SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801359

Le corpus ANCOR_Centre et son outil de requêtage : application à l'étude de l'accord en genre et nombre dans les coréférences et anaphores en français parlé

Anaïs Lefeuvre (1), Jean-Yves Antoine (1), Emmanuel Schang (2)

Affiliation : 1 Université François Rabelais de Tours, LI, E.A. 6300
2 Univ.Orléans, CNRS, LLL, UMR 7270
{anais.lefeuvre et jean-yves.antoine}@univ-tours.fr, emmanuel.schang@univ-orleans.fr

1 Introduction

Cet article rend compte d'un effort important d'annotation des anaphores et des relations de coréférence sur des corpus de français parlé spontané. La question de la coréférence est importante en linguistique comme en TALN. Elle constitue un indice fort de cohérence thématique dans un document, et elle est essentielle à la compréhension des textes et à la recherche d'information. Aussi n'est-il pas étonnant que de nombreux linguistes aient cherché à développer des modèles de la coréférence ou de l'anaphore. On dispose ainsi en linguistique de nombreux modèles explicatifs de la cohérence discursive qui rendent compte des reprises coréférentielles ou anaphoriques : théories du centrage (Grosz et al, 1995), de l'accessibilité (Ariel 1990, 2001), de l'optimalité (Prince et Smolensky, 1993). De son côté, l'importance de la résolution des coréférences et anaphores a été reconnue par le TALN dès les travaux pionniers de Hobbs (1978). Aux travaux initiaux basés sur des méthodes symboliques (Lappin et Leas, 1994) ont succédé des approches plus heuristiques (Mitkov, 1994), avant que les approches par apprentissage sur données (Soon et al., 2001, Ng et Cardie, 2002, Haghighi & Klein 2009) ne deviennent prédominantes.

En dépit de ces multiples recherches, coréférence et anaphore suscitent encore de nombreuses interrogations. D'une part, la validité de la plupart des modèles développés par la théorie linguistique prête à questionnement. Par exemple, de nombreux auteurs ont étudié les limites ou la validité de la théorie du centrage (Cornish, 1999, Kleiber 2002). Ces études se sont toutefois basées le plus souvent sur de courts textes illustratifs à la représentativité limitée. La validation expérimentale de ces modèles sur de grandes masses de données n'a pas vraiment été abordée par la linguistique de corpus (on citera cependant Poesio et al. 2004, Chiarcos 2009). De leur côté, les méthodes d'apprentissage développées par le TAL ont recours à de grands corpus annotés en coréférence. L'apprentissage repose toutefois sur des traits linguistiques dont la pertinence n'a pas toujours été démontrée. Nous verrons ainsi dans cet article que les attributs de genre et nombre associés aux mentions potentiellement coréférentielles reposent sur des hypothèses d'accord qui ne sont pas systématiquement vérifiées.

Il nous apparaît ainsi important de disposer de données d'observations suffisamment représentatives pour mener des études quantitatives de corpus utiles aussi bien à la linguistique qu'au TAL. Malheureusement, il n'existe pas en français de corpus d'envergure annoté en coréférence. Le tableau 1 présente la liste des principaux corpus annotés en coréférence disponibles au niveau mondial : le français est complètement absent de ce panorama. A notre connaissance, le seul corpus disponible en français est DEDE, centré sur l'étude des descriptions définies. Il ne comporte malheureusement que 48 kMots (Gardent et Manuelian, 2005), ce qui limite sa représentativité et le rend inutilisable pour les besoins de l'apprentissage automatique. De même, le corpus du CRISTAL, de grande envergure, ne peut qu'être partiellement utilisé car il ne code que certaines formes particulières d'anaphore[1] (Tuttin et al., 2000).

Le corpus ANCOR_Centre (ANCOR par la suite) vise à répondre à cette absence de ressource utilisable sur le français. En se concentrant exclusivement sur la modalité orale, ANCOR représente par sa taille (488 000 mots) ce qui est à notre connaissance le plus grand corpus de parole spontanée annoté en coréférence. Il s'accompagne d'une annotation linguistique riche afin de satisfaire aussi bien aux besoins du TAL qu'à ceux des sciences du langage. Afin de favoriser son appropriation par la communauté scientifique, le corpus ANCOR est distribué librement sous licence Creative Commons et s'accompagne

Article available at http://www.shs-conferences.org or http://dx.doi.org/10.1051/shsconf/20140801359

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801359

d'un outil de requêtage qui permet l'interrogation de la ressource pour des personnes ne disposant pas de connaissances informatiques.

