Création d'un corpus annoté pour le traitement des descriptions définies

Claire Gardent* — **Hélène Manuélian****

* *CNRS*
Laboratoire Lorrain de recherche en informatique et ses applications
Campus Scientifique BP 239
54506 Vandoeuvre-lès-Nancy Cedex
Claire.Gardent@loria.fr

** *Laboratoire MétaDIF*
Université de Cergy Pontoise
33 boulevard du port
95 000 Cergy Pontoise
helene.manuelian@lsh.u-cergy.fr

RÉSUMÉ. *L'évaluation et l'entraînement d'algorithmes pour le traitement automatique d'anaphores nécessitent le développement de corpus annotés. Dans cet article, nous proposons une méthodologie pour l'annotation des descriptions définies qui permet d'annoter environ 5 000 descriptions définies de façon consistante et utile pour les modules de résolution. Nous présentons ensuite les résultats de cette annotation et discutons de leurs implications pour la résolution automatique.*

ABSTRACT. *The training and the evaluation of anaphora resolvers require the development of annotated corpora. In this paper, we propose a methodology for annotating definite descriptions that supports the consistent annotation of a large corpus (roughly 5 000 definite descriptions). We then present the results of the annotation work and discuss the implications for the automatic resolution of definite descriptions.*

MOTS-CLÉS : *Descriptions définies,annotation de corpus,résolution automatique d'anaphores*

KEYWORDS: *Definite descriptions, corpus annotation, anaphora resolution*

1. Introduction

L'évaluation et l'entraînement d'algorithmes pour le traitement automatique d'expressions anaphoriques passent par le développement de corpus annotés (cf. Fig 1). Pour l'anglais, on recense les corpus MUC-6 et MUC-7, le corpus issu du PennTreebank annoté par Poesio et Vieira et plus récemment, le corpus GNOME.

Les corpus MUC-6 et MUC-7 (Message Understanding Conference) sont annotés en vue du traitement de la coréférence entre noms, groupes nominaux et pronoms (personnels et démonstratifs). Ces corpus ont permis d'élaborer un schéma d'annotation pour la coréférence (Chinchor *et al.*, 1997) puis d'entraîner et d'évaluer les systèmes de résolution des chaînes de coréférence. Sur des ensembles de 30 articles de presse, les meilleurs systèmes de résolution des coréférences ont ainsi acquis une mesure F de 62,3%.

Le corpus de (Poesio *et al.*, 1998) est un corpus fait d'articles de presse, dans lequel sont annotées environ 1 400 descriptions définies. Le taux inter-annotateurs étant très bas, le corpus n'est pas considéré comme fiable et n'est donc pas diffusé. Son annotation a néanmoins permis d'une part, de mieux quantifier les différentes catégories de descriptions définies et d'autre part, d'identifier les difficultés soulevées par l'annotation et la résolution des descriptions définies.

Enfin, le corpus GNOME (Poesio, 2004a) contient environ 18 000 groupes nominaux (pronoms, démonstratifs, possessifs et définis) annotés selon 14 dimensions syntaxiques et sémantiques. Ce corpus a permis d'établir un manuel d'annotation détaillé pour ces différents phénomènes (Poesio, 2004b) ainsi que d'entraîner et évaluer des résolveurs d'anaphores (Poesio, 2003, Poesio *et al.*, 2004a, Poesio *et al.*, 2004c) et des générateurs (Cheng *et al.*, 2001, Cheng, 2001, Karamanis, 2003).

Pour le français, le corpus ARCADE (Tutin *et al.*, 2000), diffusé par l'ELRA, contient un million de mots, et les expressions anaphoriques annotées sont : les pronoms personnels, possessifs, démonstratifs, indéfinis, et mentionnels, les adverbes anaphoriques, et les ellipses nominales. Tous les types de relations anaphoriques y sont représentés mais les descriptions définies ne sont pas annotées.

D'autres corpus existent en accès libre dans la base Freebank [1] avec en particulier, le corpus OZKAN de Nadine Ozkan et Jean Caelen où les expressions nominales (indéfinis, définis, pronoms, démonstratifs) de 33 dialogues (11 500 mots) ont été annotées mais où seules les propriétés anaphoriques des expressions référentielles en "autre" sont prises en compte.

Par ailleurs, divers articles mentionnent des corpus annotés mais non diffusés tels (Salmon-Alt *et al.*, 2002) dont le corpus contient environ 500 descriptions définies et 300 descriptions démonstratives en portugais et en français ; et (Tutin *et al.*, 2004) pour les pronoms personnels et les déterminants possessifs avec environ 80 000 mots

1. http ://freebank.loria.fr/corpus.php

Corpus	Phénomène annoté	Taille	Langue
MUC-7	Noms, GN, Pronoms Coréférence	info indisponible	Anglais
GNOME	GN Coref. et Assoc.	18 000 SN dont 554 DD	Anglais
Poesio et Vieira 98	Desc. définies Toutes relations	1 400 DD	Anglais
ARCADE	Pronoms, adverbes Ellipses nominales 5 relns anaphoriques dont Coref. et Assoc.	1 million de mots	Français
OZKAN	GN et Pronoms Anaphores en “autre”	1 344 GN et Pronoms	Français
Salmon-Alt 2004	D. Définies Anaph.Infidèles	9 000 mots 741 DD	Français
Vieira et Salmon-Alt 2002	D. Déf. et Dém Toutes relations	500 DD 300 Dem	Français Portugais
ARCADE multilingue	Pronoms pers. et dem. Coréférence	80000 mots 800 pronoms	Français Anglais

Figure 1. *Corpus annotés au niveau anaphorique*

et 800 expressions anaphoriques. La figure (1) résume l’ensemble des corpus annotés au niveau anaphorique.

Dans cet article, nous nous intéressons à l’annotation de corpus en vue de l’interprétation automatique des descriptions définies, c’est-à-dire des expressions de la forme *le/la/les N*. En effet, comme le montre le récapitulatif donné par la Figure 1, il n’existe pour le français aucun corpus annoté qui soit de taille suffisante et de qualité suffisament fiable pour entrainer ou évaluer les résolveurs de descriptions définies.

Pour l’anglais, (Poesio *et al.,* 1998) présente les résultats de plusieurs tâches d’annotation visant à classifier les descriptions définies. Les conclusions tirées à partir des résultats obtenus sont largement négatives : pour les classifications un peu fines, le taux d’accord entre annotateurs est relativement bas et seule une classification binaire semble possible. Poesio et Viera en concluent qu’il est difficile voire impossible d’annoter les descriptions définies d’un corpus de façon suffisament fiable pour que ce corpus puisse servir à l’évaluation et à l’entraînement de modules d’interprétation des expressions référentielles. (Salmon-Alt *et al.,* 2002) décrit une étude similaire pour les descriptions définies et démonstratives du français et du portugais et aboutit également

à un taux d'accord entre annotateurs très bas (un kappa de 0.52 pour les descriptions définies du français[2]).

Nous reprenons ici le travail amorcé par Poesio et Viera et décrivons l'annotation d'un corpus contenant près de 4 900 descriptions définies. Nous commençons (section 2) par définir une méthodologie pour l'annotation des descriptions définies qui permet d'annoter un gros corpus de descriptions définies de façon consistante et utile pour les modules de résolution. Nous présentons ensuite (section 3) les résultats de cette annotation et en discutons les implications pour les résolveurs automatiques.

2. L'annotation des descriptions définies : méthodologie, schéma et outils

2.1. *Méthodologie*

(Poesio *et al.*, 1998, Salmon-Alt *et al.*, 2002) montrent qu'il est difficile d'annoter les descriptions définies de façon cohérente : typiquement, une même description définie sera catégorisée différemment par les annotateurs soit parce qu'ils sont en désaccord sur la catégorie à attribuer, soit parce que la catégorisation est de fait ambigüe (deux catégories sont possibles).

Pour pallier ces difficultés, nous proposons de modifier la méthodologie d'annotation adoptée par (Poesio *et al.*, 1998) de la façon suivante.

Premièrement, le travail de Poesio et Viera était fait entièrement manuellement ce qui rendait difficile la multiplication des expérimentations et partant, la mise au point d'un schéma par itération. Nous utilisons en revanche les outils de prétraitement et d'annotation qui sont actuellement disponibles (cf. section 2.3) . De cette façon, la tâche d'annotation est facilitée et la fiabilité indirectement accrue.

