Recital 2008, Avignon, 9-13 juin 2008

Un système d'annotation des entités nommées du type personne pour la résolution de la référence

Elzbieta GRYGLICKA

(1) Laboratoire Mathématiques et Aide à la Décision – Thales R&T, RD 128 Palaiseau
elzbieta.gryglicka@thalesgroup.com
(2) ALPAGE –Paris 7, 30 rue Château des Rentiers, 75 013 Paris

Résumé Dans cet article nous présentons notre démarche pour l'annotation des expressions référentielles désignant les personnes et son utilisation pour la résolution partielle de la référence. Les choix effectués dans notre implémentation s'inspirent des travaux récents dans le domaine de l'extraction d'information et plus particulièrement de la reconnaissance des entités nommées. Nous utilisons les grammaires locales dans le but d'annoter les entités nommées du type Personne et pour construire, à partir des annotations produites, une base de connaissances extra-linguistiques. Les informations acquises par ce procédé sont ensuite utilisées pour implémenter une méthode de la résolution de la référence pour les syntagmes nominaux coréférentiels.

Abstract The aim of this paper is to describe our approach for annotating of the referential mentions that refer to the entities which are the instances of Person. Our method is inspired by the recent work in information extraction and particularly the named entities recognition and classification task. Local grammars are used to identify this category of named entities and to generate an extra-linguistic knowledge base which is further used for the process of reference resolution.

Mots-clés : entités nommées, annotation, grammaires locales, Nooj, base de connaissances

Keywords: named entities, entity recongnition, local grammars, Nooj, knowledge base

1 Introduction

La notion d'entité nommée (EN) est intimement liée à la tâche d'extraction d'information définie dans le cadre de Message Understanding Conferences (MUC) dans les années 90 (cf. MUC – 6, 1995, MUC – 7, 1998). Cette série de conférences a permis l'émergence de la tâche de la reconnaissance et de l'annotation de ces entités, ainsi que la définition de leur hiérarchie de base. Elle a également mis en avant leur importance pour l'extraction d'informations et la résolution des anaphores. Depuis, un grand nombre d'outils, comme par exemple les systèmes Question/Réponse (Gaizauskas et al., 2005), ou les applications de veille stratégique (Saggion et al., 2007), utilisent cette technologie devenue entre temps mature. La reconnaissance des entités nommées consiste en effet en deux sous-tâches : la reconnaissance des noms propres et les expressions numériques dans le texte brut, et la classification des formes repérées. L'application décrite dans le présent article s'appuie sur une interprétation spécifique de ces expressions. Nous proposons une stratégie d'annotation qui élargit les motifs annotées à des syntagmes nominaux étendus, par exemple, dans une proposition comme: *Mais paradoxalement,* ***le ministre d'État chargé de la Communication, Guillaume Soro,*** *qui aurait dû présenter devant l'assemblée nationale le 20 octobre, le projet de loi (...)* nous annotons le syntagme entier ***le ministre d'État chargé de la Communication, Guillaume Soro***. Compte tenu de la diversité des catégories des entités nommées et des structures qui peuvent les contenir, l'approche présentée se restreint uniquement à des expressions du type Personne, qui correspondent à des individus dont le nom propre a été cité dans le texte. Les noms propres présents dans le corpus ayant été préalablement extraits, les annotations produites par notre module nous servent principalement à acquérir des informations. Ces informations sont ensuite structurées et modélisés sous la forme d'une base de connaissances XML. Notre objectif inclut également l'annotation des expressions se référant à des personnes et qui ne sont pas accompagnées par un nom propre, et qui correspondent, selon les cas, aux descriptions définies autonomes, ou descriptions coréférentielles dans (Gardent, Manuélian, 2005). Une approche de la résolution de la référence de ces expressions constitue donc un autre aspect du travail présenté ici. La motivation première qui a inspiré le travail présenté dans cet article est la nécessité d'améliorer une application d'extraction de relations sémantiques existante en introduisant comme une étape de prétraitement, la résolution de la coréférence (expressions purement anaphoriques, mais aussi des expressions coréférentielles dont la référence n'est pas spécifiée). Mais la méthode utilisée nous rapproche des applications créées pour les besoins de peuplement d'ontologies (OPTM, Ontology Population from Textual Mentions), comme défini dans (Magnini et al., 2006).

