres, Callisto, MMAX2, Knowtator ou encore Cadixe¹² qui a été utilisé dans notre expérience. Les fonctionnalités et l'expressivité du langage d'annotation doivent être adaptées à la tâche d'annotation visée : selon les cas, il faut typer des segments de textes ou les mettre en relation, on a des annotations concurrentes, disjointes ou superposables, etc. Dans notre expérience en microbiologie, par exemple, le fait de ne pas pouvoir annoter les coordinations d'EN a posé problème. Les critères ergonomiques sont également importants : on sait que les fonctionnalités difficiles d'accès ne sont pas utilisées, ce qui biaise là aussi les résultats. Dans notre cas, les annotateurs ont souvent négligé de mentionner leur incertitude parce qu'ajouter l'attribut correspondant à leurs annotations était malaisé.

Évaluation de l'annotation Il est important de mesurer la qualité de l'annotation. (Gut & Bayerl, 2004) distingue l'accord inter-annotateur, qui permet de mesurer la stabilité de l'annotation, et l'accord intra-annotateur, qui donne une indication de la reproductibilité de l'annotation. S'il est important de calculer l'accord inter-annotateur, il n'est pas nécessaire de le réaliser sur tout le corpus, ne serait-ce que pour des raisons de coût. Il est en revanche conseillé de le calculer tôt, afin d'identifier les problèmes rapidement et de modifier l'annotation en conséquence. Il en va de même en ce qui concerne l'accord intra-annotateur.

Un autre moyen d'évaluation consiste à faire ajouter à l'annotateur, si nécessaire, une note d'incertitude, de préférence typée, sur ses annotations. En l'absence de calcul de l'accord inter ou

¹²respectivement : <http://callisto.mitre.org/>, <http://mmax2.sourceforge.net/>, <http://knowtator.sourceforge.net/>, <http://caderige.imag.fr/Cadixe/>

intra-annotateur pour cette annotation nous avons ainsi fait annoter les incertitudes des annotateurs *a posteriori* sur un sous-ensemble du corpus. Pour s'assurer de la représentativité de l'échantillon, les annotateurs ont identifié une typologie des fichiers du corpus (six "types"). Nous avons ensuite extrait 25 fichiers de types variés parmi les 499 notices du corpus (soit 5%). L'objectif de cette seconde validation étant d'évaluer la confiance des annotateurs en leur validation, nous avons choisi de profiter du tag *uncertainty* proposé dans Cadix et de l'utiliser de manière systématique en cas de doute. Nous avons donc redéfini les types d'*uncertainty* en fonction de l'expérience des annotateurs. Cette évaluation n'a nécessité que quelques heures de travail et nous a permis de mieux qualifier et quantifier leurs doutes. Au final, sur 555 tags, les annotateurs ont déclaré 113 *uncertainty*, soit environ 20% des tags. On observe que plus de 75% des incertitudes concernent les noms communs de type *bacteria*, et que ces incertitudes sont très largement (77%) liées à une difficulté à distinguer les noms communs des noms propres.

Plus généralement, pour s'assurer à la fois de la qualité du guide d'annotation et des possibilités d'évaluation, une bonne méthode consiste, en tout début de projet, à faire travailler les annotateurs chacun de leur côté sur un échantillon du corpus. Cela permet d'identifier rapidement les désaccords, puis de les trancher, soit en faisant intervenir un autre expert, soit par concertation des annotateurs. Ces décisions sont ensuite reportées dans le guide d'annotation.

6 Conclusion et perspectives

Au final, une démarche rigoureuse et efficace d'annotation d'EN en corpus doit prêter attention principalement à deux points. Au regard du contenu tout d'abord, il importe de se focaliser non pas tant sur *comment* annoter, mais sur *quoi* annoter, en fonction de l'application visée. Nous avons vu que chaque famille d'application spécifie un certain nombre de critères linguistiques, et que les mentions à annoter diffèrent relativement à ces derniers. Une fois spécifié ce qu'il faut annoter, il faut être prudent en termes de méthodologie et envisager dès le départ l'évaluation de l'annotation. Il est avantageux, du point de vue de la qualité de l'annotation, de réaliser un ou plusieurs galops d'essai, afin d'ajuster certains éléments du guide d'annotation et d'évaluer les biais éventuels introduits par l'outil d'annotation et/ou une mauvaise compréhension de la tâche par les annotateurs. Relativement aux enjeux et aux difficultés de l'annotation des EN en corpus, l'attention portée à la spécification du contenu et au respect d'une méthodologie d'annotation permet d'assurer un cadre de travail stable et cohérent à chacune des étapes de la tâche de REN. Nous comptons appliquer cette méthodologie dans le cadre des campagnes d'annotation de Quaero, en pharmacologie et en économie. Couvrant la terminologie et l'extraction de relations sémantiques, ces campagnes dépassent largement le cadre des EN, nous allons donc élargir notre méthode à ces applications.