Deuxièmement, Poesio et Vieira avaient pour objectif de définir un schéma d'annotation utilisable par des non experts si bien que l'annotation étaient faite par des étudiants non linguistes. L'annotation présentée dans cet article a été en revanche faite par des linguistes confirmés de la même façon que les annotateurs de la Penn Treebank étaient constitués d'une groupe de linguistes experts dont les discussions et décisions contribuèrent à l'établissement du schéma d'annotation finalement adopté (Marcus *et al.*, 1993).

Troisièmement, Poesio et Vieira utilisent deux schémas d'annotation fixés une fois pour toute. Nous utilisons en revanche un schéma affiné par plusieurs itérations : à partir d'un premier schéma, une phase d'annotation est entreprise au terme de laquelle les désaccords entre annotateurs sont discutés, résolus et les modifications nécessaires intégrées au schéma d'annotation. Ainsi le but n'est pas tant d'avoir un taux d'accord

2. Le coeffi cient kappa permet de mesurer l'accord entre annotateurs lorsque la tâche consiste à attribuer à des items une classe parmi un ensemble non ordonné de classes (Carletta, 1996). Un Kappa supérieur à 0.8 indique un schéma de classifi cation fi able et un Kappa entre 0.68 et 0.8, un schéma moyennement fi able.

inter-annotateurs élevé que d'aboutir à un consensus sur les annotations effectuées. Nous rejoignons en ceci la stratégie adoptée actuellement dans divers travaux d'annotation tels que le FrameNet allemand (Erk *et al.,* 2003) ou le TigerTreebank (Brants, 2000), travaux où les phases dites d'adjudication et de négociation font partie intégrante du processus d'annotation et visent la spécification d'annotations consensuelles plutôt que d'un taux d'accord inter-annotateurs élevé.

2.2. *Schéma d'annotation*

Comme l'ont clairement montré (Christophersen, 1939, Hawkins, 1978), les usages faits des descriptions définies sont multiples[3] . Ainsi une description définie peut coréférer avec une entité déja introduite (1a, anaphore *coréférentielle*). Elle peut avoir une fonctionnement autonome (1b, description *autonome*) ou encore, elle peut être interprétée en association avec un élément discursif antécédent (1c, anaphore *associative*).

(1) a. Un homme et une femme entrent dans la pièce. *L'homme* porte un chapeau.
b. *La terre* tourne autour *du soleil*.
c. La maison de Paul est magnifique. *Les fenêtres* sont en chêne.

Pour l'annotation de textes réels, il importe de définir un schéma d'annotation, et donc des catégories d'usage, qui couvre l'ensemble des usages effectivement réalisés. Ce schéma doit en particulier rendre compte aussi bien des usages coréférentiels, que des usages associatifs et des usages autonomes. Il doit également permettre de distinguer entre descriptions définies référentielles et non référentielles, ces dernières n'étant généralement pas prises en compte par les algorithmes de traitement des descriptions définies. Enfin, il doit fournir un base d'évaluation et d'entrainement précise pour les outils de TAL (résolveur ou générateurs de descriptions définies).

C'est dans cette optique que nous proposons un schéma basé sur une classification fine qui contraste en particulier, avec les classifications très générales, parfois binaires, des corpus annotés existants tels que (Fraurud, 1990) ; et que nous couplons ce schéma

3. Il est possible toutefois d'avoir une analyse uniforme de cette multiplicité. Ainsi par exemple, (Cooper, 1979, Kadmon, 1990, Groenendijk *et al.,* 1995, Westerstahl, 1991) proposent une analyse uniforme où l'unicité d'une description défi nie résulte de l'effet combiné de la description et du contexte. Dans (Cooper, 1979), l'interprétation d'une description défi nie fait ainsi intervenir une variable libre dont la valeur est déterminée par le contexte d'énonciation ; dans (Kadmon, 1990) la propriété identifi ante peut être présupposée, impliquée ou donnée par le contexte et dans (Westerstahl, 1991, Groenendijk *et al.,* 1995), le défi ni est similaire aux quantifi eurs en ce que sa dénotation fait intervenir une restriction contextuelle. Néanmoins pour un résolveur d'anaphores, l'essentiel est de déterminer d'où proviennent les restrictions contribuant à déterminer l'unicité. C'est dans cette optique que la classifi cation proposée est discriminante plutôt qu'unifi ante. Notons en outre que pour rendre l'annotation faisable, il n'est pas tenu compte dans cette classifi cation des phénomènes de portée qui peuvent intervenir. En d'autres termes, les descriptions défi nies sont traitées comme référentielles plutôt que quantifi antes.

avec une stratégie de gestion des conflits de catégorisation d'une part (une même description définie peut en effet être catégorisée de plusieurs facons, toutes correctes) et une annotation des liens coréférentiels impliquant les descriptions définies d'autre part. Dans ce qui suit, nous commencons par definir les catégories du schéma d'annotation utilisé. Nous présentons ensuite la stratégie de gestion des conflits et montrons en quoi cette stratégie, couplée avec l'annotation des liens coréférentiels permet une annotation adaptée aux besoins de l'évaluation des résolveurs.

Les catégories de base du schéma sont les suivantes :

Description autonome : après résolution des anaphores ou ellipses éventuellement présentes dans les modifieurs du nom tête, le référent de la description définie est identifiable indépendamment du contexte linguistique et extra-linguistique[4](cf *La terre* dans l'exemple 1b et *La maison de Paul* dans l'exemple 1c).

Description coréférentielle : le référent de la description définie est identique à un référent introduit dans le contexte linguistique antérieur (cf *L'homme* dans l'exemple 1a).

Description contextuelle : le référent de la description définie est lié par une relation autre que l'identité à un référent introduit dans le contexte linguistique ou extra-linguistique antérieur (cf *Les fenêtres* dans l'exemple 1c).

Description non référentielle : la description définie ne décrit pas de référent de discours mais introduit une prédication ou fait partie d'une expression figée (e.g., "faire la cour")

Nous détaillons maintenant chacune de ces catégories, nous les mettons en relation avec les théories sémantiques existantes et nous formulons les critères de catégorisation utilisés pendant l'annotation.

2.2.1. *Descriptions autonomes*

Les descriptions dites "autonomes" sont des descriptions qui permettent d'identifier le référent désigné indépendamment du contexte linguistique et de la situation d'énonciation.

Cette classe regroupe les unica de Russell (2a), les descriptions contenant des noms propres ou des items (nombre, mot) fonctionnant comme des noms propres au sens où ils exhibent le comportement des désignateurs rigides de (Kripke, 1949) (2b), les descriptions décrivant des concepts abstraits (2c), les noms suivis d'une complétives (2d),

4. La notion de *référence* adoptée dans cet article est celle issue des travaux sur la sémantique discursive (cf. (Webber, 1978, Kamp, 1981)). Dans ces travaux, l'interprétation d'un discours (texte, dialogue) implique la construction d'un modèle du discours peuplé par des objets discursifs (appelé aussi référents ou entités du discours) dont les propriétés sont spécifi ées par un ensemble de conditions simples ou complexes. Ce sont ces objets discursifs auxquels réfèrent les SN référentiels indépendamment de l'existence, réelle ou non, de l'objet décrit.

les noms suivis d'un modifieur permettant d'identifier le référent désigné indépendamment du contexte (2e)[5].

(2) a. Le soleil, la lune, etc.
b. L'année 1984, le mot "le", la République Populaire de Chine, le président Chirac, l'article défini
c. La sécheresse, le pouvoir
d. Le fait que Marie soit partie, la question de savoir si Marie est partie, etc.
e. Les championnats du monde de cyclisme sur route, la force multinationale de sécurité à Beyrouth, le référendum par lequel les Turcs devaient se prononcer pour ou contre ...etc.

Nous incluons également dans cette classe toute une catégorie de descriptions définies qui n'est pas discutée dans l'approche de Poesio et Vieira, à savoir les descriptions où le modifieur de la tête nominale contient des anaphores ou des ellipses. Pour ces descriptions, l'identification du référent est autonome seulement après résolution des anaphores et/ou des ellipses :

(3) a. La femme qu'*il* a rencontré, la même chose que *lui*
b. Kodak/l'ensemble de *sa* gamme nouvelle
c. les problèmes juridiques que *cela* va poser

2.2.2. *Descriptions coréférentielles*

Une description coréférentielle spécifie un référent déjà introduit dans le contexte linguistique par un groupe nominal antécédent. (Fraurud, 1990, Hawkins, 1978, Prince, 1981) parlent respectivement d'emploi en mention subséquente, de co-spécification, de coréférence ou de référents *évoqués textuellement (textually evoked)*.