Dans la première partie de cet article nous présentons brièvement la problématique des entités nommées, ses principaux aspects et les travaux récents qui nous ont amenés à proposer notre approche. La seconde partie est consacrée à la description de l'algorithme de notre application. Il est basé sur l'annotation des syntagmes nominaux faisant référence à une personne à l'aide de grammaires locales. Dans cette partie nous expliquons aussi de quelle manière nous représentons les informations acquises et le procédé de la résolution de la référence mis en œuvre. Les résultats de cette approche sont analysés dans la troisième partie de notre exposé.

2 Reconnaissance des entités nommées et ses enjeux actuels

La définition classique d'une entité nommée correspond à celle donnée par (Poibeau, 2003) : *Entité nommée : ensemble des noms de personnes, d'entreprises et de lieux présents dans un texte donné. On associe souvent à ces éléments d'autres syntagmes repérables par des*

grammaires locales comme les dates, les unités monétaires, les pourcentages repérables par les mêmes techniques à base de grammaires locales.

L'importance des entités nommées est principalement le résultat de leur caractère purement référentiel, qui leur permet de jouer un rôle fondamental dans les applications à visée industrielle (comme la veille stratégique, le résumé automatique). Depuis que la communauté TAL a pris conscience de ce fait, un grand effort de travail a été consacré à leur bonne catégorisation. Certains auteurs font même l'impasse sur la reconnaissance de noms propres : (Maynard et al., 2001) en utilisant des listes de ces formes. La hiérarchie de base a été proposée pendant les conférences (MUC – 6) : les entités du type ENAMEX correspondent à des noms propres de sous-catégories : Personne, Organisation et Lieu ; TIMEX concerne les expressions temporelles, et NUMEX, les autres expressions numériques comme poids, mesures et valeurs monétaires. Beaucoup de travaux proposent des améliorations pour ce classement. Certains correspondent à une perception plus large de ce phénomène, comme par exemple (Ferret et al., 2000) qui propose une catégorie supplémentaire: Fonction, qui n'est pas un nom propre, ni une expression numérique. Les travaux de (Sekine, 2004), sous l'influence de ACE (Automatic Content Extraction) développent même une hiérarchie d'environ 200 catégories, qui d'une part, augmente le nombre de types d'entités reconnues, en ajoutant par exemple : Théorie, Crime, Minerai. Et propose, d'autre part, une classification plus précise pour les types existants. D'autres auteurs (Ehrmann, Jacquet, 2006) abandonnent même ce jeu d'étiquettes classiques et proposent un système d'annotation basé sur les relations syntaxiques qui produit des cliques d'étiquettes correspondant aux différents contextes syntaxiques d'une entité nommée.

Les travaux cités correspondent à une prise de conscience de la polysémie des expressions appelées entités nommées, exprimée dans (Poibeau, 2005). Dans notre démarche, qui réduit fortement le nombre de phénomènes étudiés, nous pouvons nous concentrer sur quelques cas de polysémie qui concernent particulièrement les noms propres des personnes. Cela nous permet d'annoter uniquement les expressions qui font référence à des entités du type Personne, et à éviter les phénomènes de transfert de sens, afin de ne pas reconnaître les motifs comme par exemple *le prix Félix Houphouët Boigny*. Une approche semblable est décrite dans (Magnini et al., 2006), dont les auteurs définissent une tâche de peuplement d'ontologies restreinte à des mentions référant à la catégorie Personne. Le travail présenté dans cet article correspond à une volonté d'étudier la variété des formes qui peuvent correspondre à un seul type d'entité. Une étude de corpus a précédé la création des grammaires locales. Dans le même esprit, pour l'évaluation de notre approche nous avons procédé à une relevé complet des expressions désignant les personnes. Une autre problématique se rattachant à celle des entités nommées est celle de la référence. Malgré une grande variabilité des formes et malgré un certain flou dans la définition de cette notion (certains auteurs restreignent les entités nommées à des noms propres et des expressions numériques, d'autres y ajoutent des termes spécifiques pour un domaine, etc.) le fait que ces entités font référence, soit à des concepts, soit à des objets réels du monde, ne semble pas sujet à la discussion. Notamment, le phénomène de la coréférence interdocumentaire pour les noms propres de personnes, a été pris en compte par (Saggion et al., 2007), qui propose un algorithme de clustering afin de délimiter les groupes de documents qui mentionnent la même personne. Dans une autre perspective, notre approche permet de rendre explicite le lien coréférentiel existant entre une entité nommée de catégorie Personne et le concept de Fonction (*le président de la République française*). C'est précisément ce lien qui est utilisé dans l'algorithme de la résolution de la référence implémenté.

