Annotation en relations anaphoriques d'un corpus de discours oral spontané en français

Judith Muzerelle, Emmanuel Schang, Jean-Yves Antoine, Iris Eshkol, Denis Maurel, Aurore Boyer-Pelletier, Damien Nouvel

Annotations en chaînes de coréférences et anaphores dans un corpus de discours spontané en français

Judith Muzerelle (1), Emmanuel Schang (2), Jean-Yves Antoine (3,4), Iris Eshkol (2), Denis Maurel (3), Aurore Boyer (1) et Damien Nouvel (3)

1 Université François Rabelais – Tours (LLL-Tours)
2 Université d'Orléans (LLL-Orléans)
3 Université François Rabelais – Tours (LI)
4 Lab-STICC CNRS – Lorient (Université Européenne de Bretagne)

Résumé : cet article présente une analyse des relations anaphoriques d'un corpus de dialogue oral spontané en français. Il exposera plus particulièrement l'étude pilote CO2, qui a conduit à une procédure d'annotation de corpus, puis deux expériences issues du corpus (accord en genre et en nombre, descriptions des définis en première mention), et enfin les travaux à venir du projet ANCOR. L'objectif de celui-ci est d'évaluer la pertinence et de modéliser les processus de résolution de ces anaphores complexes en discours spontané.

1 Introduction

Cet article propose un effort important en matière d'annotation des anaphores en corpus de français parlé spontané, ainsi que des études distributionnelles. Celles-ci démontrent l'intérêt de la linguistique de corpus pour orienter les recherches sur la résolution des anaphores par le traitement automatique des langues naturelles (TALN).

Au cours des deux dernières décennies, l'ingénierie des langues a connu des avancées spectaculaires qui ont permis l'émergence de nombreuses applications opérationnelles destinées aussi bien au grand public qu'aux professionnels. Parmi ces technologies langagières, la recherche d'information et l'indexation de documents constituent sans nul doute un des champs applicatifs promis au plus bel avenir. En effet, la croissance exponentielle des ressources textuelles ou multimédias accessibles sur Internet nécessite la mise en place d'outils de structuration et d'interrogation automatique intelligents. Pour ne citer qu'un exemple, la quasi-intégralité des articles publiés sur la Toile par les quotidiens de la presse nationale ou régionale font l'objet d'une indexation automatique.

La qualité des outils d'indexation ou d'interrogation développés pour ces tâches dépend dans une large mesure de leur robustesse en matière de détection des entités nommées. Une entité nommée est une unité linguistique qui désigne un élément précis de l'univers du discours. Ce peut-être un nom propre (*Sarkozy* ; *France*), un mot polylexical (*le président du directoire*), mais également une mesure (*un prix*, par exemple) ou encore une date. La détection des entités nommées, désignant le plus souvent les éléments sur lesquels porte le discours, est donc essentielle dans les applications d'extraction ou de recherche d'information textuelle : elles répondent aux questions principales (qui ? quoi ? où ? quand ?) que se pose l'utilisateur en quête d'une information précise.

Les meilleurs systèmes actuels sont désormais capables de détecter et de typer les entités nommées (nom de personne, de lieu, d'organisation…) avec des taux de précision supérieurs à 90%. Le prochain saut technologique vers lequel convergent tous les travaux actuels du TALN est celui du suivi des entités dans un document donné. Considérons le texte suivant, dans lequel sont soulignées les entités nommées :

> (1) *« Nicolas Sarkozy a rencontré samedi Angela Merkel en préambule au sommet des pays les plus industrialisés (G8). Le Président de la République a une fois de plus évoqué auprès de la chancelière allemande la question du processus d'adhésion de la Turquie à l'Union Européenne. Il a réaffirmé…»*

La recherche d'information requiert la résolution des anaphores entre, par exemple, le pronom « *il* » ainsi que les termes « *Nicolas Sarkozy* » et « *Président de la République* ». Dans le cadre de cet article, nous

parlerons d'anaphore (ou de relation anaphorique) entre deux entités linguistiques lorsque l'interprétation de l'une dépend de l'autre.

L'importance de la résolution des anaphores pour les technologies langagières a conduit à l'émergence de nombreux travaux qui ont fait l'objet de campagnes d'évaluation internationales telles MUC (www.itl.nist.gov/iaui/894.02/related_projects/muc/) et SemEval (semeval2.fbk.eu/semeval2.php), ou nationales comme DEFT (deft.limsi.fr/index.php?id=1&lang=fr) au cours de la dernière décennie. Ces recherches ont toutefois porté majoritairement sur des documents ou des messages électroniques (langage écrit). A l'opposé, la communauté parole s'est surtout focalisée sur la problématique de l'anaphore pronominale, très présente en dialogue oral homme-machine (serveurs vocaux interactifs). Les avancées continues du traitement de la parole (reconnaissance vocale en particulier) amènent désormais les chercheurs à s'intéresser à une recherche d'information dans des flux oraux ou vidéos (émissions radio ou télédiffusées par exemple) équivalente à celle réalisée sur les documents électroniques. Dans cette perspective, il est essentiel de développer des techniques capables de traiter toutes les formes d'anaphores de l'oral spontané.

L'adaptation à l'oral spontané de systèmes développés initialement pour le langage écrit ne va pas de soi et nécessite une étude approfondie qui fait l'objet des travaux présentés dans cet article. La présence de multiples disfluences orales (reprises, hésitations, incises) pose en effet des problèmes scientifiques qui n'ont été qu'imparfaitement résolus à l'heure actuelle. Par ailleurs, l'oral spontané rend invisible certains accords morphosyntaxiques facilitant à l'écrit la résolution des anaphores (par exemple : « il mange » et « ils mangent » sont homophones), ce qui demande une résolution de haut niveau (sémantique et pragmatique) et non plus syntaxique (recherche d'un antécédent s'accordant en genre et en nombre). Enfin, le dialogue oral interactif facilite également l'occurrence de raccourcis métonymiques qui ont une influence directe sur la reprise anaphorique. C'est par exemple le cas de l'énoncé « *l'hôtel de la gare ils sont tous désagréables* » où il faut comprendre que le pronom « *ils* » désigne de manière métonymique les employés de l'hôtel. Il est donc important d'étudier expérimentalement sur des corpus de parole spontanée la validité d'algorithmes de résolutions conçus pour l'écrit.

La finalité de cet article est double. Il expose tout d'abord le travail d'annotation effectué dans le cadre du projet pilote CO2 interne au PRES Orléans-Tours. Celui-ci a conduit à l'annotation de relations anaphoriques dans un corpus de français parlé de référence (corpus ESLO). Par ailleurs, il présente des études conduites sur ce corpus et qui visent à évaluer la pertinence sur le français parlé spontané de certains présupposés sur lesquels reposent généralement les processus de résolution des anaphores en TAL.

En deuxième partie, nous présenterons le projet CO2 en insistant sur la méthodologie et les formats d'annotation retenus. Cette annotation pilote est un préambule à un projet de plus grande ampleur, nommé ANCOR (programme APR-IA 2011 de la région Centre) et décrit en quatrième partie, qui conduira à terme à la constitution du plus grand corpus de français parlé annoté en anaphores. La troisième partie décrit les résultats de deux études de corpus. La première évoque les descriptions définies en première mention, tandis que la seconde évalue le respect des accords en genre et en nombre dans les chaînes anaphoriques.