La relation entre la description fournie par l'antécédent et celle fournie par la description définie coréférentielle varie. (Clark, 1977, Strand, 1997) différencient les reprises directes (ou fidèles) des autres. Dans une reprise directe, la tête nominale de la DD est la même que celle de l'antécédent (4a). Les reprises indirectes incluent les reprises via une relation lexicale (de synonymie (4b), d'hyperonymie (4c) ou d'hyponymie en (4d)) et les redescriptions (Fraurud, 1990) (ou épithètes, (Strand, 1997)). Dans ce cas (4e), la description définie n'entretient aucune relation formelle avec celle de son antécédent et la détection de la coréférence résulte d'un processus d'inférence permettant de déterminer la compatibilité des deux descriptions.

(4) a. un homme/l'homme
b. un putsch/le coup d'Etat

5. Cette dernière catégorie correspond aux *containing inferrables* de (Prince, 1981) et est incluse (avec les noms suivis d'une complétive) dans les premières mentions de (Corblin, 1987).

c. deux malfaiteurs/les hommes

d. six hommes/les truands

e. UCAR .. Duracell/les deux concurrents

2.2.3. *Descriptions contextuelles*

Nous regroupons sous le terme "descriptions définies contextuelles", les descriptions associatives et situationnelles, c'est-à-dire les descriptions dont l'interprétation est déterminée par une relation de non identité avec une entité accessible dans le contexte d'énonciation.

Les *descriptions associatives* s'interprètent en relation avec un entité explicitement introduite dans le contexte linguistique[6]. Ainsi dans (5), la description définie *les deuxième et troisième places* est interprétée comme signifiant *les deuxième et troisième places aux championnats du monde de cyclisme sur route*.

(5) *les championnats du monde de cyclisme sur route*. Les Néerlandaises Heleen Hage et Connie Meijer qui occupent *les deuxième et troisième places* (...)

Les études existantes sur l'annotation des descriptions définies soulignent la difficulté d'annoter les descriptions associatives de façon consistante (Poesio *et al.*, 1998). Afin de limiter ces difficultés, nous considérons comme descriptions associatives uniquement les descriptions ayant un antécédent nominal clairement identifié et du bon type sémantique. Ainsi, *le gouvernement* sera annoté comme associatif en (6a) mais non en (6b).

(6) a. Italie : Le gouvernement a décidé...

b. le gouvernement *italien* a decidé, vendredi, d'envoyer une flottille de dragueurs de mines dans la région. Cette décision, qui devra être enterinée lundi et mardi par le *Parlement*, ...

Pour rendre compte de cas tels que (6b) ainsi que de ceux où la description ne permet d'identifier le référent désigné qu'en relation avec une entité accessible dans le contexte d'énonciation, nous introduisons une nouvelle catégorie, la catégorie des *descriptions situationnelles*. Ainsi, dans un texte portant sur l'Italie mais où l'Italie n'est pas mentionnée explicitement (ou dans l'exemple 6b ci-dessus), la description *le Parlement* sera annotée comme situationnelle pour refléter le fait que le référent désigné est *le parlement de l'Italie*.

6. Dans la littérature, les descriptions contextuelles avec antécédent ont également été dénommées anaphores associatives (Kleiber, 1997, Kleiber, 2001), *bridging anaphora* (Clark, 1977) et *inferrables* (Prince, 1981). Le premier à utiliser la notion d'association est (Guillaume, 1919). (Corblin, 1987) note que le terme est repris dans (Blanche-Benveniste *et al.*, 1966) et que (Hawkins, 1978) l'utilise de façon indépendante pour l'anglais. Le phénomène d'anaphore associative en français est étudié par (Fradin, 1984).

L'introduction de cette nouvelle catégorie permet d'une part, de faciliter l'annotation des anaphores associatives (l'antécédent doit être un GN du type sémantique attendu) et d'autre part, d'éviter la sous spécification des liens anaphoriques. En effet, dans les deux schémas d'annotation proposés par (Poesio *et al.*, 1998), les catégories utilisées sont : première mention, coréférentiel, associatif et infidèle, non référentiel. N'ayant pas d'antécédent textuel clairement identifiable, une DD qui dans notre schéma, sera classée en situationnelle, sera catégorisée comme *première mention* dans le schéma de Poesio et Vieira échouant ainsi à identifier la dépendence contextuelle de ces expressions pour leur interprétation.

2.2.4. *Descriptions non référentielles*

Certaines descriptions définies ne sont pas référentielles au sens où elles ne pointent pas sur un référent discursif.

C'est le cas en particulier, lorsque qu'une DD apparaît dans un usage attributif (7a), dans une structure prédicative (7b), une apposition (7c). Dans les corpus MUC pour le traitement de la coréférence, les SN têtes de structures prédicatives ou d'une apposition sont annotés de la même façon que les autres. Néanmoins, cette stratégie est criticable (van Deemter *et al.*, 2000, venex, 2005). Ces SN servent en effet à construire une prédication. Ils ne dénotent ni un argument ni un modifieur et leur interprétation est indépendente du contexte. Nous adoptons donc la stratégie selon laquelle, dans une structure prédicative dont la tête est une DD, soit le sujet soit l'objet sera annoté tandis que dans une structure avec apposition, seul un des SN de cette structure sera annoté.

Une DD sera également catégorisée comme non référentielle lorsqu'elle fait partie d'une expression figée (7d), d'une conjonction (7f) ou d'un quantifieur (7e) .

(7) a. Jean cherche *la meilleure méthode d'annoter les pronoms*.
b. Les Etats-Unis et le Japon continuent a être *les principaux partenaires étrangers du régime de Manille*.
c. Ce vicomte parle comme un " ketje " des Marolles, *le quartier populaire de Bruxelles*.
d. Ce témoignage (...) redonnerait *du corps* à une hypothèse.
e. *La plupart* des activités commerciales et administratives ont été interrompues.
f. Une situation qui ne pourrait que s'aggraver *du fait* des vents importants et de la sécheresse persistante.

2.2.5. *Gestion des conflits*

L'accord inter-annotateur obtenu par les expériences de Poesio et Vieira est très bas : K = 0.68 pour la première et K = 0.58 pour la seconde (où 0.68 est le seuil minimal pour pouvoir qualifier un schéma d'annotation de "moyennement fiable" et 0.8 pour pouvoir le qualifier de fiable). Il est donc essentiel de minimiser les désaccords et en particulier, ceux résultant d'une réelle ambiguité. En effet, comme le remarque

(Poesio *et al.*, 1998), certains désaccords dans l'annotation résultent d'une ambiguité de catégorisation : dans le contexte considéré, une même description définie peut être catégorisée de plusieurs façons, toute correctes. Ainsi dans (8), la description *la direction* peut être annotée soit comme CORÉFÉRENTIELLE avec le GN *les directeurs du groupe* soit comme ASSOCIATIVE avec le GN *la Lainière*.

(8) "Nous ne connaissons pas nous-mêmes les intentions de Jérome Seydoux", se défend l'un des *directeurs du groupe*. Les ouvriers de *la Lainière*, eux, font des pronostics : "Si jamais les Chargeurs rachètent Prouvost, ce sera *la direction* qui risquera d'être virée."

Dans ce cas, les deux catégorisations sont sémantiquement équivalentes au sens où l'interprétation finale du GN annoté (*la direction*) est la même. Cependant, l'annotation est différente si bien que l'évaluation d'un module de TAL donnera des résultats différents selon la décision d'annotation prise : une annotation du GN comme *coréférentielle* favorisera les systèmes détectant pour cet exemple une relation de coréférence tandis qu'une annotation *associative*, favorisera ceux qui détectent une relation du même type. Il importe donc d'adopter une stratégie de désambiguisation des conflits possibles aussi bien pour minimiser les désaccords entre annotateurs que pour permettre une évaluation non biaisée des systèmes de traitement de descriptions définies. Pour ce faire, nous adoptons la stratégie suivante :

– Nous combinons catégorisation et résolution des coréférences. Dans une première passe, les descriptions définies sont catégorisées comme autonome, coréférentielle, contextuelle ou non référentielle. Dans une deuxième passe, les descriptions définies entrant dans un lien de coréférence avec un antécédent textuel sont repérées et les coréférences annotées[7].

– En cas de conflit, les catégories non coréférentielles ont préférence puisque les chaînes de coréférence seront annotées quoi qu'il arrive, en supplément de toutes les autres relations anaphoriques.

