

Manuel d'annotation pour le corpus Dedé : le
corpus de descriptions définies.

Claire Gardent et Hélène Manuélian

22 juin 2006

Table des matières

| | | |
|----------|---|-----------|
| 1 | Introduction | 3 |
| 2 | Le corpus et les outils de traitement | 3 |
| 2.1 | Le corpus PAROLE | 3 |
| 2.2 | G-search tool | 4 |
| 2.3 | MMAX | 4 |
| 3 | Descriptions à annoter | 6 |
| 3.1 | Définition et délimitation des descriptions à annoter | 6 |
| 3.2 | Exemples des différentes configurations syntaxiques où les descriptions définies apparaissent | 7 |
| 4 | Typage des descriptions définies | 7 |
| 4.1 | Descriptions autonomes | 8 |
| 4.1.1 | Anaphorique : | 8 |
| 4.1.2 | Unique : | 8 |
| 4.1.3 | Termes généraux : | 9 |
| 4.1.4 | Identifiante : | 9 |
| 4.2 | Descriptions coréférentielles | 9 |
| 4.2.1 | Coréférence directe | 10 |
| 4.2.2 | Coréférence Lexicale | 10 |
| 4.2.3 | Redescription | 11 |
| 4.3 | Descriptions contextuelles | 11 |
| 4.4 | Descriptions définies associatives | 12 |
| 4.5 | Descriptions définies situationnelles. | 14 |
| 4.6 | Descriptions non référentielles | 15 |
| 4.7 | Deux phases d'annotation | 16 |
| 5 | Repérage des antécédents | 18 |
| 5.1 | Identification de l'antécédent | 18 |
| 5.2 | Deux relations (pointer et member) | 18 |

1 Introduction

Le but de ce corpus est d'aider à l'élaboration de systèmes de résolution d'anaphores. L'annotation a été réalisée pour permettre l'interprétation des descriptions définies. Notre schéma d'annotation étiquette les descriptions définies et non simplement les relations entre les expressions référentielles. En effet, l'étiquetage des liens entre les expressions référentielles correspond à un autre type de tâche (comme dans le projet MUC par exemple), qui implique la totalité de la chaîne de référence.

L'étiquetage permet de dire si un antécédent doit être recherché pour la description définie et de quel type il est s'il existe : est-il dans une relation d'identité ou d'association avec la description définie ?

Il y a alors trois cas de figure possibles pour que le référent soit identifié :

1. Soit la description définie n'a pas besoin d'antécédent pour être interprétée. Ce sont les définis sémantiques de Lobner, ils incluent entre autres les noms propres. Le référent fait partie des connaissances de l'auditeur ou alors il est accommodé. Les connaissances de la situation ou le contexte ne sont pas utilisés pour interpréter la description.
2. Pour que le référent de la description définie soit identifié, il est nécessaire de la relier à une autre description du contexte référant au même objet.
3. Pour que le référent de la description définie soit identifié, il est nécessaire de la relier à une autre description du contexte référant à une autre entité du discours introduite dans le texte ou saillante dans la situation d'énonciation.

2 Le corpus et les outils de traitement

2.1 Le corpus PAROLE

Le corpus que nous avons annoté est une sous-partie du corpus PAROLE¹ et comprend 48 3600 mots annotés au niveau morphosyntaxique. Il est composé d'une série d'articles du journal *Le Monde* datant de septembre 1987. Ces articles appartiennent à toutes les rubriques du journal (Politique nationale et internationale, Economie, Sport, Culture et Loisirs). Le corpus est balisé mot à mot suite à l'annotation pour le projet PAROLE. Ceci signifie que dans le fichier qui contient le texte, on trouve des indications morphosyntaxiques concernant chaque mot du corpus.

Chaque balise comporte plusieurs informations, en nombre différent selon la catégorie du mot annoté. Ces étiquettes sont héritées du schéma d'annotation Multext/Multitag pour l'action GRACE [15, 1].

¹Corpus 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.

Les déterminants sont annotés sur 7 positions, indiquant respectivement la catégorie, le type, la personne, le genre, le nombre, le possesseur, la quantification. Les champs peuvent être vides s'ils ne sont pas pertinents. Les définis seront annotés de la façon suivante :

| |
|--------------------------|
| <w msd Da-ms- d> ce </w> |
|--------------------------|

Dans l'exemple ci dessus, on doit lire que le mot (w) a la description morphosyntaxique (msd) suivante : c'est un déterminant (D), de type article (a), qu'il n'indique pas de personne (-), qu'il est masculin (m), singulier (s), qu'il n'y a pas d'indication du possesseur (-), et qu'il est défini (d).