2 Le projet CO2

L'objectif du projet CO2 était de mener une étude systématique des différentes formes de réalisation des anaphores aussi bien pronominales que nominales. Ce projet a permis l'annotation d'un corpus de 3h30 de dialogue oral spontané en français : 35000 mots issus de trois dialogues du corpus français ESLO (corpus ESLO2 du Laboratoire Ligérien de Linguistique d'Orléans-Tours), transcrit avec *Transcriber* (Barras *et alii*, 2001). La particularité de ce corpus est son caractère conversationnel : la morphosyntaxe de la parole spontanée ainsi observée diffère de celle de l'écrit par la présence de disfluences (répétitions, reprises, hésitations...), de chevauchement dans les tours de parole, d'incises... (Blanche-Benveniste *et alii*, 1991). Les segments sont également de natures différentes (groupes de souffle/énoncés). Dans le cadre de ce projet, les annotations sont vues comme un outil pour la compréhension des phénomènes et

non comme une fin en soi et constituent un préalable incontournable à la modélisation par le TAL des processus de résolution de ces anaphores complexes.

2.1 Traits linguistiques utilisés pour le schéma d'annotation

Dans un premier temps, nous allons décrire en détail les informations linguistiques qui ont été ajoutées au corpus lors de l'annotation réalisée dans le cadre du projet CO2. Nous présentons tout d'abord quelles entités référentielles ont été considérées, puis les traits linguistiques apportés à chaque annotation.

2.1.1 Repérage des entités référentielles

Nous avons annoté l'ensemble du groupe nominal (désormais GN), déterminants et adjectifs compris. Les disfluences, venant du caractère oral du corpus, ont été traitées de la même manière que les autres entités. Les chaînes anaphoriques ont également mis en jeu des pronoms ou des groupes prépositionnels. Dans ce dernier cas, la préposition introductive n'a pas été intégrée à l'annotation, mais a été néanmoins prise en compte. Ainsi, si l'élément considéré de la chaîne anaphorique était intégré à un groupe prépositionnel, l'annotateur devait le mentionner. Nous avons, en outre, exclu le pronom *ça* et ses dérivés *cela*/*c'*/*ce* car ils reprennent très fréquemment l'ensemble d'un groupe verbal, comme c'est le cas dans l'exemple suivant où l'item « *ça* » reprend le groupe « *a encore cassé sa voiture* » :

(2) L1 : *Pierre a encore cassé sa voiture ce matin.*

L2 : *Venant de lui, ça ne m'étonne pas.*

Ces reprises correspondent à des anaphores abstraites (Dipper et Zinmeister, 2010), qui ne sont pas considérées dans notre étude. Nous avons annoté par contre les formes explétives de *il* et *ça*, comme dans *il pleut* et *ça fait longtemps*. Il est en effet important de repérer ces usages non référentiels qui peuvent tromper les systèmes de résolution des anaphores.

Enfin, dans le cas de structures coordonnées (Mazur et Dale, 2007) et des structures enchâssées, nous avons choisi d'identifier le groupe ainsi que chaque membre de la structure pouvant potentiellement donner lieu à une anaphore, comme dans l'exemple suivant :

(3) *Pierre et Marie Curie* annoté *[[Pierre] et [Marie Curie]]*

En effet, la reprise anaphorique peut aussi bien concerner l'ensemble du groupe (par exemple, « *ces deux chercheurs* ») qu'un de ses membres (comme dans « *la chercheuse d'origine polonaise* »).

2.1.2 Caractérisation des relations anaphoriques et de leurs entités

L'annotation à proprement parler a consisté à décrire par différents traits les entités référentielles et leurs éventuelles relations anaphoriques. Nous avons suivi une procédure d'annotation détaillée afin de fournir des données utiles pour évaluer la pertinence de divers dispositifs linguistiques liés à l'anaphore. On peut noter que les différents traits linguistiques, qui vont être décrits pour l'annotation, correspondent aux propriétés qui sont généralement considérées à la fois par les théories linguistiques explicatives de l'anaphore et par les systèmes de résolution des anaphores développées par le TAL (Mitkov, 2002 ; Recasens, 2010). Ces traits concernent aussi bien les entités que les relations anaphoriques par elles-mêmes. Ce sont (tableaux 1 et 2) :

- Les parties du discours :
    - P : Pronoms ;
    - N : Noms (les entités nommées, notées EN, sont des sous-types de noms, cf. infra) ;
    - NULL : réservé aux artefacts, notamment les disfluences comme les chevauchements des tours de parole. Dans l'exemple 4 ci-dessous, le groupe nominal « *la margarine* » est artificiellement partagé entre deux tours de parole du fait de la présence d'un chevauchement. Cette situation est modélisée en caractérisant deux entités (« *la* » d'une part et « *margarine* » d'autre part), qui sont ensuite reliées pour indiquer qu'il s'agit d'une

seule entité. Ainsi, pour ne pas comptabiliser deux entités, la partie qui ne contient pas la tête lexicale du groupe (le déterminant) sera typée comme artefact et ne recevra aucune caractérisation lors de l'annotation.

(4) L1 : *Oui alors je voudrais maintenant de la/*

L2 : *Oui/*

L1 : *margarine et des œufs*

- Les autres traits liés aux entités :
  - GP : l'inclusion potentielle dans un Groupe Prépositionnel. Si certaines théories suggèrent de privilégier le GN comme antécédent potentiel d'une relation anaphorique, nous avons vu précédemment qu'un GP peut également ancrer une relation anaphorique ;
  - EN : le type des entités nommées tel que défini dans la typologie utilisée au cours de la campagne d'évaluation ESTER2. Cette annotation distingue par exemple les patronymes, les toponymes et autres géonymes, les organisations politiques etc. (Galliano *et alii*, 2009). On peut en effet imaginer que le type d'entité influe sur la réalisation des anaphores. Lorsque l'item considéré n'est pas une entité nommée, cet attribut reçoit la valeur NO ;
  - DEF : la définitude est l'expression du caractère défini, indéfini, explétif ou démonstratif du déterminant et donc de l'entité.

Tableau 1 : Présentation synthétique des traits d'annotation utilisés pour caractériser les entités référentielles dans le projet CO2.

Partie du discours	N (nom) ; PR (pronom) ; NULL (artefact)
Entité nommée	PERS (humains, animaux)
	FONC (fonctions politiques, administratives, etc.)
	LOC (lieu)
	ORG (organisations de divers types)
	PROD (productions humaines : films, œuvres d'art, livres, etc.)
	TIME (heure et date)
	AMOUNT (âge, durée, poids, etc.)
	EVENT (tous types d'événements : Fête nationale, etc.)
	NO (l'item n'est pas une entité nommée selon la convention ESTER2)
Définitude	INDEF (indéfini)
	DEF_DEM (défini démonstratif)
	DEF_SPLE (défini simple)
	EXPL (formes explétives de *il* et *ça*)
Présence dans un GP	YES (présence effective d'une entité dans un GP)
	NO (l'entité n'est pas située dans un GP)
Nouvelle entité du discours	YES (nouvelle entité) ; NO (entité coréférente)

Il est à noter que le type d'entité nommée EVENT n'est pas défini dans le schéma d'annotation ESTER2, comme nous l'avions annoncé. Il a en effet été ajouté pour la campagne ETAPE à venir, qui fait précisément suite à l'évaluation ESTER2.