Cette double stratégie permet d'une part, de simplifier l'annotation au sens où toutes les descriptions autonomes pourront être catégorisées comme telles sans prendre en compte les liens de coréférence qui peuvent intervenir entre ces descriptions autonomes et le contexte d'énonciation.

7. L'ordre "catégorisation avant identification des antécédents coréférentiels" est arbitraire et on aurait pu choisir l'ordre inverse. En pratique cependant, il permet d'éviter la tentation d'annoter systématiquement comme coréférentielle, une description ayant un antécédent (ce qui est important puisque comme nous l'avons mentionné plus haut, certaines descriptions sont simultanément coréférentielles et contextuelles et doivent être annotées comme telles pour permettre une évaluation impartiale des résolveurs automatiques). Notons en outre que la seconde passe n'inclut pas la première. En d'autres termes, il ne s'agit pas en deuxième passe uniquement d'identifier les antécédents des descriptions coréférentielles mais également d'identifier les cas où une description autonome est également en relation de coréférence avec un élément du cotexte.

Elle permet d'autre part, d'aplanir les différences d'annotation qui peuvent intervenir entre les catégories ASSOCIATIVE/SITUATIONNELLE d'une part, et CORÉFERENTIELLE d'autre part. En effet, dans les cas comme (8) ci-dessus où une description peut être annotée soit comme ASSOCIATIVE/SITUATIONNELLE, soit comme CORÉFERENTIELLE, la double annotation ASSOCIATIVE/SITUATIONNELLE+CORÉFERENTIELLE de fait annule toute divergence d'annotation possible. Par ailleurs, dans de nombreux cas, un entité est désignée plusieurs fois dans le même texte par une expression référentielle identique à celle utilisée en première mention, (*Le Parti Socialiste*), ou par une expression interprétable sans contexte, mais coréférent à un nom propre (*Chirac ... Le Président de la République Française*). Aussi, nous décidons d'annoter ces expressions comme autonome, puisqu'on les interprète sans référence au contexte antérieur, puis nous leur donnons un antécédent, de façon à ce que le système puisse repérer que le référent est le même.

La figure (2) résume les cas de conflit possibles et les critères de décisions adoptés[8].

1. Autonome > Coréférentielle :
La catégorie "autonome" est sélectionnée, les liens de coréférence étant de toute façon annotés lors de la deuxième passe.
Jacques Chirac/le Président de la République française
2. Associative > Coréférentielle :
La catégorie "associative" est sélectionnée , les liens de coréférence étant de toute façon annotés lors de la deuxième passe.
Jacques Chirac, la France/le Président
3. Situationnelle > Coréférentielle : La catégorie "situationnelle" est sélectionnée, les liens de coréférence étant de toute façon annotés lors de la deuxième passe.
Jacques Chirac/le Président
4. Associative > Situationnelle : La catégorie "associative" est sélectionnée car plus informative (l'antécédent est clairement identifié).

Figure 2. *Stratégie de résolution des conflits de catégorisation possibles*

2.3. *Corpus et outils utilisés*

Le corpus annoté est une sous-partie du corpus PAROLE[9] et comprend 48 360 mots annotés au niveau morphosyntaxique, suivant le schéma d'annotation Multext

8. L'interprétation de la plupart des noms communs est relative au temps et à l'espace. Lorqrsque le contexte fi xe ces deux paramètres à leur valeur par défaut à savoir, "maintenant" et "ici", nous annotons la DD comme autonome (plutôt que contextuelle). C'est pourquoi *le Président de la République française* est ici traité comme autonome.
9. Corpus fourni par l'ATILF dans le cadre du contrat de plan Etat-Région Lorrain sur l'ingénierie des langues intégrant la collaboration de l'ATILF - UMR 7118 CNRS-Nancy 2 - et du LORIA - UMR 7503, CNRS, INRIA, INPL, Nancy 1, Nancy 2.

(Lecomte, 1997, Beaumont *et al.,* 1998). Il est composé d'une série d'articles du journal *Le Monde* datant de septembre 1987 et appartenant à toutes les rubriques.

Afin de faciliter l'annotation, ce corpus a été pré-traité pour permettre la visualisation dans une interface d'annotation des descriptions définies à annoter. Les résultats de l'annotation ont également fait l'objet d'un post-traitement visant à faciliter la lecture et le triage des données obtenues. Plus spécifiquement :

– L'analyseur local *G-search* (Corley *et al.,* 2001, Corley *et al.,* n.d.) a été utilisé pour identifier les groupes nominaux définis.

– L'interface d'annotation *MMAX* (Muller *et al.,* 2001a, Muller *et al.,* 2001b) a été utilisée pour catégoriser les descriptions définies et identifier les antécédents le cas échéant.

– Des scripts perls ont été développés pour adapter les formats existants aux formats attendus par *G-search* et par *MMAX*.

– Des feuilles de style XSL ont été spécifiées pour produire à partir des fichiers XML produits par MMAX des fichiers HTML utilisables par un tableur et permettant ainsi de trier, classer et filtrer les données obtenues.

En sortie, nous avons donc d'une part un corpus annoté et d'autres part un ensemble de fichiers HTML permettant de visualiser par le biais d'un tableur l'ensemble des exemples illustrant chaque catégorie (autonome, coréférentielle, situationnelle, associative, non référentielle) et sous catégorie (e.g., fidèle, lexicale, redescription pour la catégorie coréférentielle) du schéma d'annotation (cf. (Manuélian, 2003) pour plus de détails).

3. Résultats et discussion

Nous présentons maintenant les résultats quantitatifs obtenus par l'analyse de corpus et discutons de leurs implications pour le traitement automatique des descriptions définies. Plus particulièrement, nous en identifions les implications d'une part, pour la *catégorisation* des descriptions définies en autonomes, contextuelles, coréférentielles et non référentielles et d'autre part, pour leur *résolution* (interprétation des antécédents et, dans le cas des descriptions associatives, identification de la relation implicite entre antécédent et description définie).

Pour environ 4 900 descriptions définies (DD) annotées, la distribution des différentes catégories est la suivante :

Auto	Assoc	Sit	Coréf	Non Réf	Total
2 871	436	585	567	451	4 910
58 %	9%	12%	11%	9%	100%

Figure 3. *Distribution des catégories de* DD

Les taux de descriptions autonomes et associatives sont comparables à ceux obtenus dans des travaux antérieurs ((Fraurud, 1990, Poesio *et al.,* 1998, Salmon-Alt *et al.,* 2002)). En revanche, la proportion de descriptions coréférentielles est faible avec seulement 11% contre environ 40% rapporté par (Poesio *et al.,* 1998) et 30% par (Salmon-Alt *et al.,* 2002). Cette différence résulte d'une différence dans la stratégie d'annotation : alors que (Poesio *et al.,* 1998) privilégie l'annotation des liens anaphoriques, nous dissocions ceux-ci (et en particulier, l'annotation des liens de coréférence) de la catégorisation des DD. Plus particulièrement, les descriptions autonomes qui sont également coréférentielles sont systématiquement catégorisées comme autonomes diminuant ainsi la proportion des coréférentielles.

3.1. *Les descriptions définies coréférentielles*

Comme nous l'avons remarqué dans la section 2.2.2, dans le cas d'une DD coréférentielle, la relation entre la tête nominale de la DD et celle de son antécédent peut être une relation d'identité (anaphore dite "fidèle"), une relation lexicale (synonymie, hyperonymie, etc.) ou une relation de redescription. La distribution de ces trois catégories est donnée dans le tableau 4.

Fidèles	Lexicales	Redescriptions	Total
244	82	241	567
43%	14%	43%	100%

Figure 4. *Distribution des* DD *coréférentielles*

On note une proportion importante d'anaphores infidèles (57% de l'ensemble des anaphores coréférentielles) c'est-à-dire, de cas où la tête nominale de l'anaphore diffère de celle de l'antécédent. Ce résultat confirme l'hypothèse avancée par (Corblin, 2004) selon laquelle les anaphores infidèles sont beaucoup plus fréquentes dans le discours dit « à interlocuteur générique » (la plupart du temps monologique et écrit) que dans la conversation courante, où l'interlocuteur est concret.

Les travaux visant la résolution de ce type de DD à partir de ressources lexicales (Poesio *et al.,* 1997, Vieira *et al.,* 2000, Salmon-Alt, 2004) aboutissent à des résultats largement négatifs (rappel d'environ 30%) tandis que les approches basées sur la notion de proximité sémantique semblent plus adaptées avec par exemple, un rappel de .67 et une précision .73 pour (Poesio *et al.,* 2004c)[10]

10. Cette notion de proximité sémantique est généralement obtenue par des méthodes statistiques dite de "clustering". En gros : deux unités linguistiques dénotent des entités proches si leur cooccurence est "anormalement" fréquente i.e. plus fréquente que la moyenne. Des méthodes basées sur la détection de constructions typiques d'associations lexicales peuvent également être mises en oeuvre. Par exemple, (Poesio *et al.,* 2002) utilise la construction *the* N_{ante} *of the* N_{ana} pour la détection de la relation partie/tout.