Tableau 1 – Principaux corpus annotés en coréférence

Langue	Corpus	Genre	Taille (mots)
Allemand	TüBa-D/Z *(Hinrichs et al., 2005)*	Informations (News)	800 000
Anglais	OntoNotes *(Pradhan et al., 2007)*	News, dialogue oral, conversation téléphonique, weblogs, flux radio ou télédiffusés	50 000
Chinois	OntoNotes *(Pradhan et al., 2007)*		400 000
Catalan	AnCora-*Ca (Recasens & Marti, 2010)*	Informations	400 000
Espagnol	Ancora-Es (*Recasens, 2010*)	Informations	400 000
Japonais	NAIST Text *(Idia et al., 2007)*	Informations	970 000
Hollandais	COREA *(Heindrickx et al., 2008)*	Informations, parole, encyclopédie	325 000
Tchèque	PDT *(Nedouluzhko et al., 2009)*	Journaux d'information	800 000
Polonais	PCC (*Ogrodniczuk et al., 2013*)	Nombreux genres oraux et écrits	514 000

Cet article a pour objectif de décrire la ressource et son outil de requêtage, puis de présenter une étude de corpus portant sur la question de l'accord en genre et nombre lors de la reprise coréférentielle. Cette étude questionnera directement certaines hypothèses acceptées sur le langage écrit mais jamais étudiées sur l'oral, tout en fournissant une première illustration des capacités d'analyse qu'offrent le corpus et son outil d'interrogation.

2 Présentation du corpus ANCOR

2.1 Contenu : corpus audio sources

Le corpus ANCOR ne concerne que la modalité orale. Sans constituer une ressource équilibrée comme le corpus PCC polonais, il a pour ambition de représenter une réelle diversité de situations discursives orales. Il regroupe ainsi l'annotation de quatre corpus de parole spontanée transcrits sous *Transcriber* (Barras et al., 2001). Ces corpus sont présentés dans le tableau 2. Deux d'entre eux ont été extraits du corpus ESLO, qui regroupe des entretiens sociolinguistiques présentant un degré d'interactivité faible (Baude et Dugua 2011, Eshkol-Taravella et al. 2012). A l'opposé, les deux autres corpus, OTG et Accueil_UBS (Nicolas et al., 2002), concernent des dialogues homme-homme interactifs. Ces deux derniers corpus diffèrent par le média utilisé : le corpus OTG regroupe des conversations de visu au sein d'un office de tourisme pour OTG, tandis qu'Accueil_UBS a été enregistré dans un standard téléphonique. Au total, le corpus regroupe 488 000 mots et correspond à une durée d'enregistrement de 30,5 heures.

Tableau 2 – Contenu du corpus ANCOR : corpus audio sources

Corpus	Situation discursive	Finalisation[2]	Interactivité	Taille & Durée
ESLO_ANCOR	Interview	Modérée	Faible	417 kMots – 25h
ESLO_CO2	Interview	Modérée	Faible	35 kMots – 2,5 h
OTG	Dialogue oral	Très forte	Forte	26 kMots – 2h
Accueil_UBS	Dialogue téléphonique	Assez forte	Forte	10 kMots – 1 h

2.2 Méthodologie d'annotation

L'annotation a été réalisée sur le logiciel *GLOZZ* (Mathet et Widlöcher, 2009). Dans sa version actuelle (1.0), le corpus est distribué sous format *GLOZZ*. A terme (été 2014), il sera également proposé sous format *MMAX2* (Müller et Strube, 2006), et sera normalisé suivant les recommandations de la TEI (Text

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801359

Encoding Initiative). Les annotations réalisées sous *GLOZZ* sont séparées du corpus source avec lequel elles sont synchronisées. Une telle annotation déportée permet un enrichissement multi-niveaux du corpus, ce qui est intéressant en termes d'évolutivité. Afin de limiter la charge cognitive des experts et pour favoriser la cohérence intra-annotateurs, le processus d'annotation a été divisé en quatre étapes successives :

1. Caractérisation des mentions (annotateurs : étudiants de Master ou de doctorat en linguistique)
2. Vérification de la phase 1 par un superviseur
3. Caractérisation des relations de coréférence ou anaphoriques (annotateurs identiques)
4. Vérification de la phase 3 par un superviseur.