3 Présentation de l'application

Notre idée de départ a été inspirée par (Poibeau, 2005), qui propose déjà l'utilisation de l'extraction d'information dans le but de résoudre les anaphores nominales. Notre méthode est basée sur l'utilisation des patrons lexico-syntaxiques couplés avec les ressources lexicales disponibles. La pierre angulaire de notre travail est l'idée que nous pouvons retrouver dans les textes un certain nombre de connaissances sur le monde extralinguistique. De nombreux travaux proposent des techniques d'apprentissage à partir des sources de connaissances externes comme WordNet (Poesio et al., 1997, Meyer and Dale, 2002), ou même comme des informations disponibles sur Internet, la Wikipédia par exemple (Kazama, Torisawa, 2007). Ces approches ont été développées principalement pour traiter les textes en langue anglaise, et aucune ressource disponible pour le français ne permet d'obtenir la couverture en relations sémantiques égale à celle de WordNet. Dans le cadre de notre travail appliqué sur le corpus en français, nous proposons une démarche dans la lignée de travaux de (Hearst, 1992). Cette approche est basée sur l'extraction d'informations à partir de corpus, perçu comme une source de connaissances immédiatement disponibles et qui a comme principal avantage d'éviter la surgénération des entités supplémentaires (homonymes) présentes par exemple sur Internet. Les informations extraites à partir du corpus sont ensuite analysées et présentées sous la forme d'une base de connaissances qui permet de décrire les différents types de liens entre les entités de catégorie Personne et les autres types d'entités, principalement les Organisations (partis politiques, structures gouvernementales, lieux géographiques). Notre chaîne de traitement se décompose en trois étapes, chacune faisant l'objet d'une section :

- L'annotation des syntagmes nominaux accompagnés par un nom propre. Les formes visées pendant cette phase de traitement sont les syntagmes nominaux définis, les appositions et les structures prédicatives, qui constituent pour nous une source d'informations sur les personnes.

- L'extraction du contenu des annotations produites à l'étape précédente, sa transformation et sa structuration sous la forme d'un fichier XML, qui sert de base de connaissances pour les étapes ultérieures.

- La mise en œuvre de la stratégie de la résolution partielle de la référence. Elle débute par une autre phase d'annotation. Les formes visées dans cette étape sont, en plus de celles reconnues précédemment : les noms propres seuls et les syntagmes nominaux non accompagnés par un nom propre. Cette stratégie nous permet de reconstituer des chaînes de coréférence partielles (à l'exclusion de formes pronominales) et, en utilisant les informations contenues dans notre base de connaissances, d'attribuer un nom à des expressions du second type.

3.1 Stratégie d'annotation des entités nommées

Comme nous l'avons déjà précisé, nous avons fait le choix d'annoter les entités nommées de la catégorie Personne en incluant les appositions et les prédications qui les accompagnent. Ceci diffère de la pratique courante pour les applications de catégorisation des entités nommées. Ce choix a été motivé, entre autres, par le fait que dans les textes journalistiques, qui constituent le corpus pour notre application, la plupart des expressions qui réfèrent à des personnes font partie des syntagmes étendus (environ 50%, contre 29% pour les noms propres apparaissant seuls). D'autres auteurs utilisent cette caractéristique pour aider la catégorisation des entités nommées, par exemple déjà mentionnées (Ehrmann, Jacquet, 2006).