Si cette différence s'explique en partie par des éléments ou des relations manquantes dans les ressources lexicales utilisées[11], l'analyse du corpus PAROLE conforte l'idée que l'utilisation isolée des connaissances lexicales (synonymie, hyperonymie, etc.) ne peut suffire à permettre la résolution d'anaphores infidèles. En effet, dans ce corpus, la proportion de cas où une relation lexicale lie anaphore et antécédent est relativement faible (25% des DD coréférentielles infidèles). L'utilisation de ressources en sémantique lexicale (même si celles ci étaient complètes) est donc de faible utilité.

En revanche, on note une proportion importante (environ la moitié) de descriptions infidèles où l'antécédent est un nom propre. Dans ce cas, la reprise dénote la plupart du temps un type sémantique de base associé à l'entité nommée. Ainsi dans les exemples (9a-c), *Kodak* est repris successivement par les termes *firme, entreprise* et *marque*.

(9) a. Est-ce la seule raison qui a conduit *Kodak* à se lancer dans l'arène ? Il semble bien que pour *la firme de Rochester* ce lancement très coûteux ...

b. Le seul "plus" apporté par Kodak dans les piles classiques est en définitive son nom. Mais *l'entreprise* emploie les grands moyens pour réussir ce lancement.

c. *Kodak* vient de signer un accord avec Fuji pour vendre ses produits dans les circuits photo-ciné-son et *la marque française* se prépare pour la fin de septembre à lancer une diversification majeure dont le détail est tenu secret.

Pour ce type de cas, il n'y a pas de relation lexicale directe entre le type dénoté par le nom tête de l'anaphore et celui de l'antécédent. Il faut d'abord déduire le type sémantique de base de l'antécédent (e.g., *entreprise industrielle* pour *Kodak*) puis inférer une relation lexicale entre ce type et celui de l'anaphore (e.g., *entreprise, firme* et *marque*). Pour bien en traiter, il faudra donc combiner la détection de relations lexicales (données par WordNet, par un dictionnaire de synonymes ou par des méthodes statistiques) avec la reconnaissance d'entités nommées (Fourour, 2001) afin d'associer au nom propre antécédent un type sémantique reconnu par les relations lexicales.

Enfin, dans un petit nombre de cas de descriptions infidèles coréférentielles, il n'existe aucune relation lexicale, dérivée ou non, entre l'antécédent et l'anaphore si bien que l'identification de l'antécédent (et la catégorisation de la DD comme coréférentielle) implique un raisonnement par rapport au contexte d'énoncation. Ces cas, illustrés en (10), sont hors de portée des résolveurs actuels puisqu'ils exigent un traitement sémantique profond.

11. (Poesio, 2003) montre qu'aucun des liens méréologiques nécessaires au traitement des 58 anaphores infidèles du corpus GNOME n'est présent dans WordNet 1.6 tandis que (Garcia-Almanza, 2003) en trouve 16 par le biais de procédures de recherche plus sophistiquées. De même, (Salmon-Alt, 2004) montrent un rappel plafonné à 29.8% pour l'identification des relations lexicales portant sur 15 paires à partir de 3 ressources combinées à savoir, des listes de similarités obtenues par des méthodes statistiques, WordNet et le Petit Robert.

ANAPH	CIRC	GÉN	UNIQUES	IDENTIF	TOTAL
423	81	575	703	935	2 717
15.37 %	2.94%	20.89%	25.54%	33.96%	100%

Figure 5. *Distribution des* DD *autonomes*

(10) a. *La Lainière va peut-être supprimer des cars de ramassage !* Pour ces ouvrières du bassin houillier dont quelques-unes ont déjà trois heures de transport par jour, *la nouvelle* a rélégué au second plan les manoeuvres boursières dont leur entreprise fait l'objet.

b. *UCAR* sert le marché européen à partir de deux usines – en Grèce et en Suisse – sur un parc total de trente unités à travers le monde. Même schéma chez *Duracell* (filiale elle aussi d'un groupe alimentaire américain, Kraft), qui compte huit usines au total, dont trois en Europe : Belgique, Italie et Grande-Bretagne. Forts de ces bastions, *les deux concurrents* observent avec un grand calme l'arrivée de Kodak.

3.2. *Les descriptions définies autonomes*

Les descriptions autonomes représentent 56% des descriptions annotées confirmant ainsi les études de (Fraurud, 1990, Poesio *et al.,* 1998, Salmon-Alt *et al.,* 2002) où la proportion de première mention était très elevée (60.9% pour (Fraurud, 1990), 48.37% pour (Poesio *et al.,* 1998) et 49.6% pour (Salmon-Alt *et al.,* 2002)).

Pour le succès d'un résolveur automatique, il est donc essentiel de bien classifier ces DD et plusieurs travaux récents visent à développer des algorithmes capables d'identifier les DD autonomes. Certains sont basés sur des méthodes symboliques (cf. (Vieira *et al.,* 2000)), d'autres sur des méthodes statistiques (Poesio *et al.,* 2004c, Ng *et al.,* 2002, Uryupina, 2003)), mais tous se basent sur des critères syntaxiques : présence de modifieurs, de noms propres, etc. Une analyse plus détaillée des résultats de l'annotation (cf. tableau 5) montre cependant qu'environ 46% des descriptions autonomes sont des descriptions brèves, ne contenant ni modifieur, ni nom propre. Il s'agit des "uniques"[12] (25.54% des autonomes) et des "termes généraux" (20.89%) qui dénotent pour l'essentiel des concepts abstraits (11a), des noms d'espèces (11b) ou des génériques (11c)[13].

12. Le terme d'unique en association avec les descriptions défi nies remonte à (Russell, 1905) qui décrivait ainsi les descriptions référant à des objets uniques par défi nition tel par exemple, le soleil ou le pape.

13. L'usage générique d'une DD ne réfère pas à proprement parler. Dans "le lion a une crinière", la DD "le lion" ne réfère ni à un individu spécifi que, ni à une espèce (puisque seuls les lions mâles ont une crinière). Nous traitons néanmoins les génériques dans notre annotation en partant de l'hypothèse qu'il s'agit non de déterminer si une DD réfère ou non, mais plutôt d'identifi er les référents du discours qui peuvent entrer dans une chaîne de référence. Cette pra-

(11) a. En tout, vingt et un Etats de l'Union sur vingt-cinq sont victimes en tout ou partie de *la sécheresse*.

b. Plus proche de nous, Jean-Noël Escudier parle également de la "blanchaille de poissons", y compris sardines et anchois, et cite *le melet* et son cousin *le pissalat*, frai de poissons appelé la poutina.

c. Mais, si le nom de pissalat, qui amuse notre lecteur, est bien peu connu des Français du moins est-il un nom que *le touriste* connaît (ou apprend) sur la Côte : "pissaladière" !

Afin de repérer ces cas de descriptions autonomes, il importe donc qu'un résolveur puisse les différencier des cas de descriptions coréférentielles, dont les réalisations sont généralement également brèves. La discrimination pourra se faire sur la base de la présence ou non d'un antécédent ainsi que sur le typage sémantique du nom tête. Comme nous venons de le remarquer, les DD autonomes courtes dénotent des objets relativement abstraits ce qui est moins systématiquement le cas pour les descriptions coréférentielles.

Concernant l'interprétation (par opposition à la catégorisation) des DD autonomes, les résultats de l'annotation soulèvent plusieurs remarques.

Premièrement, une bonne proportion des DD dites "autonomes" sont en fait indirectement dépendantes du contexte pour leur interprétation puisque dans 15.37% des cas, ces descriptions contiennent des anaphores ou des ellipses. Catégoriser une DD comme "autonome" ne dispense donc pas d'un travail d'interprétation qui peut porter sur d'autres DD bien sûr (et tel est fréquemment le cas) mais également sur d'autres types d'anaphores (pronom, démonstratifs, possessifs, etc.) ou d'ellipse. En d'autres termes, de même que la catégorisation des DD autonomes ne peut systématiquement se baser sur des critères syntaxiques, leur interprétation n'est pas non plus systématiquement indépendante du contexte.