- Les relations anaphoriques :
    - NEW : ce trait caractérise les nouvelles entités du discours, c'est-à-dire que l'interprétation de l'entité considérée ne dépend d'aucune expression mentionnée précédemment. En elle-même, cette caractéristique concerne les entités référentielles et non les relations anaphoriques. Cependant, on notera que si toute chaîne anaphorique commence obligatoirement par une entité de type NEW, la réciproque n'est pas vraie ;
    - TYPE : type de relation anaphorique. Nous en distinguons quatre :
        - Anaphore Directe : *d* (une description définie) est en relation anaphorique avec une précédente expression nominale *a*, son antécédent ; *d* et *a* ont la même tête nominale.

            (5) *La voiture rouge… Cette belle voiture…*
        - Anaphore Indirecte : *d* est en relation anaphorique avec une précédente expression nominale *a* ; *d* et *a* ont des têtes nominales différentes.

            (6) *Le cabriolet… cette décapotable… la voiture…*
        - Anaphore Pronominale (cas particulier de l'Anaphore Indirecte).
        - Anaphore Associative : *d* n'est pas en relation anaphorique avec une précédente expression nominale *a*, mais son interprétation dépend de *a*.

            (7) *La voiture … la porte…*
    - Les traits d'accord morphosyntaxique : la plupart des théories considèrent l'accord en genre (masculin/féminin) et en nombre (singulier/pluriel) comme une contrainte forte de réalisation des relations anaphoriques. La procédure d'annotation que nous avons suivie ne décrit pas le genre et le nombre des entités annotées : nous nous sommes restreints à décrire le respect ou non de l'accord entre l'anaphore et son antécédent. Cependant, cette caractérisation en genre et en nombre des relations mais également des entités deviendra systématique dans le cadre du projet ANCOR (cf. infra 2.3).

Tableau 2 : Présentation synthétique des traits d'annotation utilisés pour caractériser les relations anaphoriques dans le projet CO2.

Types de relations	DIRECTE (anaphore directe)
	INDIRECTE (anaphore indirecte)
	ANAPHORE (anaphore pronominale)
	ASSOCIATIVE (anaphore associative)
Accord en genre	YES (accord) ; NO (absence d'accord)
Accord en genre	YES (accord) ; NO (absence d'accord)

2.1.3 Comparaison avec l'état de l'art sur l'annotation des anaphores

Le choix de ce modèle d'annotation repose sur la formalisation des relations anaphoriques par van Deemter et Kibble (2000) et sur les travaux de Vieira *et alii* (2002) pour les GN définis et démonstratifs, à la suite du travail de Poesio et Vieira (2000). Il présente à la fois des avantages et des difficultés bien identifiés.

Parmi les avantages se trouve la possibilité de comparer nos résultats sur le français oral à des travaux multilingues sur l'écrit (voir 3.2). De plus, nous avons déjà connaissance des points délicats concernant l'accord inter-annotateur. Un accord inter-annotateur assez faible pour l'écrit semble être un inconvénient qui ne peut que présager des difficultés pour l'annotation de l'oral. Dans le cadre d'un projet faisant suite à CO2 (le projet ANCOR présenté en dernière partie du présent article), il nous semble donc nécessaire de tester sur notre corpus cet accord inter-annotateur suivant les différentes propositions citées sur un extrait de notre corpus. Il nous sera alors possible de voir quel est le gain obtenu globalement.

Plusieurs travaux ont cherché à pallier ce désaccord inter-annotateur en proposant des schémas d'annotations plus précis. On citera notamment Gardent et Manuélian (2003) qui ont développé un schéma d'annotation des relations anaphoriques (bridging) selon une typologie sémantique précise afin d'augmenter l'accord inter-annotateur et les travaux de Recasens *et alii* (2011) introduisant la notion de quasi-identité pour les cas qu'ils décrivent comme de la quasi-coréférence et qui sont pris en compte dans notre schéma sous l'étiquette 'anaphore associative'.

Toutefois, ces propositions alternatives n'envisagent que les cas de désaccord liés à l'hésitation entre les anaphores associatives et les autres types d'anaphores, mais rien n'est dit des désaccords entre 'nouvelle entité du discours' et 'anaphores associatives' qui dépendent du sentiment qu'ont les annotateurs de réaliser une inférence et qui, comme l'ont montré Vieira *et alii* (2002), représentent un pourcentage important. Ainsi, pour reprendre un exemple de Vieira *et alii* (2002), un annotateur a codé un lien anaphorique entre *a* et *d*, tandis que l'autre a codé *a* comme 'nouvelle entité du discours' :

(8) a. *L'important flux de <u>réfugiés albanais</u> en Italie…*

d. *La Commission s'efforce de venir en aide <u>aux populations victimes de catastrophes</u>…*

Au regard de tout cela, il nous semble donc que le schéma d'annotation du projet CO2 présente comme avantage majeur la possibilité de comparaison de données avec des corpus écrits déjà annotés avec les mêmes catégories, tout en proposant une description déjà fine des relations anaphoriques.