L'annotation des entités nommées du type Personne

Nous considérons ces expressions comme une seule entité constituée du couple (concept – entité nommée). Pour notre application, le concept correspond à une fonction politique (ou métier) occupée, dans le présent ou par le passé, par l'entité nommée. Une autre motivation pour utiliser cette stratégie est le fait qu'elle permet facilement de se restreindre à un seul phénomène, dans notre cas les personnes, et d'exclure de cette manière un certain nombre d'ambiguïtés.

Pour produire les annotations, nous avons utilisé NooJ (Silberztein, 2002), qui est un environnement de développement linguistique basé sur les transducteurs à états finis. Il se charge de toutes les étapes de prétraitement nécessaires à un texte brut : la tokenisation, la segmentation et l'application des dictionnaires. NooJ offre également une interface graphique facilitant la création des patrons de reconnaissance sous la forme de graphes à plusieurs niveaux, permettant ainsi de créer des grammaires de reconnaissance. Le principal atout de cette application, exploité dans notre cas, est sa capacité à produire des informations différentes en fonction des parcours dans les graphes de reconnaissance. Cette propriété nous a permis d'utiliser certaines informations syntaxiques (comme les relations en sein d'un groupe nominal), sans avoir a introduire d'analyse syntaxique dans la chaîne de traitement.

Nous allons présenter maintenant les ressources linguistiques mises en œuvre dans la phase d'annotation. Elles sont de deux types : les dictionnaires et les graphes de reconnaissance. Nous avons fait le choix d'utiliser les dictionnaires généraux du français. Par contre les graphes de reconnaissance ont été créés spécifiquement pour les besoins de l'application et en s'appuyant en grande partie sur les traits sémantiques présents dans ces dictionnaires. Nous utilisons entre autres :

- Le dictionnaire DELA distribué par Laboratoire d'Automatique Documentaire et Linguistique (LADL), qui comporte des traits sémantiques comme : +Hum, +HumColl, +Profession pour décrire les lexies susceptibles de désigner les entités du type humain.

- Le dictionnaire Prolex-PaysCapitales diffusé par le portail Tln du Laboratoire d'Informatique de l'Université François-Rabelais de Tours, qui contient les 191 pays indépendants (avec les capitales, les gentilés et les adjectifs toponymiques). Il utilise, entre autres, les traits sémantiques suivants : +Toponyme, +Pays, +Hum.

- En plus des noms propres inclus dans le dictionnaire de LADL, nous disposons également d'une liste de 259 formes (noms des personnes, lieux géographiques, organisations) adaptée à notre corpus, extraites manuellement.

Le format de ces dictionnaires étant adapté pour Unitex nous avons été obligé d'effectuer quelques adaptations. Nous avons également enrichi les listes des lexies portant le trait sémantique +Hum, pour garantir à nôtres grammaires une plus grande couverture. Nous avons créée deux grammaires de reconnaissance indépendantes : une pour la première étape de traitement, constituée de 9 graphes à plusieurs niveaux (Figure 1), et la deuxième qui permet de créer les chaînes de coréférence (cf. chapitre 3.3). Les annotations produites par ces grammaires ont la forme suivante :

1. *le Premier ministre Alassane Dramane Ouattara,* <personne titre = "Premier ministre"nom = "Alassane Dramane Ouattara" genre = "masculin">

2. *le chef de l'Etat nigérian, Olusegun Obasanjo qui est également le président en exercice de l'Union africaine,* <personne titre = "chef de l'Etat" nationalite = "nigérian" nom = "Olusegun Obasanjo" titre = "président en exercice" org = "union africaine" genre = "masculin">

3. *Tiémoko Sanogo, électricien à Bouaké,* <personne nom = "Tiémoko Sanogo" titre = "électricien" geo = "Bouaké" genre = "masculin">