Deuxièmement, et cette observation renforce la précédente, les DD autonomes sont souvent incluses dans des chaînes de coréférence. Par exemple dans (12), *La libération de Pierre-Andre Albertini* est une DD autonome qui coréfère avec *la libération de Pierre-Andre Albertini, le coopérant francais condamné à quatre ans de prison par un tribunal sud-africain*.

(12) Les services de M. Jacques Chirac ont confirmé, samedi 5 septembre, que des négociations avaient lieu pour obtenir *la libération de Pierre-André Albertini, le coopérant français condamné à quatre ans de prison par un tribunal sud-africain*. De bonne source, on laissait entendre que cette libération pourrait intervenir incessamment. L'accord qui devrait aboutir à la libération de Pierre-André Albertini a été negocié par M. Fernand Wibaux, conseiller diplomatique

tique, standard en sémantique computationelle, permet par exemple de mieux traiter de tâches telles que l'extraction d'information où il s'agit d'extraire d'un texte l'information pertinente à un individu (concret, abstrait ou générique).

du gouvernement. *La libération de Pierre-André Albertini* constituerait un "bon point" pour M. Chirac.

Pour un traitement discursif complet des descriptions définies, il importe donc de ne pas s'arrêter à la catégorisation des DD dites "autonomes" mais également d'identifier les liens de coréférence qui peuvent y être attachés.

3.3. *Les descriptions définies contextuelles*

Les descriptions contextuelles représentent 24% des descriptions annotées, 9% ayant un antécédent textuel et 15% n'en n'ayant pas.

3.3.1. *Descriptions définies associatives.*

La relation qui lie le référent d'une description définie associative à celui de son antécédent est de nature variable. (Gardent *et al.*, 2003) identifient à partir de la littérature l'ensemble de relations suivant : ensemble/sous ensemble, ensemble/élément, évènement/participant, individu/fonction, object/attribut, tout/partie, tout/morceau, objet/matière, collection/membre, endroit/lieu, évènement/sous-évènement, endroit/objet, temps/objet, predicat/argument.

On peut cependant regrouper ces différents types de cas en quatre grandes catégories :

REL : la relation entre antécédent et description définie associative est une relation prédicat/argument. Cette relation est véhiculée par le nom tête de la DD ou de l'antécédent qui est, soit un nom relationnel, soit un nom prédicatif.

(13) Deux complices *des deux malfaiteurs* qui, le 1er septembre, avaient pris en otage six personnes après *l'attaque* à main armée d'une agence bancaire à Alencon (Orne) ont été inculpés.

CIRC : la relation entre antécédent et description définie associative est une relation modifieur/modifié.

(14) [...] les moyens mis en oeuvre depuis près d'une semaine à *Besancon* et dans *la région*.

MERO : le nom tête de la DD n'est ni un nom relationnel, ni un nom prédicatif et la relation entre antécédent et description définie associative est une relation partie/tout généralisée.

(15) [...] une campagne de boycottage *des bombes à aérosols*.*Le gaz propulseur* est, en effet, fait de chlorofluorocarbones dont on pense qu'ils détruisent l'ozone de la haute atmosphère.

CIRC	MERO	MOD	REL	TOTAL
70	87	22	257	436
16.05 %	19.95%	5.04%	58.94%	100%

Figure 6. *Distribution des* DD *associatives*

MOD : la relation entre antécédent et description définie associative est donnée par un modifieur de la description définie.

(16) Victorieuse en *juillet* du Tour de France féminin, puis *le mois suivant* de la Coors Classic américaine, la sportive grenobloise a terminé détachée sur le circuit autrichien de Villach.

La sous-catégorisation des descriptions définies associatives du corpus suivant ces trois dimensions donne les résultats indiqués dans le tableau (6).

On observe que le pourcentage de descriptions définies où le nom tête (de l'antécédent ou de la description définie) est un nom relationnel ou prédicatif est élevé. Inversement, la proportion de cas impliquant la méronymie sans impliquer de nom relationnel ou prédicatif est relativement faible. Ces observations confirment l'analyse de (Löbner, 1985) selon laquelle les DD associatives font intervenir une tête nominale "FC2", c'est-à-dire, un nom dénotant une fonction à deux arguments l'un situationnel associé par défaut à tous les noms et l'autre donné par le contexte. Elles suggèrent en outre qu'un traitement adéquat des DD associatives passe par l'utilisation de ressources comme FrameNet (Baker *et al.,* 1998) où les relations prédicat/argument font l'objet de descriptions fines et qu'en contrepartie WordNet (Fellbaum, 1998) qui contient de l'information sur la synonymie, l'hyperonymie et la méronymie est quantitativement de moindre utilité[14].

Ces données montrent également que les méthodes statistiques visant à déterminer le degré d'association entre deux mots (Poesio *et al.,* 2004b, Meyer *et al.,* 2002, Bunescu, 2003) ne peuvent couvrir l'ensemble des cas de DD associatives puisque dans 16% des cas de DD associatives, le degré d'association lexicale entre anaphore et antécédent n'est pas contraint par la relation anaphorique. En effet, pour les descriptions de type CIRC, la relation entre antécédent et anaphore est une relation de type circonstant qui par définition, n'est pas spécifique au modifié (un circonstant par définition a un champ d'application très général).

14. Diverses études (Poesio *et al.,* 1997, Salmon-Alt, 2004) ont en outre montré que WordNet ne contient pas l'information méronymique de façon suffi sament systématique pour être utile à la résolution des anaphores associatives.

3.3.2. *Descriptions définies situationnelles.*

Les descriptions que nous appelons situationnelles sont des descriptions qui n'ont pas d'antécédent identifié dans le texte, mais qui doivent être rattachées au contexte pour être résolues. Les éléments du contexte qui ancrent ces descriptions définies peuvent être de deux types : soit ils appartiennent au cotexte, et c'est l'information contenue dans l'article qui permet de reconstituer l'ancre de la description définie, même s'il est impossible de trouver une ancre unique, ou même de délimiter cette ancre linguistiquement ; soit l'élément qui permet de résoudre la description appartient au contexte d'énonciation, à la situation dans laquelle le texte est produit. Dans notre cas, il s'agira systématiquement de référence à la France de septembre 1987, dans la mesure où tous nos articles sont extraits d'un journal français de septembre 1987.

Le corpus annoté contient 739 descriptions définies situationelles (soit environ 15% des DD annotées), que nous avons sous-typées en trois grandes catégories : TOPIC, LIEU et DATE (cf. figure 7).

La première catégorie (TOPIC) contient des descriptions qui n'ont donc pas d'ancre textuelle, mais dont on peut identifier le référent grâce au sujet (au sens le plus large du terme) de l'article dans lequel elles apparaissent. Dans ce cas, nous n'aurons bien entendu jamais de référence au contexte d'énonciation. Ainsi dans l'exemple ci-dessous, l'expression référentielle *l'étranger* ne signifie pas "l'étranger" mais "les investissements étrangers aux Philippines", interprétation qui n'est reconstructible qu'à partir de la situation décrite par le texte.

(17) Sur un total d' investissements représentant 210 millions de dollars pour les sept premiers mois de 1987, la part de *l'étranger* avait augmenté de 53 %.

La catégorie LIEU est la plus représentée dans notre corpus. En effet, elle contient non seulement toutes les descriptions définies qui ne peuvent être rattachées explicitement à un nom de lieu mentionné dans le texte mais plutôt à des noms de capitale, ou à des adjectifs de nationalité (exemple 18), mais également toutes les descriptions définies dont l'ancre est donnée par le contexte d'énonciation (exemple 19).

(18) (...) les Turcs devaient se prononcer, dimanche 6 septembre, pour ou contre la levée de l'interdiction de participer a la vie politique qui frappe les anciens dirigeants. La campagne pour le "non" (...) s' est intensifiée : distribution par camionnettes de photos de cadavres ensanglantés rappelant les années précédant *le coup d' Etat* (...)

(19) L'accord qui devrait aboutir à la libération de Pierre-André Albertini a été negocié par M. Fernand Wibaux, conseiller diplomatique *du gouvernement*.

Enfin, la catégorie DATE contient essentiellement des expressions temporelles dont l'interprétation est déterminée par la situation d'énonciation (dans l'exemple 20, il s'agit du 17 septembre 1987)).

TOPIC	LIEU	DATE	TOTAL
122	463	154	739
16,38%	62,15%	20,67%	100%

Figure 7. *Distribution des* DD *situationnelles*

(20) Ses fonctions exactes ne sont pas encore arrêtées (elles seront precisées lors du conseil de surveillance *du 17 septembre*), mais il est clair qu'il va renforcer la direction générale.