La précision de l'annotation morphosyntaxique nous a été particulièrement utile dans l'utilisation des outils présentés dans les sections suivantes.

Nous décrivons ici brièvement les outils que nous avons utilisés dans notre travail sur corpus, puis le détail des traitements effectués pour passer du format de départ du corpus à un format permettant son annotation au niveau référentiel².

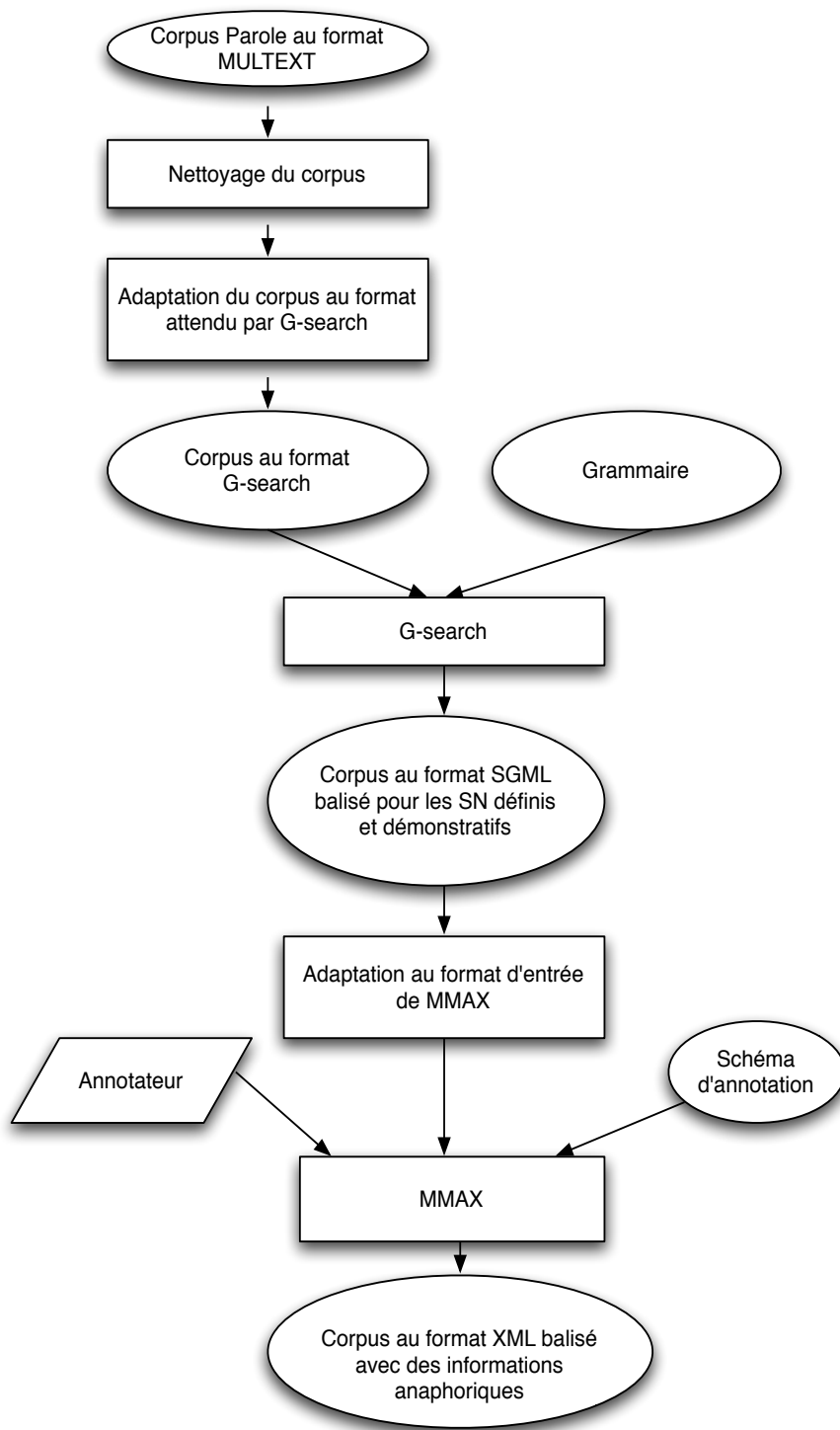
2.2 G-search tool

G-search tool [6, 5] est un outil qui permet d'identifier des structures syntaxiques dans un corpus annoté au niveau morphosyntaxique (i.e. dans lequel on a associé une partie du discours à chaque mot). Il permet de retrouver ces structures grâce à une grammaire définie par l'utilisateur (cf. section suivante). Les éléments terminaux de la grammaire sont des expressions régulières sur les éléments du corpus (mots, lemmes, étiquettes morphosyntaxiques). Le corpus dont nous disposons ayant été annoté au niveau morphosyntaxique de façon très fine, nous avons pu identifier les groupes nominaux nous intéressant et les isoler de façon à les repérer facilement dans une fenêtre de notre outil d'annotation, MMAX.