2.3 Schéma d'annotation

Le schéma d'annotation du corpus ANCOR cherche de manière classique à identifier pour chaque entité référentielle (ou mention) si elle introduit une nouvelle entité du discours, si elle réfère à une entité précédemment mentionnée (coréférence) ou si la référence à une entité précédemment mentionnée dans le texte est nécessaire pour son interprétation (anaphore associative). Ce paragraphe décrit la philosophie générale qui a présidé à la mise en place du schéma d'annotation retenu.

Mentions : repérage des entités référentielles – Il est important de noter que l'annotation se limite strictement aux entités nominales ou pronominales. Un groupe nominal tel que *le lendemain* sera ainsi annoté comme mention intéressant l'annotation, alors que l'adverbe *demain* ne le sera pas. Ce choix fort peut induire l'omission de certaines coréférences, particulièrement dans le cas de la référence temporelle. Il a été décidé afin de s'assurer d'une fiabilité maximale des données. Notre expérience a en effet montré que les codeurs éprouvent des difficultés à savoir ce qui doit être ou non considéré comme une mention si on ne se limite pas à une définition purement syntaxique (noms et pronoms) des mentions. ANCOR propose ainsi une annotation fiable définie sur des critères précis. Le format ouvert d'annotation permet à ceux qui le désirent de compléter cette annotation à l'aide de phénomènes non décrits actuellement.

L'annotation prend en compte l'ensemble du groupe nominal et pas uniquement sa tête. Elle concerne également les pronoms et les groupes prépositionnels (GP). Dans ce dernier cas, la préposition introductive n'est pas intégrée à l'annotation mais prise en compte par un attribut associé (GP=YES). Ont été en outre exclus le pronom *ça* et ses dérivés lorsqu'il reprend l'ensemble d'un groupe verbal, comme dans l'exemple : *Pierre a encore cassé sa voiture. Venant de lui, ça ne m'étonne pas.* Ces reprises correspondent à des anaphores abstraites[3], qui dépassent largement les objectifs d'annotation du corpus. Nous avons par contre annoté les formes explétives de *il* (cf. *il pleut*). Il est en effet important de repérer ces usages non référentiels qui peuvent tromper les systèmes de résolution. Enfin, dans le cas de structures coordonnées ou enchâssées, nous avons choisi d'identifier le groupe ainsi que chaque membre le composant. Tous ces éléments peuvent en effet ancrer une reprise coréférentielle.

Anaphore ou coréférence : délimitation des relations – La délimitation des relations consiste à relier les éléments coréférentiels ou anaphoriques. Certains travaux privilégient une annotation en chaînes (Gardent et Manuélian, 2005 ; Amsili et al. 2007) c'est-à-dire en séquences d'expressions référent au même élément du discours. Pour le corpus ANCOR, il a été décidé au contraire de relier toutes les relations à la première mention de l'entité référentielle trouvée dans le texte. Nous avons en effet estimé que l'annotation en chaîne posait des problèmes délicats dans le cas de dialogues interactifs : la notion de chaîne, pertinente dans la linéarité de l'écrit, devient à l'oral beaucoup moins évidente à caractériser pour les annotateurs. Des arguments d'ordre linguistique ou computationnel peuvent toutefois être trouvés en faveur de chaque représentation. C'est pourquoi les évolutions futures du corpus ANCOR offriront dès l'été 2014, outre le codage en première mention actuel :

- un codage en chaîne, sous forme de séquences de mentions coréférentes au long du texte,
- un codage en clusters de co-référents, qui se contente d'associer le même identificateur aux mentions qui réfèrent à une même entité du discours. Ce codage n'est compatible qu'avec le format MMAX2.

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801359

Caractérisation des relations et de leurs entités – Un des objectifs du projet ANCOR est de proposer une annotation fine permettant une interrogation du corpus suivant un grand nombre de propriétés linguistiques. Les mentions sont caractérisées par un **TYPE** qui correspond aux parties du discours de ces entités : P (pronom), N (Nom) ou NULL (artefact lié à certaines disfluences orales). Elles sont plus finement décrites par les propriétés linguistiques suivantes :

- G : Genre et N : Nombre
- GP : inclusion dans un GP – Valeur YES (si l'entité est un GP) ou NO (si c'est un GN)
- EN : types d'entités nommées – Les types retenus sont ceux dans la campagne d'évaluation ESTER2 (Galliano et al., 2009). On utilise le type NO si l'entité n'est pas une entité nommée.
- DEF : définitude – cet attribut sert à distinguer le caractère défini (DEF), indéfini (INDEF), démonstratif (DEM) ou explétif (EXP) de l'entité.
- GENE : généricité – Permet de décrire si l'entité considérée dénote un référent générique ou spécifique (*j'ai le pain* vs. *le pain que tu m'as offert était délicieux*)
- NEW : attribut binaire qui précise si la mention constitue ou non une nouvelle entité du discours.