Dans la majorité des cas, le repérage des descriptions définies situationnelles pourra s'appuyer sur des critères syntaxiques et/ou sémantiques. En effet, les DD situationelles de la sous-catégorie DATE pourront être identifiées par les reconnaisseurs d'expressions temporelles ; les DD de sous-type LIEU sont généralement des descriptions courtes dénotant des concepts uniques par rapport à un lieu géographique ou conceptuel donné (e.g., *la police, l'Etat, les finances, le tourisme, etc.*) ; et les DD de sous-type TOPIC sont généralement des cas d'anaphores associatives sans ancres textuelles identifiables avec en particulier une forte proportion de noms relationnels ou prédicatifs (e.g. *la concurrence, les décideurs, le rythme annuel, etc.*).

Ces heuristiques nécessitent néanmoins d'être intégrées dans les résolveurs de descriptions définies et leur interprétation modélisée. En effet, dans les résolveurs existants (Poesio *et al.*, 2004a) les descriptions situationnelles sont traitées soit comme des descriptions associatives (pour lesquels le taux de succès dans la résolution est notoirement bas), soit comme des descriptions autonomes (et donc seulement partiellement interprétées).

L'interprétation de ces descriptions nécessite en outre une modélisation adéquate de la situation d'énonciation ainsi que de son sujet (topique de discours).

3.4. *Descriptions définies non référentielles*

La distribution des sous-catégories des expressions non référentielles est donnée dans la figure 8. On constate que les expressions idiomatiques sont les plus nombreuses, suivies des expressions qui appartiennent à la catégorie conjonction. Dans la mesure où toutes les expressions que nous avons classées dans la catégorie "non référentielle" sont soit des expressions (quasi) figées, soit des expressions pour lesquelles la coréférence est exprimée syntaxiquement, elles devraient poser moins de problèmes pour les résolveurs d'anaphores.

APPOSITN	CONJ.	IDIOME	PRÉDICATN	QUANTIFIEUR	TOTAL
58	126	214	34	19	451
12,75%	27,69 %	47,03 %	7,47%	4,18%	100 %

Figure 8. *Distribution des* DD *non référentielles*

4. Conclusions et perspectives

L'annotation réalisée porte sur un corpus de 48 360 mots et 4 910 descriptions définies. Elle a abouti à la définition d'une nouvelle stratégie d'annotation pour les descriptions définies ainsi qu'à la création d'un corpus proche d'un corpus de référence pour la résolution des descriptions définies. Pour finaliser la création de ce corpus de référence, il reste à annoter les chaînes de référence impliquant les descriptions définies afin de rendre compte des ambiguités de catégorisation mentionnées en section 2.1. Le corpus, ainsi que les tableaux HTML extraits à partir des annotations seront ensuite mis en ligne[15]. Par sa taille et la richesse de ses annotations, ce corpus et les fichiers associés (qui répertorient l'ensemble des exemples illustrant une catégorie ou sous-catégorie donnée) devraient permettre aussi bien l'analyse théorique linguistique que l'entraînement et l'évaluation de résolveurs ou de générateurs.

Cette étude de corpus a par ailleurs mis en relief plusieurs éléments importants pour les résolveurs de descriptions définies.

Pour les descriptions autonomes, des facteurs à la fois syntaxiques (complexité des descriptions) et sémantiques (concept abstrait, général ou unique) devraient être de bons critères de décision pour la catégorisation. L'interprétation de ces descriptions est cependant complexifiée à la fois par la présence relativement fréquente d'anaphores ou d'ellipse dans la description et par la proportion élevée de cas impliquant une DD autonome dans une chaîne de coréférence (la DD peut être interprétée hors contexte mais coréfère avec un SN antérieur).

Les descriptions coréférentielles se caractérisent par une proportion relativement élevée d'anaphores infidèles (57% des DD coréférentielles) et parmi ces anaphores infidèles, de cas où l'antécédent est un nom propre. Le traitement de ces cas est donc essentiel pour un bonne précision dans la catégorisation et passe probablement par l'utilisation conjointe d'un reconnaisseur d'entités nommées avec une ressource permettant d'approximer la notion de proximité sémantique.

Une forte proportion de noms prédicatifs ou relationnels comme têtes nominales des DD associatives (environ 59% des DD associatives) suggère qu'un bon traitement des relations prédicat/arguments devrait améliorer les performances des résolveurs sur ce type de cas.

15. Très probablement sur le site de l'ATILF, fournisseur du corpus, et en accord avec la politique de mise en place de centres de compétence par le CNRS

Enfin la catégorisation des DD situationelles requière en outre de bien reconnaître les expressions temporelles, les noms relationnels ou prédicatifs ainsi que les "uniques relatifs" c'est-à-dire les concepts dont l'unicité est relative à un lieu (conceptuel ou géographique) donné.

Remerciements

Nous remercions le programme interdisciplinaire CNRS TCAN d'avoir en partie financé les travaux présentés dans cet article. Nous remercions également les trois relecteurs de la revue T.A.L. pour leurs commentaires détaillés.

5. Bibliographie

Baker C., Fillmore C., Lowe J., « The Berkeley Framenet Project », *Proceedings of the thirty-sixth Annual Meeting of the ACL and Seventeenth International Conference on Computational Linguistics*, ACL, 1998.

Beaumont C., Lecomte J., Hatout N., « Etiquetage morpho-syntaxique du corpus "Le Monde" pour les besoins du projet PAROLE », 1998.

Blanche-Benveniste C., Chervel A., « Recherches sur le syntagme substantif », *Cahiers de lexicologie*, 1966.

Brants, « Inter-Annotater Agreement for a German Newspaper Corpus », *Proceedings of the Second International Conference on Language Resources and Engineering (LREC 2000)*, p. 1435-1439, 2000.

Bunescu R., « Associative Anaphora Resolution : A Web-Based Approach », *Proceedings of the EACL 2003 Workshop on The Computational Treatment of Anaphora*, 2003.

Carletta J., « Assessing agreement on classifi cation tasks ; the Kappa statistic », *Computational Linguistics*, vol. 22, n° 2, p. 249-254, 1996.

Cheng H., Modelling aggregation motivated interaction in descriptive text generation, PhD thesis, University of Edinburgh, 2001.

Cheng H., Poesio M., Henschel R., Mellish C., « Corpus-based NP modifi er generation », *Proceedings of the second NAACL*, Pittsburgh, 2001.

Chinchor N., Hirschmann L., « MUC-7 Coreference Task Defi nition (Version 3.0) », *Actes de MUC-7*, 1997.

Christophersen P., *The articles : A study of their theory and use in English*, Munksgaard, Copenhagen, 1939.

Clark H., « Bridging », *in* , J.-L. P.N., , W. P.C. (eds), *Thinking : Readings in Cognitive Scienc*, Cambridge University Press, 1977.

Cooper R., « The interpretation of pronouns », *in* , F. Heny, , H. Schnelle (eds), *Syntax and Semantics 10*, Academic Press, p. 61-92, 1979.

Corblin F., *Indéfini, Défini et Démonstratif*, Droz, Genève, Paris, 1987.

Corblin F., « Chaînes référentielles et communautés épistémiques », Cahiers d'Acquisition et Pathologie du Langage (Calap, N° 24, LEAPLE, Université René Descartes, 2004.

Corley M., Corley S., Crocker M., Keller F., Trewin S., « Gsearch User Manual, Revision 1.3 », n.d., http ://www.hcrc.ed.ac.uk/gsearch/.

Corley S., Corley M., Keller F., Crocker M., Trewin S., « Finding Syntactic Structure in Unparsed Corpora : The Gsearch corpus query system », *Computer and Humanities*, vol. 35, n° 2, p. 81-94, 2001.

Erk K., Kowalski A., Pado S., Pinkal. M., « Towards a Resource for Lexical Semantics : A Large German Corpus with Extensive Semantic Annotation », *Proceedings of at ACL 2003*, Sapporo, 2003.

Fellbaum C., *Wordnet. An electronic lexical database*, MIT Press, Cambridge, Mass., 1998.

Fourour N., « Identification et catégorisation automatiques des anthroponymes du Français », *Actes de Récital*, 2001.

Fradin B., *Anaphorisation et stéréotypes nominaux*, Lingua Elsevier Science Publishers, North Holland, 1984.

Fraurud K., « Definiteness and the processing of noun phrases in natural discourse », *Journal of Semantics*, 1990.