2.2 Choix du logiciel pour l'annotation

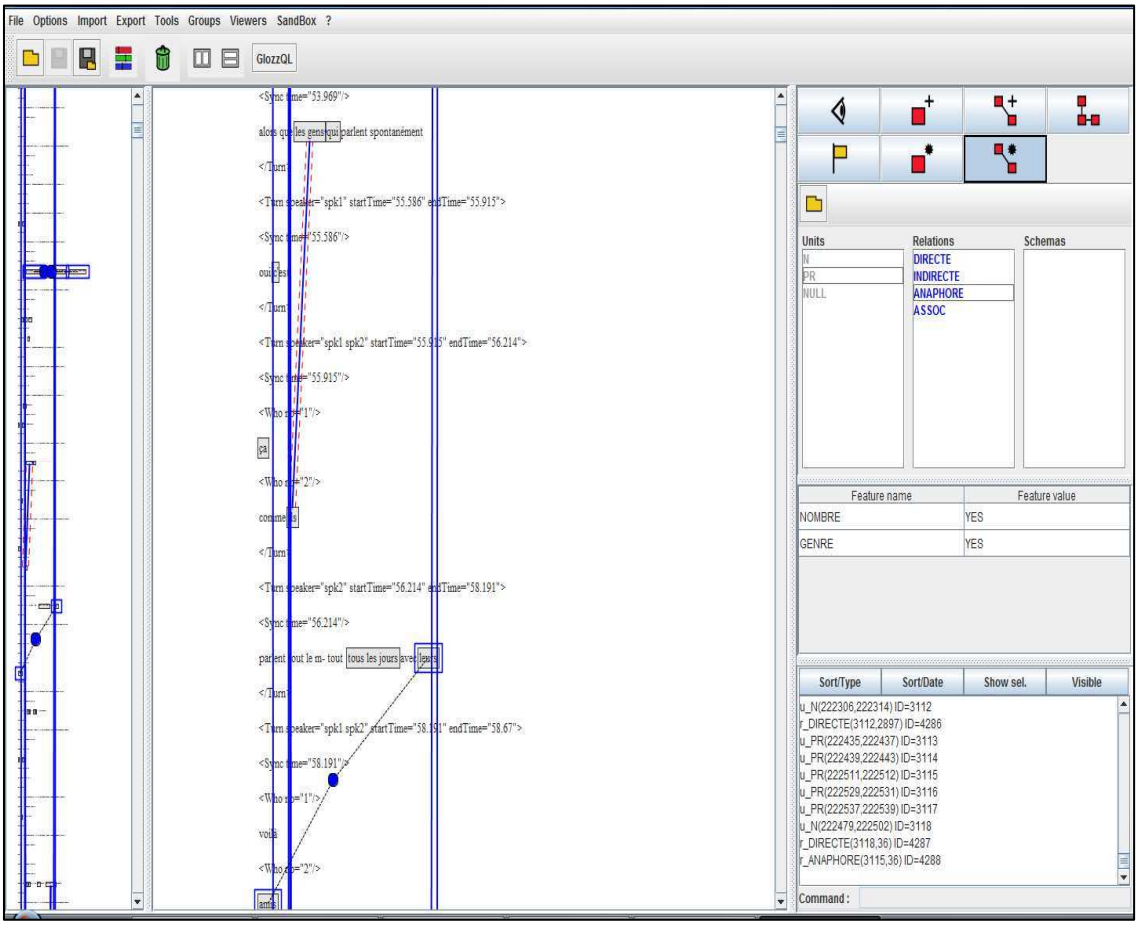
Au préalable à la constitution du corpus annoté CO2, nous avons mené une étude comparative de différents logiciels d'annotation au regard de nos besoins, qui ne sont pas uniquement l'annotation des entités référentielles, mais également les relations qu'elles sont susceptibles de partager entre elles. Trois logiciels ont particulièrement retenu notre attention. Le logiciel *Cadixe* (Bessieres *et alii*, 2001), s'il permet d'annoter les entités, ne peut en revanche annoter les relations anaphoriques. La plateforme *MMAX2* (Müller et Strube, 2006) possède une interface peu conviviale rendant le travail d'annotation difficile. Elle n'est par ailleurs plus maintenue. Nous avons finalement fait le choix du logiciel *Glozz* (Mathet et Widlöcher, 2009) qui nous permet à la fois d'annoter les entités et leurs relations.

De façon plus précise, l'un de nos principaux besoins et critères dans le choix de l'outil était que celui-ci devait être flexible en s'adaptant à nos besoins, d'autant plus que CO2 relevait d'un projet pilote où de nombreuses évolutions étaient à prévoir. *Glozz* produit ainsi une annotation au format XML reposant sur une DTD entièrement personnalisable que nous avons pu faire évoluer en même temps que nos recherches. Autre flexibilité permise par le logiciel, les annotations sont proposées dans des dossiers séparés au format XML, synchronisés avec le discours transcrit (annotation déportée, plus connue sous le terme anglais *stand-off annotation*). Cette annotation déportée ouvre la porte à la réalisation d'annotations multi-niveaux très intéressantes.

Glozz propose, en outre, un outil de recherche dans le corpus (GlozzQL) et la définition de schémas permettant de rechercher des motifs récurrents dans le corpus. C'est cet outil qui nous a permis d'extraire les résultats que nous présentons en troisième partie.

Enfin, la nouvelle version de *Glozz* comprend un calcul d'accord inter-annotateur intégré ainsi qu'une fonction de visualisation des relations anaphoriques sous la forme d'un graphe affiché dans une fenêtre dynamique (Mathet et Widlöcher, 2011). Cette dernière fonctionnalité est particulièrement séduisante pour la révision des annotations.

Figure 1 : Illustration des relations anaphoriques sous *Glozz* (traits bleus).



2.3 Procédure d'annotation

Nous avons modifié la DTD de l'outil *Glozz* afin de l'adapter à notre schéma d'annotation. Le corpus CO2 a fait l'objet d'un codage par deux annotateurs et selon une procédure en quatre phases successives :

1) Repérage et caractérisation manuel des Entités Nommées, et, plus largement, des éléments principaux d'une chaîne anaphorique (pronoms et Groupes Nominaux) par un annotateur. Les entités nommées ont été automatiquement identifiées en utilisant la ressource *CasEN* (tln.li.univ-tours.fr/Tln_CasEN.html) avec le logiciel *CasSys*, disponible sur la plateforme *Unitext* (Friburger et Maurel, 2004 ; Maurel *et alii*, 2011), puis corrigées (uniquement pour le corpus ESLO) par un expert humain selon la convention ESTER2 (Galliano *et alii*, 2005). Les autres GN, dont les pronoms, ont été identifiés semi-automatiquement avant la tâche d'annotation des anaphores ;
2) Révision croisée du repérage par l'autre annotateur et recherche de consensus entre annotateurs ;
3) Repérage et caractérisation des relations anaphoriques par un annotateur ;
4) Révision croisée des relations ainsi caractérisées par l'autre annotateur.

La phase (3) consiste à relier respectivement l'anaphore à l'antécédent. Certains travaux privilégient une annotation en chaînes anaphoriques (Tardiff, 2005 ; Gardent et Manuélian, 2005 ; Amsili *et alii*, 2007). A la suite de Chastain (1975) et de Corblin (2005), nous « appelons chaîne anaphorique une séquence d'expressions singulières apparaissant dans un contexte telles que si l'une de ces expressions réfère à quelque chose, toutes les autres y réfèrent également ».

Dans le projet CO2, il a été décidé de relier au contraire l'anaphore à la première mention de l'entité référentielle trouvée dans le texte. Des arguments d'ordre linguistique ou computationnel peuvent être trouvés en faveur de chacune de ces représentations qui ont donc chacune leurs avantages et leurs inconvénients. C'est pourquoi nous avons pris garde, autant que possible, à ce que notre codage en première mention puisse être transformé automatiquement en une annotation en chaîne de coréférence. C'est la raison pour laquelle, l'accord en genre et nombre ne sera pas codé au niveau de la relation dans le projet ANCOR qui fera suite à CO2 en 2012. Ce seront les entités qui porteront une marque de genre et de nombre, qui permettra ensuite la vérification automatique de cet accord, quelle que soit la représentation choisie. Le type de relation anaphorique (directe, indirecte, pronominale, associative) dépend toutefois directement du choix d'annotation effectuée, sans qu'une solution alternative ne nous paraisse envisageable.

Le schéma d'annotation du projet CO2 ne comprend pas d'accord inter-annotateur ; celui-ci sera en revanche pleinement intégré au schéma mis en œuvre lors du projet ANCOR. La raison tient au fait que la révision des annotations était effectuée par les annotateurs eux-mêmes, l'annotation était donc déjà relativement lissée.

3 Résultats

Le corpus CO2 a été conçu comme corpus pilote pour évaluer la pertinence de plusieurs contraintes habituellement prises en considération par les processus de résolution des anaphores en TAL. La richesse des annotations sur corpus permet de tester ce qui devrait concerner potentiellement une gamme étendue de traits linguistiques. Pour le moment, deux expériences ont été conduites sur le corpus. Avant de les présenter, nous allons décrire le matériel de test que constitue le corpus.