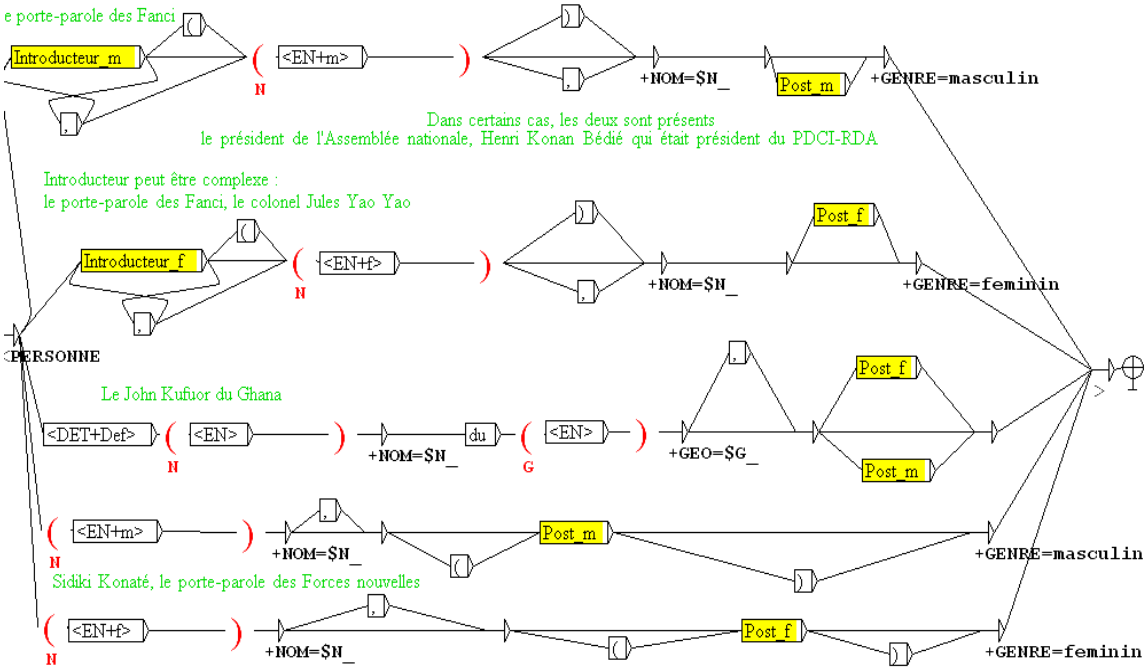

Figure 1 : Graphe principal de la grammaire

3.2 Création d'une base des connaissances

Un appel système vers NooJ qui applique les ressources décrites précédemment produit comme résultat une concordance avec toutes les expressions reconnues ainsi que tous les

```
– <personne nom="alassane dramane ouattara" nationalite="ivoirien">
    <fonction titre="premier ministre" desc="ancien" nationalite="ivoirien"/>
    <fonction titre="premier ministre" desc="ancien"/>
    <fonction titre="premier ministre"/>
    <fonction titre="chef" org="rassemblement des républicains" desc="ancien"/>
    <fonction titre="opposant" desc="principal"/>
    <fonction titre="chef" respde="opposition politique"/>
    <fonction titre="président" org="rassemblement des républicains"/>
  </personne>
```

Figure 2 : La représentation d'informations dans la base de connaissances

chemins possibles dans les graphes pour chacune de ces expressions. Nous analysons ces résultats afin de produire un fichier XML, qui correspond à notre base de connaissances. Les caractéristiques de l'information produite par ces grammaires, comme le fait que les informations à propos d'une personne sont éparses, souvent redondantes et parfois

incomplètes, nous obligent de l'analyser en deux passes. La première étape consiste à transformer les annotations produites sous la forme d'un fichier XML structuré et à regrouper toutes les annotations à propos d'une personne au sein d'un même élément XML, ensuite dans la seconde étape nous supprimons les annotations redondantes.