De leur côté, les relations sont caractérisées par un **TYPE** qui distingue classiquement les classes :

- ***DIR : coréférence directe***, dans le cas d'une coréférence entre mentions de même tête nominale (exemple : *le bus rouge… ce grand bus*),
- ***IND : coréférence indirecte***, si les entités coréférentes ont des têtes nominales différentes (exemple : *le cabriolet… cette décapotable*),
  - ***PR : coréférence pronominale***, dans le cas particulier de la coréférence indirecte où la reprise est un pronom, (exemple : *le cabriolet… il roulait*). Notons que nous sommes bien ici dans un cas de coréférence, même si la littérature parle parfois improprement d'anaphore pronominale,
- ***ASSOC : anaphore associative*** (*bridging anaphora* en anglais) si les mentions ne sont pas coréférentes mais que l'interprétation de l'une dépend de l'autre (exemple : *le village … son clocher*),
- ***ASSOC_PR : anaphore associative pronominale***, dans le cas où la reprise associative est portée par un pronom comme dans le cas de métonymies : *le café Jeanne d'Arc, ils sont tous désagréables*.

On notera ainsi que le type code à la fois le statut référentiel de la reprise (coréférence ou anaphore) et sa forme morpho-syntaxique.

Cette classification nécessite quelques explications, qui correspondent aux consignes données aux annotateurs. Le typage des relations repose en large mesure sur la classification proposée dans Vieira & alii (2002). Elle repose sur la distinction faite par Van Deemter & Kibble (2000) entre la coréférence et l'anaphore : deux SN *a* et *b* coréfèrent ssi Référent(a)=Référent(b). Le lien anaphorique est d'une autre nature : il n'y a pas coréférence mais l'interprétation d'une expression *b* dépend d'une expression *a*. Dans la classification ci-dessus, DIR, PR et IND correspondent à une relation de coréférence, tandis que ASSOC et ASSOC-PR correspondent à une relation anaphorique. Le classement proposé intègre donc des critères de forme (pour des raisons pratiques d'exploitation des données) à l'intérieur de critères sémantico-référentiels qui dominent et qui guident les annotateurs.

Par ailleurs, les traits linguistiques suivants sont associés à chaque relation :

- GENRE et NOMBRE, valeurs booléennes qui indiquent si la relation respecte ou non une contrainte d'accord en genre et nombre en l'antécédent et sa reprise coréférentielle ou anaphorique,
- ID_LOC : cet attribut précise si la reprise est le fait du locuteur qui a fait la première mention au référent (antécédent), ou s'il s'agit au contraire d'un autre interlocuteur.

2.4 Estimation de la fiabilité des données

La fiabilité du corpus a été estimée par une expérience qui a consisté à mesurer l'accord entre 4 experts ayant participé à l'annotation sur un sondage de 10 fichiers. L'estimation de l'accord inter-annotateur

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801359

reste une question ouverte dans le cas de la coréférence, du fait de problèmes d'alignement entre annotations (Passoneau, 2004 ; Artstein et Poesio, 2008 ; Matthet et Widlöcher, 2011). Nous contournons ce problème par la mesure de l'accord sur la délimitation des relations, puis sur celui du typage de ces relations. Trois métriques ont été utilisées : κ (Cohen, 1960), π (Scott, 1955) et α (Krippendorff, 2004).

Tableau 3 – Mesures de fiabilité sur le corpus ANCOR

Tâche	Kappa	Pi	Alpha
Délimitation : accord inter-annotateur	0.45	0.45	0.45
Délimitation : accord intra-annotateur	0.91	0.91	0.91
Typage : accord inter-annotateur	0.80	0.80	0.80

On observe dans le tableau 3 un excellent accord inter-annotateur (0,80 sur toutes les métriques) sur la tâche de typage des relations. A l'opposé, l'accord est bien plus faible sur la tâche de délimitation (0,45 sur toutes les métriques). Cette fiabilité est en dessous du seuil de confiance de 0,64. Il faut toutefois comprendre que cette mesure est pénalisée par notre codage en première mention : une divergence sur la première mention entraîne un désaccord sur tous les coréférents qui la suivent. Cette difficulté a déjà été repérée par Vieira et al. (2002). Dès que le corpus ANCOR sera codé en chaînes, il nous sera possible d'estimer l'accord inter-annotateur de manière moins biaisée. Dans l'immédiat, nous avons procédé à une expérimentation avec la superviseuse principale du corpus, en lui demandant de reproduire l'annotation d'un extrait du corpus. Nous avons comparé son annotation avec celle acceptée dans le cadre de la révision du corpus. Les mesures d'accord, que nous qualifierons cette fois d'intra-annotateur, que nous obtenons (0,91) nous montrent que l'annotation est très fortement cohérente.