2.3 MMAX

L'outil d'annotation que nous avons utilisé, MMAX, a été conçu spécifiquement pour annoter manuellement les corpus au niveau référentiel, et plus précisément les relations de coréférence, d'anaphore associative et les corpus multimodaux [16, 17]. Dans une fenêtre de l'application, on peut lire le texte à annoter. Par un simple système de sélection à la souris et de clic, on insère des balises XML qui permettent d'identifier le type de relation anaphorique et l'antécédent du syntagme considéré. L'intérêt d'utiliser ce type d'outil est d'une part qu'il facilite la tâche de balisage en XML, et d'autre part qu'il permet de transformer le corpus dans un format électronique standard, XML, qui permet la réutilisation du corpus.

²Cette partie du travail a été réalisée en collaboration avec Eric Kow, qui est l'auteur de l'intégralité des scripts conçus pour le traitement du corpus.



3 Descriptions à annoter

3.1 Définition et délimitation des descriptions à annoter

Les descriptions définies à annoter ont été identifiées par Gsearch sur la base de l'annotation morphosyntaxique issue de l'action GRACE décrite précédemment et de la grammaire présentée en figure 1. Les éléments à annoter seront donc tous les groupes nominaux dont le déterminant est soit *le*, soit *la*, soit *les*, ainsi que les groupes nominaux précédés des prépositions contractées (au, aux, des, du), à côté desquelles on a fait figurer le mot TRACE afin de matérialiser le déterminant. Les groupes nominaux sont annotés systématiquement, même lorsqu'ils figurent dans du discours rapporté.

```
File:          GrammarFrenchDefNP
Purpose:       Simple Gsearch grammar for French (NPs)
               without embedded NPs
Time-stamp:   <2002-13-04 13:41:32 kowey>

VP --> V1
VP --> V1 NP
VP --> V1 PP
V1 --> ADV* VERB ADV*

NP --> defNP
NP --> demNP
NP --> otherNP
defNP --> defDET N1+
demNP --> demDET N1+
otherNP --> otherDET N1+
NP --> N1+ PP*
NP --> NP CONJ NP

N1 --> AP* NOUN+ AP*
N1 --> N1 CONJ N1

AP --> ADV* ADJ
AP --> AP CONJ AP

PP --> PREP NP

----- terminals

defterm "msd"          Saves writing

VERB --> <"V.*">
NOUN --> <"N.*">
PREP --> <"S.*">

ADV --> <"R.*">
ADJ --> <"A.*">

defDET --> <"Da.*d">
demDET --> <"Dd.*">
otherDET --> <"D[~ads].*[~d]">

CONJ --> <"Cc.*">
```

FIG. 1 – Grammaire pour G-search

3.2 Exemples des différentes configurations syntaxiques où les descriptions définies apparaissent

Les groupes nominaux imbriqués sont annotés normalement. Ainsi, dans une séquence comme [*le malaise₁ dans l'armée₂ ... le malaise₁ dans l'armée₂*], on annote les deux liens de coréférence indiqués par les indices.

Les phrases à copules sont elles aussi annotées le plus simplement possible : on annote le groupe nominal défini attribut du sujet normalement et on ne l'annote pas en coréférence avec le sujet (on considère que la relation est donnée syntaxiquement). [*Sannier , dont le sport favori est le velo*] est un exemple type de cette configuration.

Appositions Dans la séquence [*Mme Valade , épouse du ministre de la recherche et de l'enseignement supérieur , a baptisé le 4 septembre a Concarneau (Finistère), l' Alis, le nouveau navire oceanographique de l' ORSTORM*], le syntagme [*le nouveau navire oceanographique de l' ORSTORM*] est annoté comme une description définie, mais ne sera pas considérée comme référentielle pour les mêmes raisons que l'attribut du sujet mentionné précédemment.

Expressions temporelles Des expressions comme [*au cours de..., au moment où...*] seront elles aussi annotées.

Coordination et Listes On annotera séparément les deux expressions définies dans une séquence comme [*Le Président de la République et le Premier Ministre...*], ou dans la séquence [*Le Président de la République, le Premier Ministre et le reste du gouvernement...*]

4 Typage des descriptions définies

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-linguistique³(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).

³La notion de *référence* adoptée dans ce travail est celle issue des travaux sur la sémantique discursive (cf. [23, 12]). 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 spécifiées 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.

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")

- (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.