Le résultat de ces transformations est un fichier structuré qui présente toutes les informations à propos d'une personne sous la forme d'une liste d'éléments "fonction", comme dans notre exemple à propos de Henri Konan Bédié (Figure 2).

3.3 Traitement de la référence

Pour la troisième et la dernière étape de traitement un nouveau jeu de graphes a été développé (une grammaire constituée de 12 graphes). Cette fois-ci nous sommes à la recherche : des noms propres seuls et des syntagmes nominaux non accompagnés par un nom propre, en plus des expressions annotées par la première grammaire. Les graphes sont prévus pour ne pas annoter des cas où le nom propre fait partie d'une entité nommée d'un autre type que Personne, comme dans l'exemple : *(...) des agents doubles infiltrés pour prêter main forte <groupe> aux partisans du sergent Ibrahim Coulibaly </groupe>*. Les annotations produites par cette grammaire constituent une liste d'expressions candidates pour créer une chaîne de coréférences, et elles ont la même forme que celles des exemples 1. et 2. La résolution est basée sur les fonctions des personnes, comme dans (Figure 2), et repose ainsi en partie sur le phénomène de la reprise lexicale qui correspond à des titres comme : *général*, *colonel*, etc.

L'algorithme se déroule de la façon suivante. La première expression référentielle d'un texte ouvre une nouvelle chaîne de coréférences. Les étapes de recherche commencent par la vérification du nom propre. Si l'annotation n'en contient pas nous récupérons à partir de notre base des connaissances une liste de personnes dont au moins une des fonctions possède un attribut titre identique à celui de l'annotation. Dans le but de réduire la liste de candidats la recherche se poursuit parmi d'autres attributs de la même fonction, en priorité *org*, *geo*, qui sont considérés comme fiables, contrairement à par exemple *desc* qui contiennent des adjectifs qualificatifs. Dans le cas où nous ne disposons pas d'autres informations que le titre, nous exigeons que la personne soit déjà présente dans une chaîne de coréférences existante, dans le cas contraire nous abandonnons le processus de la résolution.

4 Évaluation des résultats

Deux modules de notre application nécessitent une évaluation spécifique et indépendante. Nous voulons souligner le fait que les résultats de la résolution de référence sont directement dépendants du contenu de la base de données et par conséquent de l'efficacité du processus d'annotation dans la première étape du traitement. Pour cela nous avons procédé à un relevé manuel des expressions de ce type dans le corpus d'acquisition (pour vérifier l'exactitude de la résolution de la référence nous avons procédé de la même manière).

Commençons par donner quelques informations à propos de ce corpus. Il est constitué de 46 articles de presse et dépêches journalistiques de tailles différentes, en provenance des sites Web (les grammaires décrites dans le chapitre précédentes ont été crées à partir de 5 premiers articles). Les articles sont d'origines diverses et une partie provient de sites étrangers francophones. L'élément unificateur de ce corpus est sa thématique, la crise en Côte-d'Ivoire, d'où la présence des mêmes personnes et organisations tout au long des articles.

4.1 Évaluation de la couverture des grammaires

Sur un total de 505 annotations produites par les grammaires, nous obtenons le bruit de 16 résultats, contre le silence pour 15 expressions, soit un rappel et précision d'environ 0,95. Nous pouvons distinguer 4 types d'erreurs dans l'annotation :

- **L'omission pure et simple.** Par exemple dans : *"La Côte d'Ivoire n'est pas un quartier de Paris" avait lancé à la télévision ivoirienne* ***le "général patriote" Charles Blé Goudé,*** *au soir du 6 novembre,(…)* le syntagme nominal n'est pas annoté et cette erreur est provoquée par la présence des guillemets que nous n'avons pas prévu de traiter à cet endroit.