3.1 Matériel de test

Dans sa version actuelle, le corpus CO2 correspond à 208 minutes d'enregistrement de la parole et 35192 mots. Le corpus d'essai qui en résulte inclut 8910 entités nominales et pronominales et 3513 relations anaphoriques. Nous allons rapidement décrire certaines caractéristiques des observables du corpus.

Tableau 3 : Distribution des entités nominales et pronominales dans le corpus CO2.

	Nouvel élément du discours	Élément déjà introduit dans le discours	Total
Entités nominales	2 542 (99,6%)	1 804 (28,9%)	4 346 (49,4%)
Entités pronominales	11 (0,4%)	4 441 (71,1%)	4 452 (50,6%)

Le Tableau 3 montre la distribution des entités nominales et pronominales dans le corpus CO2. Premièrement, nous notons que ces entités nominales et pronominales apparaissent presque dans la même mesure (respectivement 49,4% contre 50,6%). Bien qu'une majorité d'entités nominales présente un nouvel élément du discours (2542 nouvelles entités du discours parmi 4346 entités nominales), les éléments nominaux représentent toujours 28,9% des éléments en relation anaphorique. Ceci prouve que les anaphores nominales doivent être prises en compte par les processus de résolution des anaphores en TAL, alors que la plupart des travaux sur les systèmes de dialogue oral se concentrent seulement sur l'anaphore pronominale. Comme attendu, les pronoms n'introduisent quasiment jamais de nouvel élément du discours (0,4% des situations).

Tableau 4 : Distribution des références dans la chaîne anaphorique dans le corpus CO2

	Nouvel élément du discours	Référence de l'élément
Entités nominales	550 (99,8%)	1 616 (55,0%)
Entités pronominales	1 (0,2%)	1 323 (45,0%)

Puisqu'il y a plus de reprises pronominales que de reprises nominales, nous nous attendions à ce que les entités nominales apparaissent plus fréquemment à l'initiale des chaînes anaphoriques. Le Tableau 4 confirme partiellement cette hypothèse. Même si nous n'avons trouvé qu'une seule chaîne anaphorique commençant par un pronom, les pronoms peuvent fréquemment agir en tant que référence à l'intérieur d'une chaîne anaphorique : ils représentent 45% des références dans ces positions. En conséquence, il est illusoire de vouloir distinguer la résolution des anaphores pronominales de celle des anaphores nominales en français parlé. De même, nous avons noté que l'antécédent d'un terme anaphorique est situé dans un groupe prépositionnel dans 27% des relations anaphoriques. Bien souvent, les processus de résolution privilégient la recherche de l'antécédent dans les groupes nominaux. Quoiqu'appropriée, cette heuristique présente un risque dans 27% des cas et ne doit être considérée que comme une préférence.

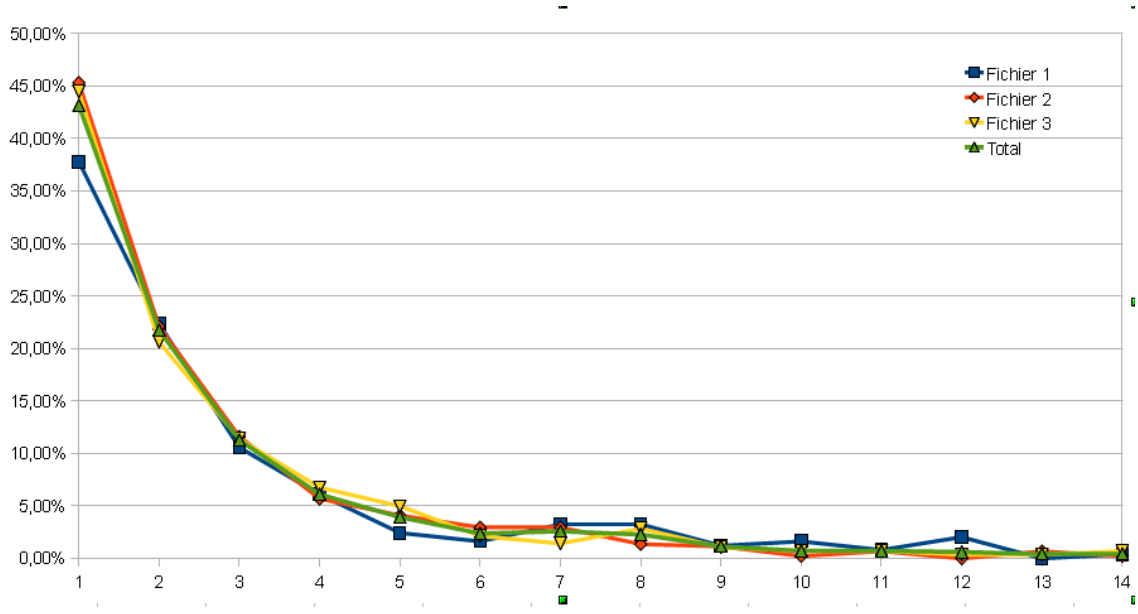
Tableau 5 : Distribution des types de relation anaphorique dans le corpus CO2.

Directe	Indirecte	Pronominale	Anaphore associative
34,2% (σ = 6,8%)	15,1% (σ = 3,5%)	37,4% (σ = 4,0%)	13,4% (σ = 7,7%)

Le Tableau 5 présente la distribution des relations anaphoriques selon les types structuraux décrits dans la section 2.1.2. L'anaphore directe, qui est aisément traitée par les processus de résolution, représente 34,2% de ces relations. L'anaphore pronominale, qui a capté l'attention des chercheurs en TAL depuis des années, représente un tiers de ces relations (37,4%). La résolution des anaphores associatives demeure un défi en TAL. Malheureusement, ces anaphores représentent 13,4% des anaphores attestées dans le corpus CO2, ce qui signifie que leur traitement ne peut être ignoré sans conséquence. La plupart de ces anaphores complexes correspondent à des cas de métonymie.

Nous nous sommes également intéressés à la longueur des chaînes anaphoriques. Cette information est importante pour la résolution des co-référence, puisque la distance entre les termes coréférents est une caractéristique essentielle pour les algorithmes de résolution.

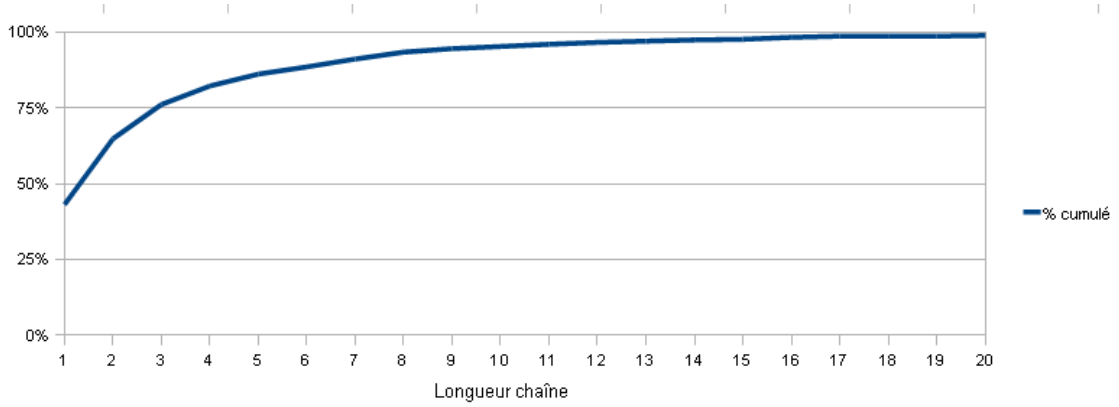
Figure 2 : Distribution des chaînes anaphoriques en fonction de leur longueur