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.

4.1 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 après résolution des anaphores ou ellipses éventuellement présentes dans les modificateurs du nom tête. La balise utilisée est la suivante `<type = autonomous>` pour toutes les descriptions définies appartenant à cette catégorie. Des balises indiquant des sous-types sont ajoutées ensuite.

4.1.1 Anaphorique :

Cette catégorie recouvre les **descriptions contenant des anaphores ou des ellipses dont l'interprétation est donnée par le contexte**, les descriptions contenant des **déictiques** ou des expressions **anaphoriques**, les descriptions contenant des descriptions situationnelles

La femme qu'il a rencontré, la même chose que lui, l'ensemble de sa gamme nouvelle, les problèmes juridiques que cela va poser

La fin de l'année,

les prix à la consommation (en France)

La balise utilisée pour cette catégorie est la suivante : `<autonomous_subtype = anaphor>`

4.1.2 Unique :

Dans cette catégorie, nous classons (1) les **unica** de Russell, c'est-à-dire, les descriptions référant à des objets uniques par définition :

Le soleil, la lune, le pape, etc.

(2) les descriptions dont la tête nominale contient un **nom propre** ou des items (nombre, mot) fonctionnant comme des noms propres au sens où ils dénotent une entité unique invariable dans le temps :

L'année 1984, le mot "le", la République Populaire de Chine, le président Chirac, le lundi 4 septembre 1984

La balise utilisée est la suivante <autonomous_subtype = unique>

4.1.3 Termes généraux :

Ce type de description autonome regroupe les descriptions dénotant des **concepts abstraits** : *La sécheresse, le pouvoir, la loyauté*, les **noms d'espèces** : *le pissalat*, les **termes génériques** paraphrasables par *tous les X* : *le touriste, les Français, le français, les libéraux*

La balise utilisée est la suivante <autonomous_subtype = gen>

4.1.4 Identifiante :

Dans cette catégorie, sont classés (1) les **noms suivis d'une complétive** : *Le fait que Marie soit partie, la question de savoir si Marie est partie, etc.*, (2) les **noms suivis d'un modifieur** permettant d'identifier le référent désigné indépendamment du contexte :

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, le parlement iranien, la filière boraine, ... etc.

Les concepts abstraits modifiés (e.g., *l'honneur soviétique*) sont classés en autonome-terme général.

La balise utilisée est la suivante <autonomous_subtype = identif>

4.2 Descriptions coréférentielles

Une description coréférentielle spécifie un référent déjà introduit dans le contexte linguistique par un groupe nominal antécédent. [8, 11, 19] 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. [3, 20] 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 (2a). Les reprises indirectes incluent les reprises via une relation lexicale (de synonymie (2b), d'hyponymie (2c) ou d'hyponymie en (2d)) et les redescriptions [8] (ou épithètes, [20]). Dans ce cas (2e), 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.

- (2) a. un homme/l'homme
- b. un putsch/le coup d'Etat
- c. deux malfaiteurs/les hommes
- d. six hommes/les truands
- e. UCAR .. Duracell/les deux concurrents

Dans l’annotation, on utilisera une balise à part pour la coréférence. L’annotateur devra choisir entre `<coref = yes>` et `<coref = no>` pour toutes les descriptions annotées. En effet, les descriptions définies peuvent être à la fois autonomes et coréférentielles, par exemple. Ainsi, dans une séquence du type [Le Parti Socialiste, le Parti Socialiste, ... le PS,... etc...], la description employée est la même. On considère donc à la fois que la description est autonome (dès la première mention, le référent est identifiable uniquement), ce qui n’empêche pas ces expressions de coréférenter à la même entité. Une autre balise sera utilisée pour le sous-typage, en supplément de `coref=yes`. Le sous-typage adopté est décrit dans les sections suivantes.

4.2.1 Coréférence directe

Une description coréférentielle directe se caractérise par la reprise d’un mot apparaissant dans son antécédent.

- (3) La radio privée Star-System, la dernière station qui continuait à émettre sans autorisation sur *la bande FM parisienne* a été saisie, le jeudi 3 septembre dans la soirée. *La bande FM* est donc maintenant complètement nettoyée entre 88 et 106.

La reprise peut faire intervenir un changement de catégorie grammaticale ou une ellipse :

- (4) a. Les élections législatives_{Adj} / Les législatives_N
 b. Les législatives / les élections (Ellipse de «élections»)

On utilisera la marque `<coref-type= direct>` pour annoter ces descriptions définies.