- **Le silence partiel.** Par exemple dans : *Dimanche soir, <personne titre = "président" respde = "parti" >* ***le président du parti*** *</personne> du présidentiel, <personne desc = "ancien" titre = "Premier ministre" nom = "Pascal Affi N'Guessan">* ***l'ancien Premier ministre Pascal Affi N'Guessan*** *</personne> demandait en effet (…).* Nous voyons ici que le syntagme a été découpé par deux annotations distinctes. Une partie de l'annotation a été effectuée de façon correcte, nous permettant d'extraire des informations pour la base de connaissances, par contre l'apposition a été traitée comme nécessitant le processus de résolution. Un autre cas de figure sont les silences provoqués par le fait que les expressions du type localisation n'ont pas été suffisamment traitées. Par exemple : *C'est ce qui explique la présence <personne titre = "Sergent" nom = "Chérif Ousmane" >****du Sergent Chérif Ousmane*** *</personne> depuis lundi aux côtés <personne titre = "Caporal" nom = "Fofana Losseny" titre = "Commandant" respde = "zone">* ***du Caporal Fofana Losseny, Commandant de la zone*** *</personne> Ouest.*

- **Le dépassement de bornes droites ou gauches du syntagme**. Nous avons trouvé cette erreur une fois dans : *Les deux hommes ont discuté pendant une heure à la résidence <personne titre = "second" respde = "présence du ministre des Affaires étrangères" nom = "Mamadou Bamba">* ***du second en présence du ministre des Affaires étrangères, Mamadou Bamba*** *</personne>*. Cette annotation a été provoquée par la présence du trait sémantique +Hum attribué à la lexie second dans le dictionnaire utilisé.

- **L'annotation non désirée**. Par exemple dans : *Mais ces bombardements visent visiblement à fragiliser <personne desc = "dispositif" titre = "militaire" org= "Forces nouvelles">* ***le dispositif militaire des FN****</personne>*. Le syntagme nominal a été annoté à tort pour les mêmes raisons que celui cité précédemment, la lexie *dispositif* étant décrite dans le dictionnaire avec un trait +Hum, couplé avec la présence d'article défini singulier dans le contexte.

4.2 Résultats de la résolution de la référence

Dans notre corpus 87 expressions ont été annotées comme ayant besoin de la résolution de la référence. Comme pour la couverture des grammaires, nous pouvons observer ici plusieurs cas de figure :

- La résolution de la référence a eu lieu et elle est correcte (35 cas, soit 40%). Par exemple l'annotation : *Mardi soir, <personne titre = "ambassadeur" geo = "France"*

org = "ONU"> ***l'ambassadeur de France aux Nations unies*** *</personne> indiquait que (..)* a été correctement interprétée comme Jean-Marc de la Sablière.

- La résolution a eu lieu, mais il y a une erreur (5 cas, soit 6%). Par exemple dans la proposition: *La trajectoire politico-militaire <personne desc = "défunt" titre ="président">* ***du défunt président****</personne> Guéï a sans doute compté.* L'expression du défunt président a été à tort interprétée comme Félix Houphouët Boigny. Nous pouvons facilement deviner que cette erreur a été provoquée par l'annotation incomplète du syntagme. Les autres cas de résolution erronée sont dus, soit au fait d'informations manquantes dans la base de connaissances, soit à l'insuffisance de l'algorithme lui-même.

- La résolution est abandonnée, parce que le système ne dispose pas d'un candidat reconnu comme suffisamment adéquat (44[1] cas, soit 50%) pour la résolution. Pour ne citer que quelques exemples, les expressions comme *le civil* ne sont pas traitées, faute de ressources lexicales suffisantes. Un autre cas de figure est l'abandon de résolution dans les situations pour lesquelles aucune stratégie n'a pas été prévue : *<personne titre = "président" geo = "Afrique du Sud">* ***Le président d'Afrique du Sud*** *</personne> a reçu mandat (...).* Cette expression aurait dû être résolue correctement, parce que dans notre base de connaissances nous possédons l'information que *Thabo Mbéki est le président sud- africain*. Par contre aucun lien sémantique n'a été implémenté pour représenter la relation entre Afrique du Sud et l'adjectif toponymique lui correspondant.