Sur l'ensemble du corpus, les chaînes anaphoriques ont une longueur moyenne de 6,14 entités. Cette moyenne masque toutefois de fortes disparités, puisque la plus longue chaîne observée comportait 44 entités nommées, tandis que les chaînes de longueur 1 (i.e. comportant uniquement une relation anaphorique entre deux entités) représentent près de 45% des cas, comme le montre la figure 2. On peut observer sur cette figure que la distribution des chaînes suivant leur longueur est remarquablement stable entre les différents dialogue du corpus. On observe que celle-ci suit approximativement une loi de Zipf qui explique que si la longueur moyenne des chaînes est supérieure à 6, la valeur médiane est inférieure à deux. La figure 3 permet de prendre la mesure de cette distribution : on compte déjà 65% des chaînes qui ont une longueur inférieure à 2, et 76% une longueur inférieure à 3). Les 90% sont atteints avec des chaînes de longueur inférieure à 7.

En conclusion, si les chaînes anaphoriques peuvent potentiellement s'étendre sur de très longues durées, la plupart d'entre elles mobilisent un nombre assez restreint d'éléments anaphoriques.

Figure 3 : Distribution cumulée des chaînes anaphoriques en fonction de leur longueur : pourcentage de chaînes de longueur inférieur à une valeur donnée



3.2 Les Descriptions Définies comme nouvelles entités de discours

Les descriptions définies sont des GN dont le déterminant est le/la/les. Ce caractère défini (comme trait du Groupe Nominal, GN) est un trait qui est largement considéré par les processus de résolution (Soon *et alii*, 2001). Il a déjà été mentionné dans divers travaux (Poesio et Vieira, 2000 ; Gundel *et alii*, 2001 ; Lyons, 1999 ; Racasens *et alii*, 2009) que les GN définis introduisent de nouvelles entités dans le discours. La quantité de Descriptions Définies (DD) utilisées en tant que Nouvelles Entités de Discours (NED) en texte écrit a déjà été étudiée. Vieira *et alii* (2002) observe que 49,6 % des DD sont classifiée comme NED dans leur corpus (la version française du Journal officiel de l'UE). Une quantité semblable est trouvée dans des textes en portugais et en portugais brésilien.

Dans une première expérience, nous avons voulu évaluer le pourcentage de DD employée comme NED dans ce corpus français pour le comparer aux résultats mentionnés par Vieira *et alii* (2002). Sans surprise (traditionnellement, les nouvelles entités ont une distribution différente en ce qui concerne le genre de discours, voir Biber *et alii*, 1998), les résultats prouvent que le taux de DD classifiée comme NED dans notre corpus est très supérieur, soit 69,8%. Cela rejoint les taux observés par Recasens (2009) sur des textes journalistiques en espagnol, qui sont de 73% de définis à l'initiale de chaîne anaphorique. Analysant 4900 descriptions définies dans le corpus journalistique du *Monde*, (Gardent et Manuelian, 2005) observent quant à elles une proportion de 56% de DD classifiées comme NED. 9% des DD sont par ailleurs jugées non référentielles. Au final, 45% des descriptions définies qu'elles étudient sont donc intégrées dans une chaîne anaphorique, dont près de la moitié correspondent à des anaphores directes.

3.3 Accord en genre et en nombre

L'accord en genre et en nombre est une contrainte très commune, toujours considérée par les processus de résolution de l'anaphore. Elle exprime une contrainte obligatoire pour les correcteurs symboliques (Lappin et Leas, 1994), alors qu'elle joue le plus souvent un rôle de préférence pour les approches heuristiques (Mitkov, 1998). Le genre et le nombre sont, enfin, des propriétés toujours considérées par les techniques basées sur un apprentissage automatique sur des corpus annotés (Recasens, 2010).

Tandis que ces deux contraintes ont prouvé leur utilité sur la langue écrite, très peu de travaux les ont examinées sur la parole conversationnelle. Cependant, la présence de disfluences (reprises, corrections, incises…) et l'usage fréquent de métonymies pouvant prêter lieu à des anaphores associatives dans le dialogue oral spontané mériteraient que l'on y porte attention.

Nous avons ainsi entrepris plusieurs études distributionnelles sur le corpus CO2 afin d'avoir une image précise de l'accord en genre et en nombre dans les chaînes anaphoriques en français parlé conversationnel.

Tableau 6 : Accord en genre dans les relations anaphoriques dans le corpus CO2.

Directe	Indirecte	Pronominales	Associatives	Total
99,0% (σ = 1,0%)	74,5% (σ = 9,3%)	98,7% (σ = 0,8%)	70,1% (σ = 9,1%)	91,3% (σ = 4,9%)

Le Tableau 6 présente les résultats de l'accord en genre. Dans l'ensemble, nous pouvons considérer que l'accord en genre est bien respecté en français parlé conversationnel : 91,3% des relations anaphoriques respectent cette contrainte. Surtout, le taux d'accord s'élève jusqu'à 99% dans le cas des anaphores directes et pronominales. Ce taux d'accord diminue de manière significative (74,5%) dans le cas des anaphores indirectes. Cela était prévisible puisque le genre est relativement arbitraire en français : même si deux têtes lexicales décrivent le même référent, elles peuvent ne pas présenter le même genre. Par exemple, « voiture » est un mot féminin, alors que son hyperonyme « véhicule » est masculin.

De même, l'accord en genre pour les anaphores associatives est assez faible (70,1%). Cette absence d'accord était attendu, puisque dans ce cas, il n'y aucune identité de référence entre l'antécédent et l'entité anaphorique. C'est donc plutôt le taux d'accord encore important qui est étonnant dans nos observations. Nous nous sommes demandés si ce résultat n'était pas dû à la prédominance d'un genre sur l'autre dans nos corpus. Une étude distributionnelle nous indique une légère prévalence des termes de genre masculin (60,72%) sur le genre féminin. Cette prévalence est remarquablement stable sur les trois dialogues étudiés (Tableau 7). Elle ne saurait toutefois pas expliquer le taux d'accord de 70% étudié. En effet, un calcul statistique montre qu'une distribution au hasard du genre de l'antécédent et de l'anaphore associative ne conduirait alors qu'à un accord de 52,3%. Il semble donc que l'accord en genre joue encore partiellement son rôle même dans le cas des anaphores associatives.

Tableau 7 : Répartition de termes suivant le genre dans le corpus CO2.

Genre	Dialogue 1	Dialogue 2	Dialogue 3	Total
Masculin	62,5%	60,5%	60,5%	62,7% (σ = 1,7%)
Féminin	37,5%	39,5%	39,5%	32,3% (σ = 1,7%)

En conclusion, cette étude sur le corpus CO2 suggère que le français parlé conversationnel obéit aux mêmes contraintes que le français écrit en ce qui concerne l'accord en genre. Ce dernier peut être considéré utilement par les processus de résolution concernant les anaphores directes et pronominales. Cependant, les contraintes de genre et de nombre ne sont pas pertinentes pour les anaphores indirectes et associatives où elles peuvent tout de même être considérées comme heuristiques faibles.