4.2.2 Coréférence Lexicale

Une description coréférentielle lexicale fait intervenir une relation lexicale entre antécédent et description définie. Dans le cas le plus simple, la relation lexicale intervient entre le nom tête de l’antécédent et le nom tête de la description définie :

- (5) a. la ville / la capitale (hyperonymie)
 b. le vélo / la bicyclette (synonymie)

On inclut également dans cette catégorie les cas où l’antécédent est un nom propre et la description définie indique le type de ce nom propre :

- (6) a. Paris / la ville
 b. Paris / la capitale

Le sous-type `lexical` inclut également les alias, les cas où la relation entre antécédent et description définie est une relation définitionnelle (la description est une définition de l’antécédent ou vice versa) et les cas de synonymie non lexicale :

- (7) a. le SRPJ/le service
- b. la libération simultanée de Pierre-André Albertini et d’un militaire sud-africain/l’échange
- c. le climat pesant/la tension latente

Cette catégorie sera marquée par la balise `<coref-type= lexical>`.

4.2.3 Redescription

La catégorie `redescription` inclut tous les autres cas : les paires nom propre/fonction, les paires sans relation spécifique entre les noms têtes de l’antécédent et de la description et les cas où la reprise se situe au niveau métalinguistique :

- (8) a. Jacques Chirac/le président
- b. Sept personnes ont été tuées/les morts
- c. célèbre/le mot

On utilisera la marque `<coref-type= redesc>` pour annoter ces descriptions définies.

4.3 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 une entité explicitement introduite dans le contexte linguistique⁴. Ainsi dans (9), 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*.

- (9) *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 [18]. 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 (10a) mais non en (10b).

- (10) a. Italie : Le gouvernement a décidé...

⁴Dans la littérature, les descriptions contextuelles avec antécédent ont également été dénommées anaphores associatives [13, 14], *bridging anaphora* [3] et *inferrables* [19]. Le premier à utiliser la notion d’association est [10]. [4] note que le terme est repris dans [2] et que [11] l’utilise de façon indépendante pour l’anglais. Le phénomène d’anaphore associative en français est étudié par [7].

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

Pour rendre compte de cas tels que (10b) 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 10b ci-dessus), la description *le Parlement* sera annotée comme situationnelle pour refléter le fait que le référent désigné est *le parlement de l’Italie*.

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 [18], 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 dépendance contextuelle de ces expressions pour leur interprétation.

Nous utiliserons respectivement pour les descriptions associatives et situationnelles les marques `<type= bridging>` et `<type= situational>`. Chacune des catégories portera ensuite une balise de sous-typage.

4.4 Descriptions définies associatives

La relation qui lie le référent d’une description définie associative à celui de son antécédent est de nature variable. [9] identifient à partir de la littérature l’ensemble de relations suivant : ensemble/sous ensemble, ensemble/élément, événement/participant, individu/fonction, objet/attribut, tout/partie, tout/morceau, objet/matière, collection/membre, endroit/lieu, événement/sous-événement, endroit/objet, temps/objet, prédicat/argument.

Le sous-typage adopté utilise le typage sémantique (individu ou événement), la relation d’implication (entre événement et individu ou entre individus). Comme par ailleurs les catégories choisies ne sont pas mutuellement exclusives, les catégories sont ordonnées et en cas de conflit, la catégorie la plus forte est choisie.

Les sous-types sont les suivants.

Relationnelles : `<bridging_subtype= REL>`

- l’antécédent ou la description définie dénote une éventualité (état ou événement) *et*
- 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.

- (11) 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.

Méronymiques : <bridging_subtype= MERO>

- l’antécédent et la description définie dénotent des individus (abstrait ou concret) *et*
- la relation entre antécédent et description définie associative est une relation d’implication.

Cette relation inclut la relation partie/tout (paraphrasable par *fait partie de*), la relation ensemble/sous-ensemble, la relation ensemble/élément, la relation individu/attribut et la relation individu/fonction.

- (12) [...] 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.

Modifieurs : <bridging_subtype= MOD> la relation entre antécédent et description définie associative est donnée par un modifieur de la description définie.

- (13) 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.

Circonstancielle : <bridging_subtype= CIRC> la relation entre antécédent et description définie associative est soit une relation modifieur/modifié, soit une relation non nécessaire mais impliquée par le contexte discursif. Concrètement, *circ* inclut non seulement les spécifications de lieu ou de temps mais également toutes les relations non recensées par les catégories relationnelle, méronymique et modifieurs mais qui sont nécessaires pour rendre compte de l’unicité des DD e.g.,