2.5 Distribution du corpus

Le corpus ANCOR est diffusé directement par téléchargement sur une des pages WWW suivantes :

- http://tln.li.univ-tours.fr/Tln_Corpus_Ancor.html (portail projet ANCOR)
- www.info.univ-tours.fr/~antoine/parole_publique/ (serveur de corpus Parole_Publique)

Il est distribué sous licence *Creative Commons* sous une licence reproduisant celle des corpus sources, à savoir CC-BY-SA-NC pour les corpus ELSO et CO2, et CC-BY-SA pour les OTG et Accueil_UBS.

3 Outil de requêtage du corpus

Les corpus annotés présentent des ressources précieuses à la description des phénomènes linguistiques propres à une langue. Néanmoins, la quantité des données disponibles dans un corpus annoté d'envergure requiert des dispositifs facilitant l'accès à celles-ci. Nous proposons à la communauté des sciences du langage ANCORQI (ANCOR Query Interface), un outil permettant d'exploiter facilement toute grande masse de données codées suivant le format Glozz, et donc en particulier le corpus ANCOR. Cet outil permet d'explorer un corpus tant en termes quantitatifs que qualitatifs, proposant un calcul sur la distribution des objets annotés répondant à un certain nombre de contraintes données par l'utilisateur, puis en lui proposant de visualiser en contexte les objets concernés sous la forme d'un concordancier.

3.1 GlozzQL et ANCORQI

La plateforme d'annotation Glozz comporte déjà un outil de requêtage utilisant un langage dédié, GlozzQL. Son objectif est directement lié aux tâches successives d'annotation et d'observation (Mathet 2011). Ainsi, l'annotateur, travaillant sur un fichier, a l'opportunité d'annoter puis d'observer le fruit de son annotation en utilisant une requête. Ce dispositif lui permet d'affiner son annotation ou d'en observer la cohérence. ANCORQI n'est pas redondant avec cet outil, dans le sens où GlozzQL ne permet d'interroger qu'un seul fichier à la fois. ANCORQI permet au contraire une exploitation de l'ensemble d'un corpus, de manière préférentielle dans une phase d'exploitation des données, c'est-à-dire une fois l'annotation terminée. C'est un outil de requêtage simple d'emploi qui permet d'exprimer graphiquement

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801359

des contraintes sur toutes les propriétés utilisées par l'annotation. Cette simplicité d'interrogation restreint toutefois la complexité des requêtes en comparaison à GlozzQL.

3.2 Applications d'ANCORQI

Notre outil est dépendant du format Glozz mais pas d'un schéma d'annotation spécifique respectant ce format. En premier lieu, l'utilisateur doit donc choisir le fichier de DTD qui décrit le schéma d'annotation retenu pour le corpus interrogé, puis le répertoire et/ou les fichiers que l'on souhaite analyser. Nous détaillons maintenant les différentes fonctionnalités d'ANCORQI.

Description générale du corpus – Lors du chargement du corpus, une première analyse est menée indépendamment d'une quelconque requête exprimée par l'utilisateur. Elle donne la distribution des principales catégories d'objets qui peuplent le corpus. Cette vue d'ensemble permet d'acquérir rapidement une vision générale du contenu du corpus interrogé (figure 1). Par exemple, la description représentée sur la figure 1 donne le nombre de Noms et de Pronoms, puis la répartition des relations par type, avant de décrire la distribution des mentions par genre etc. Ce descriptif exploite exhaustivement les propriétés d'annotation sans les combiner : les requêtes vont permettre une fouille plus précise du corpus.