L'accord en nombre amène à d'autres conclusions. En effet, les résultats présentés dans le Tableau 8 montrent que l'accord en nombre est sensiblement moins respecté que l'accord en genre en français parlé conversationnel.

Tableau 8 : Accord en nombre dans les relations anaphoriques dans le corpus CO2.

Directe	Indirecte	Pronominales	Associatives	Total
88,3% (σ = 2,8%)	85,8% (σ = 3,9%)	90,7% (σ = 5,3%)	21,9% (σ = 11,8%)	85,3% (σ = 4,0%)

Dans l'ensemble, l'accord en nombre est seulement respecté dans 85,3% des relations anaphoriques attestées. Ce résultat recoupe les observations de (Antoine, 2004) qui ne concernaient que l'anaphore pronominale. Étonnamment, cet accord modéré vaut pour chaque type de relations. En particulier, un nombre remarquable d'anaphores directes ne présente pas d'accord en nombre (taux d'accord : 88,3%). L'étude détaillée du corpus montre que, dans la plupart des situations où l'accord est absent, le référent est générique. Dans de tels cas, le pluriel ou le singulier peut être employé indifféremment en français, comme le montre l'exemple suivant :

(9) *« Sur le plan des honoraires, les malades me payent leur consultation et ils sont remboursés à 75%. (...) je n'ai pas le droit de les dépasser, sauf lorsque le malade pose des exigences ou s'il s'agit d'une urgence ».*

De telles situations peuvent également se produire avec l'anaphore indirecte et pronominale. Par exemple, l'expression référentielle « *le malade* » de l'exemple précédent peut être remplacée sans difficulté par l'expression anaphorique indirecte « *le patient* » (tête lexicale différente sans accord en nombre) ou par le pronom singulier « *il* ». Dans tous les cas, il y aura alors absence d'accord avec l'antécédent « *les malades* ». Ceci explique le faible taux d'accord que nous avons également noté avec l'anaphore indirecte et pronominale. Notons que le taux d'accord que nous observons avec les anaphores pronominales est plus élevé que celui relevé par (Barbu et al., 2002) en anglais écrit. Certains cas de désaccord observés par ces auteurs sont spécifiques à l'anglais, comme l'usage du pluriel *they* pour remplacer le traditionnel *he or she* générique dans le monde anglo-saxon. Ces cas particuliers mis à part, (Barbu et al., 2002) observent comme nous que l'utilisation de pluriels génériques ou de noms collectifs est une des principale causes de non respect de l'accord en nombre.

Enfin, l'accord en nombre chute à 21,9% dans le cas des anaphores associatives. La présence de la métonymie est l'explication principale de ce manque d'accord, comme le montre l'exemple suivant :

(10) *« A l'hôtel Caumartin généralement ils sont tous désagréables ».*

Nous avons également entrepris quelques analyses de données complémentaires pour évaluer si d'autres caractéristiques linguistiques pouvaient influer sur l'accord en nombre. Comme illustré par le Tableau 9, aucun facteur d'influence n'a pu être réellement caractérisé. Dans tous les cas (référence située dans un groupe prépositionnel, référence correspondant à une entité nommée, définitude de la référence), le taux d'accord reste situé entre 80% et 90%. L'accord en nombre tend à être légèrement inférieur avec la référence indéfinie. Ceci peut être expliqué par le fait qu'il doit correspondre plus fréquemment à une référence générique. Cependant, un test statistique montre que la dispersion des données est trop haute (écart type σ = 6,1%) pour caractériser cette diminution comme significative.

Tableau 9 : Accord en nombre selon le type de la référence.

Référence dans un GP	Référence dans une EN	Référence définie	Référence démonstrative	Référence indéfinie
84,8% (σ = 6,4%)	85,2% (σ = 2,7%)	87,8% (σ = 3,8%)	90,3% (σ = 7,0%)	80,4% (σ = 6,1%)

Pour conclure, cette étude a clairement prouvé que l'accord en nombre est très modérément respecté quel que soit le type de relations anaphoriques considérées et le type de la référence. Le taux d'accord moyen de 85% concernant les relations anaphoriques montre en effet qu'il serait risqué que les processus de résolution de l'anaphore les considèrent comme obligatoires en français parlé conversationnel. L'accord en nombre peut au mieux être considéré comme une préférence à la réalisation des anaphores, mais en aucun cas comme une contrainte. Nous avons vu que lorsque les cas de désaccords concernent particulièrement les situations où le référent est générique. D'un point de vue computationnel, cette

observation demande aux concepteurs de systèmes de résolutions des anaphores de considérer le trait d'accord conjointement à celui du type (générique ou spécifique) des entités. Cette modélisation est aisée avec les rares systèmes à bases de règles encore utilisés (Haghighi & Klein 2009). Dans le cas des systèmes centrés données reposant sur un apprentissage sur corpus annoté, on remarquera que des propositions ont été faites pour apprendre des modèles différents suivant le type d'entité considéré (Ng 2005). Ces résultats sont, de notre point de vue, de bonnes indications de l'intérêt d'une étude en corpus annoté des relations anaphoriques en français parlé conversationnel.

4. Travaux en cours : le Projet ANCOR (ANaphore dans les Corpus ORaux)

Depuis décembre 2011, cet effort mené pour une annotation minutieuse se poursuit au travers du projet sur deux ans nommé ANCOR (subvention APR-IA Région Centre). Le corpus, composé de fichiers ESLO, sera renforcé par le corpus PAROLE_PUBLIQUE (Antoine, 2002), fortement interactif. Il aboutira ainsi à un corpus de discours spontané annoté comprenant à terme un million de mots et au moins 50 000 relations anaphoriques. Il sera le plus volumineux corpus de français parlé annoté en anaphores. Ce corpus sera librement distribué et devrait être utile pour toute recherche sur la résolution des anaphores en discours spontané. Il nous permettra, en particulier, de poursuivre l'évaluation expérimentale des traits linguistiques mise en application par les processus de résolution des anaphores complexes en TAL en discours spontané.

Parmi ces autres dispositifs linguistiques, nous envisageons de prendre en compte dans le schéma d'annotation d'autres paramètres tels que le parallélisme syntaxique (Mitkov, 1994), le caractère spécifique ou générique du déterminant (Recasens *et alii*, 2010), le marquage des structures topicalisées, etc. Comme nous employons un format d'annotation déporté (fichier d'annotation indépendant synchronisé avec le fichier de transcription), il sera possible d'envisager une annotation multi-niveaux combinant, par des fichiers d'annotation séparés mais synchronisés, différents types d'informations utiles à l'analyse des procédés anaphoriques.

A ce jour, le corpus produit est de 5841 mots pour 75 fichiers annotés. Ces fichiers sont extraits du sous-corpus OTG (Office de Tourisme de Grenoble) issu du corpus Parole_Publique. Fortement interactives, ces situations discursives ont pour point commun d'être des demandes de renseignements, émanant aussi bien des habitants de la ville que des touristes de passage. Ils seront augmentés du sous-corpus UBS (Université de Bretagne-Sud), également très interactif, qui est composé d'enregistrements du standard téléphonique d'une université. Cela nous permettra de comparer les observations faites sur ESLO avec un corpus plus dynamique en termes d'oral spontané. Ces deux corpus Parole_Publique, comprenant au total 356 fichiers, seront dans un second temps renforcés par 46 fichiers du corpus ESLO, dont trois ont donc fait l'objet du travail entrepris dans le cadre du projet CO2.