- (14) Toutes les conditions paraissent réunies pour que le patronat traverse de nouveau une grave crise de paranoïa : dans quelques mois, il risque de se retrouver avec un président qu’il juge incompetent et qui n’est pas de leur bord. C’est là où commence la véritable surprise de ce sondage : *la sérénité* a remplacé la crainte. Plus des deux tiers des dirigeants pensent qu’ils n’ont rien à redouter. *sérénité (affichée par) le patronat*
- (15) Dans la soirée, conférence de presse de *M. Duon Sadia*, ministre du tourisme. Il n’ésquive pas le débat. Recourt aux aphorismes : “Lorsqu’on est devant un tam-tam, il vaut mieux battre le tam-tam plutôt que de battre son ventre.” Traduire : parlons franc, les médias répercuteront. Définit son objectif : passer de 200 000 à 400 000 touristes par an. “Et *la confiance* que voici. Il y a un masque, un masque poro qui ne sort que tous les trente ans.” *confiance (faite par) M. Duon Sadia*

N.B. Lorsque la relation entre description définie et antécédent est donnée par la syntaxe, la description n’est pas classée comme associative.

- (16) Les canalisations ont éclaté, des centraux téléphoniques se sont arrêtés de fonctionner, et *les coupures d’électricité* se sont prolongées, notamment dans [la partie sud de la ville], pendant plus de dix heures. *les coupures d’électricité dans Dehli*

4.5 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.

Les descriptions définies situationnelles sont sous-typées en trois grandes catégories : TOPIC, LIEU et DATE. Ces trois catégories seront marquées dans l'annotation par les balises `<sit_subtype= Topic>`, `<sit_subtype= Loc>` et `<sit_subtype= date>`.

La 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 reconstituable 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 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 à 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 négocié par M. Fernand Wibaux, conseiller diplomatique *du gouvernement*.

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)).

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

On inclut également dans les catégories LIEU et DATE, les expressions (e.g., *l'an dernier*) dont l'interprétation est dépendante d'un indexical de lieu (e.g., *ici*) ou de temps (*maintenant*).

4.6 Descriptions non référentielles

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

Elles seront typées par la balise <Type = non-referential>.

C'est le cas en particulier, lorsque qu'une DD apparaît dans un usage attributif (21a), dans une structure prédicative (21b), une apposition (21c). 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 critiquable [21, 22]. Ces SN servent en effet à construire une prédication. Ils ne dénotent ni un argument ni un modifieur et leur interprétation est indépendante 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 (21d), d'une conjonction (21f) ou d'un quantifieur (21e).

- (21) a. Jean cherche *la meilleure méthode d'annoter les pronoms*.
b. Les Etats-Unis et le Japon continuent à 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.

Les catégories décrites précédemment seront sous-typées grâce aux marques suivantes :

- Attributs et prédications : <nonref_subtype= Pred>
- Appositions : <nonref_subtype= apposition>
- Expressions figées : <nonref_subtype= idiom>

- Emploi conjonctif : <nonref_subtype= conj>
- Emploi quantifieur : <nonref_subtype= quant>

4.7 Deux phases d’annotation

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 ambiguïté. En effet, comme le remarque [18], certains désaccords dans l’annotation résultent d’une ambiguïté de catégorisation : dans le contexte considéré, une même description définie peut être catégorisée de plusieurs façons, toute correctes. Ainsi dans (22), 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*.

- (22) "Nous ne connaissons pas nous-mêmes les intentions de Jérôme 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 désambiguïsation 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 annotées⁵.

⁵L’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 solveurs 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 contexte.

- 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.

Elle permet d’autre part, d’aplanir les différences d’annotation qui peuvent intervenir entre les catégories ASSOCIATIVE/SITUATIONNELLE d’une part, et CORÉFÉRENTIELLE d’autre part. En effet, dans les cas comme (22) ci-dessus où une description peut être annotée soit comme ASSOCIATIVE/SITUATIONNELLE, soit comme CORÉFÉRENTIELLE, la double annotation ASSOCIATIVE/SITUATIONNELLE + CORÉFÉRENTIELLE 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 conflits possibles et les critères de décisions adoptés⁶.

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).

FIG. 2 – Stratégie de résolution des conflits de catégorisation possibles

⁶L’interprétation de la plupart des noms communs est relative au temps et à l’espace. Lorsque le contexte fixe 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.

5 Repérage des antécédents

5.1 Identification de l'antécédent

Il peut y avoir deux cas de figure. Le premier, et le plus simple, est le cas où l'antécédent est une description définie. Dans ce cas, la balise <markable> délimite déjà l'élément et peut servir d'antécédent (exemple 23a). Le second cas est plus compliqué : l'antécédent n'est pas une description définie. L'annotateur doit alors utiliser la procédure de création de balise <markable>. On n'annotera de préférence qu'un seul nom comme antécédent (exemple 23b). A partir du moment où l'antécédent n'est pas nominal, on aura tendance à privilégier l'interprétation situationnelle, sauf dans des cas où l'antécédent est identifié très clairement (exemple 23c). Les antécédents peuvent être marqués dans les titres comme dans le corps du texte de l'article.