Remerciements

Ce travail a été réalisé en partie dans le cadre du programme Quaero¹³, financé par OSEO, agence nationale de valorisation de la recherche. Nous en remercions les participants, en particulier l'équipe MIG de l'INRA. Nous remercions également F. Tisserand et B. Taliércio, les annotateurs experts de l'INIST.

¹³<http://www.quaero.org>

Références

- (1995). *Proc. of the 6th Message Understanding Conference*. Morgan Kaufmann Publishers.
- (1998). *Proc. of the 7th Message Understanding Conference*. Morgan Kaufmann Publishers.
- AKRICH M. & BOULLIER D. (1991). *Savoir faire et pouvoir transmettre*, chapitre Le mode d'emploi, genèse, forme et usage, p. 113–131. éd. de la MSH (coll. Ethnologie de la France).
- AMARDEILH F., LAUBLET P. & MINEL J.-L. (2005). Document annotation and ontology population from linguistic extractions. In *Proc. of K-CAP'05*, p. 161–168, New York : ACM.
- ARTILES J., GONZALO J. & SEKINE S. (2007). The semeval-2007 WePS evaluation : establishing a benchmark for the web people search task. In *Proc. of SemEval, ACL*, Prague.
- BONNEAU-MAYNARD H., ROSSET S., AYACHE C., KUHN A. & MOSTEFA D. (2005). Semantic annotation of the french media dialog corpus. In *InterSpeech*, Lisbonne, Portugal.
- DODDINGTON G., MITCHELL A., PRZYBOCKI M., RAMSHAW L., STRASSEL S. & WEISCHEDEL R. (2004). The ACE program tasks, data, and evaluation. In *Proc. of LREC'04*, Lisbonne, Portugal.
- DRAGOS V. & NAZARENKO A. (2009). Towards a semantic model to enhance knowledge sharing and discovery in organic chemistry. In *Proc. of the IADIS IS'09*, Barcelone, Espagne.
- EHRMANN M. (2008). *Les entités nommées, de la linguistique au TAL : statut théorique et méthodes de désambiguïsation*. PhD thesis, Université Paris 7.
- GUT U. & BAYERL P. S. (2004). Measuring the reliability of manual annotations of speech corpora. In *Proc. of Speech Prosody*, p. 565–568, Nara, Japon.
- HIRSCHMAN L., YEH A., BLASCHKE C. & VALENCIA A. (2005). Overview of biocreative : critical assessment of information extraction for biology. *BMC Bioinformatics*, **6**(1).
- KIM J.-D., OHTA T., TATEISI Y. & TSUJII J. (2003). Genia corpus—a semantically annotated corpus for bio-textmining. *Bioinformatics*, **19**, 180–182.
- KIM J.-D., OHTA T., TSURUOKA Y., TATEISI Y. & COLLIER N. (2004). Introduction to the bio-entity recognition task at JNLPBA. In *Proc. of JNLPBA COLING'04 Workshop*, p. 70–75.
- KULICK S., BIES A., LIBERMAN M., MANDEL M., MCDONALD R., PALMER M., SCHEIN A. & UNGAR L. (2004). Integrated annotation for biomedical information extraction. In *HLT-NAACL 2004 Workshop : Biolink : ACL*.
- LDC (2004). *ACE (Automatic Content Extraction) English Annotation Guidelines for Entities*. Livrable version 5.6.1 2005.05.23, Linguistic Data Consortium.
- LI X. & LIU B. (2005). Mining community structure of named entities from free text. In *Proc. of CIKM'05*, p. 275–276, New York, NY, USA : ACM Press.
- MARKERT K. & NISSIM M. (2007). Semeval-2007 task 08 : Metonymy resolution at semeval-2007. In *Proc. of SemEval, ACL*, Prague.
- NADEAU D. & SEKINE S. (2007). A survey of named entity recognition and classification. *Linguisticae Investigaciones*, **30**(1), 3–26.
- NÉDELLEC C., BESSIÈRES P., BOSSY R., KOTOUJANSKY A. & MANINE A.-P. (2006). Annotation guidelines for machine learning-based named entity recognition in microbiology. In *Proc. of the Data and text mining in integrative biology workshop*, p. 40–54, Berlin.
- PLAMONDON L., LAPALME G. & PELLETIER F. (2004). Anonymisation de décisions de justice. In *Proc. of TALN'04*, p. 367–376, Fès, Maroc.

Vers une méthodologie d'annotation des entités nommées en corpus ?

POIBEAU T. (2005). Sur le statut référentiel des entités nommées. In *Proc. of TALN'05*, p. 173–182, Dourdan, France.

SUNDHEIM B. (1995). Overview of results of the MUC-6 evaluation. In (MUC, 1995).

TANABE L., XIE N., THOM L., MATTEN W. & WILBUR J. (2005). Genetag : a tagged corpus for gene/protein named entity recognition. *Bioinformatics*, **6**.

TRAN M. & MAUREL D. (2006). Prolexbase : un dictionnaire relationnel multilingue de noms propres. *TAL*, **47**(3), 115–139.