Enfin, la dernière version de la plate-forme d'annotation *Glozz* intégrant un outil d'accord d'inter-annotateur (Mather et Widlöcher, 2009), nous serons en mesure de calculer le taux d'accord inter-annotateur sur différents segments de notre corpus et d'évaluer la fiabilité de l'annotation qui sera diffusée.

Références bibliographiques

Amsili, P., Landragin, F., Acosta, A., Bittar, A. (2007). *Résolution anaphorique : Etat d'une réflexion collective*, 1–4.

Antoine, J.-Y., Nicolas, P., Letellier-Zarshenas, S., Schadle, I., Caelen, J. (2002). Towards a large corpus of spoken dialogue in French that will be freely available : the PAROLE PUBLIQUE project. In: *Proceedings of the 3rd International Conference on Language Resources & Evaluation, LREC'2002*, 649–655.

Antoine, J.-Y. (2004). Résolution des anaphores pronominales : Quelques postulats du TAL mis à l'épreuve du dialogue oral finalisé. In: *Actes TAL2004.*

Barbu, C., Evans, R., Mitkov, R. (2002) A corpus based investigation of morphological disagreement in anaphoric relations. In: *Proceedings of LREC'2002*, volume 6, 275–280.

Barras, C., Geoffrois, E., Wu, Z., Liberman, M. (2001). Transcriber: Development and use of a tool for assisting speech corpora production. *Speech Communication 33(1-2)*, 5–22.

Bessieres, P., Nazarenko, A., Nedellec, C. (2001). Apport de l'apprentissage à l'extraction d'information : Le problème de l'identification d'interactions géniques. In: *Actes du 4e Colloque International sur le Document Electronique*, 1–11.

Biber, D., Conrad, S., Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge University Press.

Blanche-Benveniste, C., Bilger, M., Rouget, C., van den Eynde, K. (1991). *Le français parlé : Etudes grammaticales*. Paris : CNRS Editions.

Chastain, C. (1975) Reference and Context. In Gunderson, K. (éd.), *Language Mind and Knowledge*, Minneapolis : Presses universitaires du Minnesota, 194–269.

Corblin, F. (2005) Les chaînes de la conversation et les autres. In Gouvard, J.-M. (éd.), *De la langue au style*, Lyon : Presses universitaires de Lyon, 233–254.

Dipper, S., Zinmeister, H. (2010). Towards a standard for annotating abstract anaphora. In: *Proceedings of the LREC 2010 workshop on Language Resources and Language Technology Standards*, 54–59.

Friburger, N., Maurel, D. (2004). Finite-state transducer cascade to extract named entities in texts. *Theoretical Computer Science 313*, 94–104.

Galliano, S., Geoffrois, E., Mostefa, D., Choukri, K., Bonastre, J.F., Gravier, G. (2005). The ESTER phase II evaluation campaign for the rich transcription of French broadcast news. In: *9th European Conference on Speech Communication and Technology-2005*, 1149–1152.

Galliano, S., Gravier, G., Chaubard, L. (2009). The ESTER2 evaluation campaign for the rich transcription of french radio broadcasts. In: *Interspeech'09*, 2583–2586.

Gardent, C., Manuélian, H., Kow, E. (2003). Which bridges for bridging definite descriptions ? In: *Proceedings of the EACL 2003 Workshop on Linguistically Interpreted Corpora*, 69–76.

Gardent, C., Manuélian, H. (2005). Création d'un corpus annoté pour le traitement des descriptions définies. *TAL*, 46(1), 115–139.

Gundel, J., Hedberg, N., Zacharski, R. (2001). Definite descriptions and cognitive status in English: Why accommodation is unnecessary. *English Language and Linguistics 5(2)*, 273–295.

Haghighi, A., Klein, D. (2009). Simple coreference resolution with rich syntactic and semantic features. In: *Proceedings of EMNLP 2009*, 1152–1161.

Landragin, F. (2004). *Dialogue homme-machine multimodal. Modélisation cognitive de la référence aux objets*. Paris : Hermès-Lavoisier.

Lappin, S., Leas, H. (1994). An algorithm for pronominal anaphora resolution. *Computational Linguistics 20(4)*, 535–561.

Lyons, C. (1999). *Definiteness*. Cambridge University Press.

Mathet, Y., Widlöcher, A. (2009). La plate-forme GLOZZ : Environnement d'annotation et d'exploration de corpus. In: *Actes de TALN-2009*, 1–10.

Mathet, Y., Widlöcher, A. (2011). Une approche holiste et unifiée de l'alignement et de la mesure d'accord inter-annotateurs. In: *Actes TALN-2011*, 1–12.

Maurel, D., Friburger, N., Antoine, J.Y., Eshkol-Taravella, I., Nouvel, D. (2011). Cascades autour de la reconnaissance des entités nommées. A paraître *TAL 52(1)*.

Mazur P., Dale R. (2007) Handling conjunctions in named entities. *Lingvisticae Investigationes*, 30(1), 49–68.

Mitkov, R. (1994). *An integrated model for anaphora resolution*. In: *Proceedings of the 15th Conference on Computational Linguistics*, 1170–1176.

Mitkov, R. (1998). Robust pronoun resolution with limited knowledge. *Association for Computational Linguistics 98*, 869–875.

Mitkov, R. (2002). *Anaphora Resolution*. Longman.

Müller, C., Strube, M. (2006). Multi-level annotation of linguistic data with MMAX2. In: Braun, S., Kohn, K., Mukherjee, J., editors, *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*. Peter Lang, Francfort, Allemagne, 197–214.

Ng, V. (2005). Machine learning for coreference resolution: From local classification to global ranking. In: *Proceedings of ACL 2005*, 157–164.

Poesio, M., Vieira, R. (2000). An Empirically Based System for Processing Definite Descriptions. *Computational Linguistics 26(4)*, 525–579.

Recasens, M., Martí, M.A., Taule, M. (2009). First mention definites: More than exceptional cases. *The Fruits of Empirical Linguistics: Products2*:217.

Recasens, M. (2009). A chain-starting classifier of definite NPs in Spanish. In: *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, 46–53.

Recasens, M. (2010) *Coreference: Theory, Annotation, Resolution and Evaluation*. Mémoire de doctorat de l'Université de Barcelone, Espagne.

Recasens, M., Hovy, E., Martí, M.A. (2011). Identity, non-identity, and near-identity: Addressing the complexity of coreference. *Lingua 121(6)*, 1138–1152.

Soon, W., Ng, H., Lim, D. (2001). A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics 27(4)*, 521–544.

Tardif, O. (2005). Annotation de la coréférence entre expressions référentielles. *Texte et Corpus : Actes des quatrièmes Journées de la Linguistique de Corpus*, 105–110.

van Deemter, K., Kibble, R. (2000). On Coreferring: Coreference in MUC and related annotation schemes. *Computational Linguistics 26(4)*, 629–637.

Vieira, R., Salmon-Alt, S., Schang, E. (2002). Multilingual corpora annotation for processing definite descriptions. *Advances in Natural Language Processing*, 721–729.