- (23) a. Asphyxiées par les émanations de la circulation automobile et les fumées d'usines soudain bloquées au-dessus de *la ville*, quelque soixante personnes ont dû être transportées d'urgence à *l'hôpital*
- b. Une brusque montée de l'hygrométrie et une absence totale de vent ont provoqué à *Barcelone*, dans la nuit du 4 au 5 septembre, une série d'intoxications, dont deux mortelles. Asphyxiées par les émanations de la circulation automobile et les fumées d'usines soudain bloquées au-dessus de *la ville*,(...).
- c. *La Lainière va peut-etre supprimer des cars de ramassage!* Pour ces ouvrières du bassin houillier dont quelques-unes ont déjà trois heures de transport par jour, *la nouvelle - pour l' instant simple rumeur -* a relégué au second plan les manoeuvres boursières dont leur entreprise fait l'objet depuis deux mois .

5.2 Deux relations (pointer et member)

MMAX permet de distinguer deux types de relations entre les groupes nominaux. L'une de ces relations est une relation binaire, tandis que l'autre relation permet d'impliquer autant d'éléments qu'on le souhaite. La première, la relation pointer, permettra d'annoter les antécédents d'anaphores associatives, tandis que la seconde, nous permettra d'annoter les chaînes de coréférence.

Relation pointer : descriptions définies associatives La relation dénotée par le terme "pointer" est une relation intransitive, orientée et binaire. A l'écran, elle est matérialisée par un trait bleu entre l'anaphore et son antécédent. Comme il s'agit d'une relation strictement binaire et orientée, nous ne l'utilisons que pour signifier le lien entre une description associative et son antécédent.

Relation member : chaînes de coréférence La relation dénotée par le terme "member" est une relation transitive. A l'écran, elle est matérialisée par

un trait rouge entre l’anaphore et son antécédent. Comme il s’agit d’une relation transitive, elle permet d’annoter des chaînes de référence facilement. Nous l’utilisons donc pour annoter les liens de coréférence. Une seule exception à cette règle réside dans les coréférences avec un antécédent double. En effet, la relation “member” étant transitive, il est impossible de l’utiliser pour indiquer correctement qu’une description définie a un double antécédent. Nous montrons dans la figure 3 la façon dont les antécédents doubles sont annotés en coréférence.

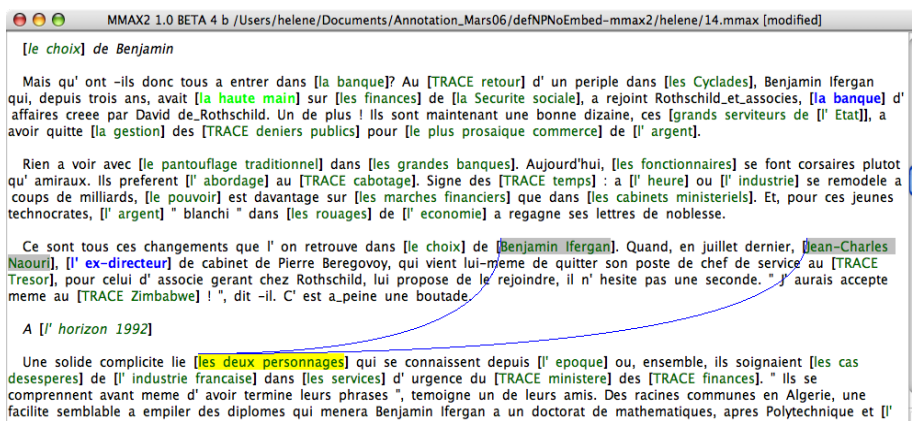


FIG. 3 – Annotation des antécédents doubles dans les relations de coréférence

La figure 4 montre la façon dont l’antécédent double serait annoté si l’on conservait la relation member dans ce cas particulier. On peut voir clairement sur l’image que la relation matérialisée par les traits rouges n’indique pas clairement que les deux noms propres sont les antécédents de la description définie.