Garcia-Almanza A., *« Using WordNet for mereological anaphora resolution »*, Master's thesis, University of Essex, 2003.

Gardent C., Manuélian H., Kow E., « Which Bridges for Bridging Definite descriptions », *Proceedings of theWorkshop on Linguistically Interpreted Corpora (LINC'03), European Chapter of the Association for Computational Linguistics (EACL)*, 2003.

Groenendijk J., Stokhof M., Veltman F., « Coreference and Modality », *Handbook of Contemporary semantic theory*, Blackwell, Oxford, 1995.

Guillaume G., *Le problème de l'article et sa solution dans la langue française*, Hachette, Paris, 1919.

Hawkins J. A., *Definiteness and indefiniteness*, Humanities Press, Atlantic Highland, NJ, 1978.

Kadmon N., « Uniqueness », *Linguistics and Philosophy*, 1990.

Kamp H., « A Theory of Truth and Semantic Representation », *in* , J. Groenendijk, , T. Janssen, , M. Stokhof (eds), *Formal Methods in the Study of Language*, Mathematisch Centrum Tracts, Amsterdam, p. 277 - 322, 1981.

Karamanis N., Entity coherence for descriptive text structuring, PhD thesis, University of Edinburgh, 2003.

Kleiber G., « Des anaphores associatives méronymiques aux anaphores associatives locatives », *Verbum*, 1997.

Kleiber G., « Anaphore associative, lexique et référence, ou un automobiliste peut-il rouler en anaphore associative ? », *Anaphores pronominales et nominales*, Walter De Mulder and Co, 2001.

Kripke S., *Naming and necessity*, Harvard University Press, 1949.

Lecomte J., « Codage Multext - GRACE pour l'action GRACE / Multitag », 1997, Rapport Interne INALF.

Löbner S., « Definites », *Journal of Semantics*, 1985.

Manuélian H., Descriptions définies et démonstratives : Analyses de corpus pour la génération de textes, PhD thesis, Université de Nancy 2, 2003.

Marcus M. P., Santorini B., Marcinkiewicz M. A., « Building a large annotated corpus of English : the Penn Treebank », *Computational Linguistics*, 1993.

Meyer J., Dale R., « Mining a corpus to support associative anaphora resolution », *Proceedings of 4th Discourse Anaphora and Anaphor Resolution Colloquium*, Lisbon, Portugal, september, 2002.

Muller C., Strube M., « Annotating Anaphoric and Bridging Relations with MMAX », *2nd SIGDial Workshop on Discourse and Dialogue*, 2001a.

Muller C., Strube M., « MMAX : A tool for the annotation of multi-modal corpora », *Proceedings of the 2nd IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems*, 2001b.

Ng V., Cardie C., « Identifying anaphoric and non-anphoric noun phrases to improve coreference resolution », *Proceedings of COLING*, 2002.

Poesio M., « Associative descriptions and salience », *Proceedings of the EACL Workshop on Computational Treatments of anaphora*, 2003.

Poesio M., « Discourse annotation and semantic annotation in the GNOME corpus », *Proc. of the ACL Workshop on Discourse Annotation*, 2004a.

Poesio M., « The MATE/GNOME Scheme for Anaphoric Annotation, Revisited », *Proc. of SIGDIAL*, Boston, 2004b.

Poesio M., Alexandrov-Kabadjov M., « A general-purpose, off-the-shelf system for anaphora resolution », *Proc. of LREC*, 2004a.

Poesio M., Ishikawa T., im Walde S. S., Vieira R., « Acquiring lexical knowledge for anaphora resolution », *Proceedings of 3rd LREC*, 2002.

Poesio M., R. Mehta A. M., Hitzeman J., « Learning to resolve bridging references », *Proc. of ACL*, 2004b.

Poesio M., Uryupina O., Vieira R., Alexandrov-Kabadjov M., Goulart R., « Discourse-new detectors for definite description resolution : A survey and a preliminary proposal », *Proc. of the ACL Workshop on Reference Resolution*, 2004c.

Poesio M., Vieira R., « A Corpus-based Investigation of Definite Description Use », *Computational Linguistics*, vol. 24, n° 2, p. 183-216, 1998.

Poesio M., Vieira R., Teufel S., « Resolving bridging references in unrestricted text », *Proceedings of the ACL workshop on Robust Anaphora Resolution*, Madrid, 1997.

Prince E., « Toward a taxonomy of given-new informatio », *Radical pragmatics*, Academic Press, 1981.

Russell B., « On denoting », *Mind*, 1905.

Salmon-Alt S., « Résolution automatique d'anaphores infidèles en français : Quelles ressources pour quels apports ? », *Acte de TALN 2004*, Fès, Maroc, 2004.

Salmon-Alt S., Vieira R., « Nominal expressions in multilingual corpora : definites and demonstratives », *3rd International Conference on Language Resources and Evaluation - LREC*, Las Palmas, Spain, 2002.

Strand K., « A Taxonomy of Linking Relations », 1997, Manuscript.

Tutin A., Haddara M., Mitkov R., Orasan C., « Annotation of anaphoric expressions in an aligned bilingual corpus », *Proceedings of LREC*, 2004.

Tutin A., Trouilleux F., Clouzot C., Gaussier E., Zaenen A., Rayot S., Antoniadis G., « Annotating a large corpus with anaphoric links », *Proceedings of DAARC 2000*, 2000.

Uryupina O., « High-precision identification of discourse-new and unique noun phrases », *Proceedings of the ACL 2003 Student Workshop*, p. 80-86, 2003.

van Deemter K., Kibble R., « On Coreferring : Coreference in MUC and related annotation schemes. », *Computational Linguistics*, 2000.

venex, « The Venezia / Essex (VENEX) corpus of Italian Anaphora : Instructions for annotating anaphora and deixis in Italia », 2005. The VENEX Project.

Vieira R., Poesio M., « An empirically-based system for processing definite descriptions », *Computational Linguistics*, 2000.

Webber B., A formal approach to discourse anaphora, PhD thesis, Harvard University, 1978.

Westerstahl D., « Determiners and context sets », *in* , J. van Benthem, , A. T. Meulen (eds), *Generalised quantifiers in natural language*, Foris, Dordrecht, 1991.

SERVICE ÉDITORIAL – HERMES-LAVOISIER
14 rue de Provigny, F-94236 Cachan cedex
Tél : 01-47-40-67-67
E-mail : revues@lavoisier.fr
Serveur web : http://www.revuesonline.com

ANNEXE POUR LE SERVICE FABRICATION

A FOURNIR PAR LES AUTEURS AVEC UN EXEMPLAIRE PAPIER
DE LEUR ARTICLE ET LE COPYRIGHT SIGNE PAR COURRIER
LE FICHIER PDF CORRESPONDANT SERA ENVOYE PAR E-MAIL

1. ARTICLE POUR LA REVUE :
   L'objet. Volume 8 – n°2/2005

2. AUTEURS :
   Claire Gardent — Hélène Manuélian***

3. TITRE DE L'ARTICLE :
   Création d'un corpus annoté pour le traitement des descriptions définies

4. TITRE ABRÉGÉ POUR LE HAUT DE PAGE MOINS DE 40 SIGNES :
   Un corpus annoté de descriptions définies

5. DATE DE CETTE VERSION :
   3 novembre 2005

6. COORDONNÉES DES AUTEURS :

– adresse postale :

* CNRS
Laboratoire Lorrain de recherche en informatique et ses applications
Campus Scientifique BP 239
54506 Vandoeuvre-lès-Nancy Cedex
Claire.Gardent@loria.fr

** Laboratoire MétaDIF
Université de Cergy Pontoise
33 boulevard du port
95 000 Cergy Pontoise
helene.manuelian@lsh.u-cergy.fr

– téléphone : 00 00 00 00 00
– télécopie : 00 00 00 00 00
– e-mail : Roger.Rousseau@unice.fr

7. LOGICIEL UTILISÉ POUR LA PRÉPARATION DE CET ARTICLE :
   LaTeX, avec le fichier de style `article-hermes.cls`,
   version 1.2 du 03/03/2005.

8. FORMULAIRE DE COPYRIGHT :
   Retourner le formulaire de copyright signé par les auteurs, téléchargé sur :
   `http://www.revuesonline.com`

SERVICE ÉDITORIAL – HERMES-LAVOISIER
14 rue de Provigny, F-94236 Cachan cedex
Tél : 01-47-40-67-67
E-mail : revues@lavoisier.fr
Serveur web : http://www.revuesonline.com