5 Conclusions

Nous avons présenté ici notre démarche pour l'analyse et l'annotation des entités nommés de catégorie Personne. L'objectif principal de cette étude était la mise en œuvre d'une stratégie d'annotation et de la résolution partielle de la coréférence des descriptions définies. Nous avons essayé d'utiliser une approche qui permet d'obtenir les résultats fiables avec le minimum des ressources. De ce point de vue, nous pouvons considérer que l'objectif a été atteint, le taux d'erreur de 6% étant satisfaisant. L'étape prochaine consistera a enrichir progressivement notre module en ressources (comme par exemple permettre de faire le lien entre les gentilés et les noms propres des lieux géographiques) sans perdre la précision du processus. Notre démarche est actuellement étendue sur la modélisation des groupes des personnes, le traitement des antécédents disjoints et des expressions pronominales.

Références

EHRMANN M., JACQUET G. (2006). Vers une double annotation des Entités Nommées. *Traitement Automatique du Langage* 47:3, 63 – 88.

FERRET O., GRAU B., HURAULT-PLANTET M., ILLOUZ G., JACQUEMIN C. (2000). QALC – the Question-Answering system of LIMSI-CNRS. Actes de *The Ninth Text REtrieval Conference (TREC 9)*. 235 – 244.

[1] Dans un certain nombre des cas (3 réponses) nous n'étions pas en mesure de vérifier l'exactitude de la réponse donnée par la résolution.

GAIZAUSKAS R., GREENWOOD M., HARKEMA H., HEPPLE M., SAGGION H., SANKA A. (2005). The University of Sheffield's TREC 2005 Q&A Experiments. Actes de *The Fourteenth Text REtrieval Conference (TREC 2005)*.

GARDENT C., MANUELIAN H. (2005). Création d'un corpus annoté pour le traitement des descriptions définies. *Traitement Automatique du Langage* 46:1, 115 – 139.

HEARST M. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th International Conference on Computational Linguistics*.

KAZAMA J., TORISAWA K. (2007). ExploitingWikipedia as External Knowledge for Named Entity Recognition. In *Proceedings of the 2007 EMNLP-CoNLL.*

MAGNINI B., PIANTA E., POPESCU O., SPERANZA M. (2006). Ontology Population from Textual Mentions: Task Definition and Benchmark. *Proceedings of the OLP2 workshop on Ontology Population and Learning, Sidney, Australia*, *Joint with ACL/Coling 2006*.

MAYNARD D., TABLAN V., URSU C., CUNNINGHAM H., WILKS Y. (2001). Named Entity Recognition from Diverse Text Types. Actes de *Recent Advances in Natural Language Processing 2001 Conference, Tzigov Chark, Bulgaria*. 257 – 274.

MEYER J., DALE R. (2002). Mining a corpus to support associative anaphora resolution. In *Proceedings of the Fourth International Conference on Discourse Anaphora and Anaphor Resolution*.

MUC – 6 (1995). *Proceedings Sixth Message Understanding Conference*.

MUC – 7 (1998). *Proceedings Seventh Message Understanding Conference*.

POESIO M., VIEIRA R., TEUFEL S. (1997). Resolving bridging references in unrestricted text. *In Proceedings of the ACL Workshop on Operational Factors in Robust Anaphora*.

POIBEAU T. (2003). *Extraction automatique d'information*. Paris : Hermès - Lavoisier.

POIBEAU T. (2005). Sur le statut référentiel des entités nommées. Actes de la *Conférence Traitement Automatique des Langues Naturelles (TALN 2005)*. 173 – 182.

SAGGION H., FUNK A., MAYNARD D., BONTCHEVA K. (2007). Ontology-based Information Extraction for Business Intelligence. Actes de *The 6th International Semantic Web Conference and 2nd Asian Semantic Web Conference (ISWC/ASWC2007), Busan, South Korea*. 837 – 850.

SEKINE S., NOBATA C. (2004). Definition, dictionaries and tagger for Extended Named Entity Hierarchy. Actes de *The 4th International Conference on Language Resources and Evaluation (LREC 2004)*, Lisbonne. 1977 – 1980.

SILBERZTEIN M. (2002 - 2008). NooJ Manual. http://www.nooj4nlp.net/.