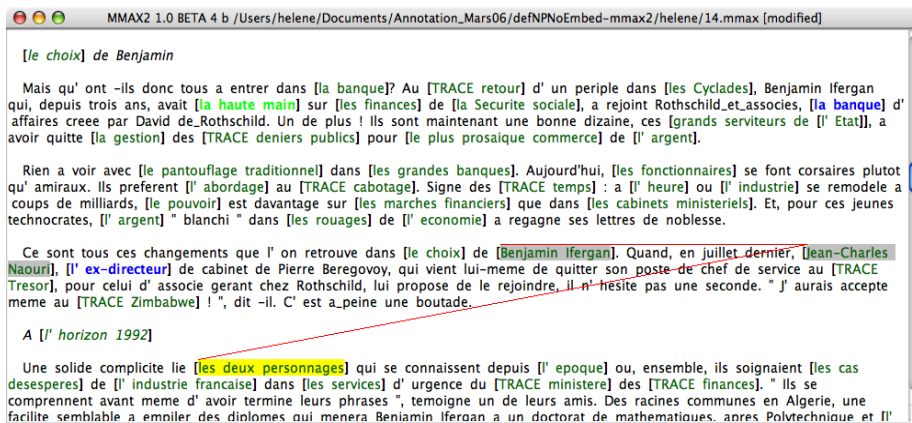


FIG. 4 – Résultat de l’annotation d’un double antécédent avec la relation member

Références

- [1] C. Beaumont, J. Lecomte, and N. Hatout. Etiquetage morpho-syntaxique du corpus *le monde* pour les besoins du projet parole. Rapport Interne INALF, 1998.
- [2] Claire Blanche-Benveniste and André Chervel. Recherches sur le syntagme substantif. *Cahiers de lexicologie*, (IX-2), 1966.
- [3] H. Clark. Bridging. In Johnson-Laird P.N. and Wason P.C., editors, *Thinking : Readings in Cognitive Scienc.* Cambridge University Press, 1977.
- [4] F. Corblin. *Indéfini, Défini et Démonstratif*. Droz, Genève, Paris, 1987.
- [5] M. Corley, S. Corley, M. Crocker, F. Keller, and S. Trewin. Gsearch user manual, revision 1.3. <http://www.hcrc.ed.ac.uk/gsearch/>.
- [6] S. Corley, M. Corley, F. Keller, M. Crocker, and S. Trewin. Finding syntactic structure in unparsed corpora : The gsearch corpus query system. *Computer and Humanities*, 35(2) :81–94, 2001.
- [7] B. Fradin. *Anaphorisation et stéréotypes nominaux*. Lingua Elsevier Science Publishers, North Holland, 1984.
- [8] Kari Fraurud. Definiteness and the processing of noun phrases in natural discourse. *Journal of Semantics*, 1990.
- [9] C. Gardent, H. Manuélian, and E. Kow. Which bridges for bridging definite descriptions. In *Proceedings of the Workshop on Linguistically Interpreted Corpora (LINC'03), European Chapter of the Association for Computational Linguistics (EACL)*, 2003.
- [10] G. Guillaume. *Le problème de l'article et sa solution dans la langue française*. Hachette, Paris, 1919.
- [11] John A. Hawkins. *Definiteness and indefiniteness*. Humanities Press, Atlantic Highland, NJ, 1978.
- [12] H. Kamp. A theory of truth and semantic representation. In J. Groenendijk, Th. Janssen, and M. Stokhof, editors, *Formal Methods in the Study of Language*, pages 277 – 322. Mathematisch Centrum Tracts, Amsterdam, 1981.
- [13] G. Kleiber. Des anaphores associatives méronymiques aux anaphores associatives locatives. *Verbum*, 1997.
- [14] Georges Kleiber. Anaphore associative, lexique et référence, ou un automobiliste peut-il rouler en anaphore associative? In *Anaphores pronominales et nominales*. Walter De Mulder and Co, 2001.
- [15] J. Lecomte. Codage multext - grace pour l'action grace / multitag. Rapport Interne INALF, 1997.
- [16] C. Muller and M. Strube. Annotating anaphoric and bridging relations with mmax. In *2nd SIGDial Workshop on Discourse and Dialogue*, 2001.
- [17] C. Muller and M. Strube. Mmax : A tool for the annotation of multimodal corpora. In *Proceedings of the 2nd IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems*, 2001.

- [18] Massimo Poesio and Renata Vieira. A corpus-based investigation of definite description use. *Computational Linguistics*, 24(2) :183–216, 1998.
- [19] Ellen Prince. Toward a taxonomy of given-new informatio. In *Radical pragmatics*. Academic Press, 1981.
- [20] K. Strand. A taxonomy of linking relations. Manuscript, 1997.
- [21] K. van Deemter and R. Kibble. On coreferring : Coreference in muc and related annotation schemes. *Computational Linguistics*, 26(4), 2000.
- [22] venex. The venezia / essex (venex) corpus of italian anaphora : Instructions for annotating anaphora and deixis in italia. The VENEX Project, 2005.
- [23] B. Webber. *A formal approach to discourse anaphora*. PhD thesis, Harvard University, 1978.