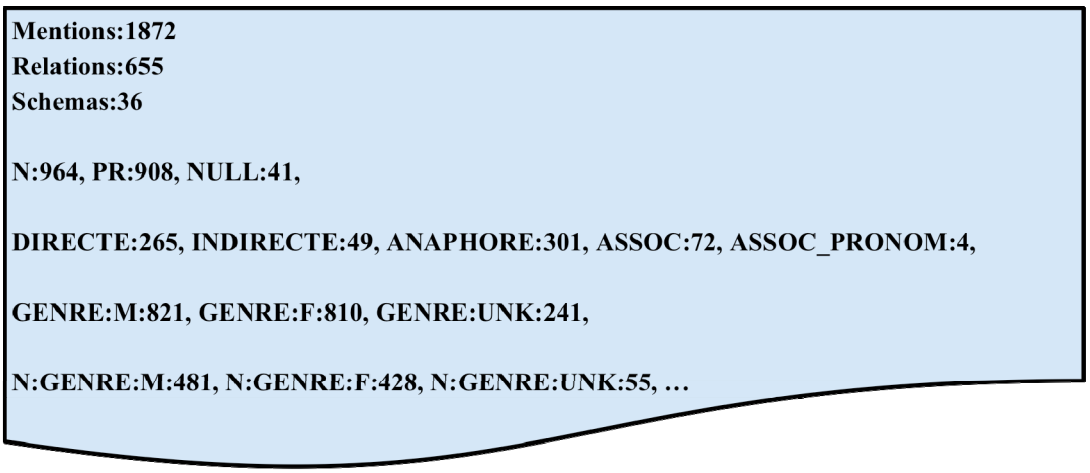

Figure 1 – Exemple extrait d'une description générale de corpus

Requêtes simples – Une fois le descriptif du corpus obtenu, l'utilisateur se voit proposer le panel des contraintes disponibles à la requête, c'est-à-dire l'ensemble exhaustif des types et traits spécifiés dans la DTD (schéma d'annotation). Deux niveaux de requêtes sont offerts : les requêtes simples et les requêtes complexes. Les premières permettent d'obtenir le décompte des objets qui répondent à la combinaison de contraintes sur les traits les caractérisant. Par exemple, il est possible d'obtenir l'ensemble des relations qui observent un accord en genre et en nombre (conjonction de traits). En revanche, il est impossible d'obtenir le nombre de mentions qui sont des entités nommées référant soit à des personnes, soit à des organisations. Pour cette requête mobilisant deux valeurs distinctes pour le même trait (disjonction de traits), il faudra formuler deux requêtes, une première pour les personnes, et une seconde pour les organisations. Cette limitation, due au choix d'une définition graphique des requêtes, ne restreint donc pas les capacités d'interrogation du corpus, tout en simplifiant leur expression. De la même manière, la négation n'est pas prise en charge par notre outil mais peut-être réalisée indirectement.

Requêtes complexes – Un module de requêtes complexes est implémenté afin de joindre aux contraintes sur les relations, des contraintes sur les mentions qu'elles associent. On peut par exemple rechercher les relations observant l'accord en nombre tout en ayant pour antécédent une mention définie et/ou pour coréférent une mention ayant le trait indéfini. La combinaison des contraintes sur les relations et sur les mentions qui les composent ouvre les perspectives d'exploitation immédiates du corpus. Après avoir

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801359

exprimé cette requête, on peut ainsi faire varier le critère de l'accord de la relation et de la définitude de l'antécédent pour contraster l'influence de la définitude des premières mentions sur l'accord.

Relation	Entité 1	Entité 2
PR	**Jeanne** aime bien son	bien son jardin. **Elle** y passe tous
PR	J'ai **un client** pour toi qui a un problème	Et bien passe-le-moi, je vais voir ça avec **lui**
…	…	…

Figure 2 – Illustration schématique de l'affichage du concordancier

Concordancier – Une fois les requêtes exécutées, il est proposé à l'utilisateur de visualiser un concordancier affichant ligne par ligne et fichier par fichier les mentions extraites par la requête, ou les paires de mentions dans le cas d'une requête portant sur les relations. Ces objets sont présentés avec leur contexte proche comme le montre la figure illustrative 2. La taille du contexte, définie en nombre de caractères, est paramétrable.

3.3 Dépendance entre le format Glozz et ANCORQI

Comme nous l'avons dit plus haut, notre outil peut être utilisé sur tout corpus respectant le format Glozz. Plus précisément, ANCORQI permet de formuler des requêtes sur les annotations qui observent le méta-modèle URS (pour Unité-Relation-Schéma) issu de (Widlöcher, 2008) et qui sert de base à Glozz. Afin de contrôler la cohérence de l'annotation, Glozz s'appuie sur une DTD, fichier XML unique, qui décrit la structure du schéma d'annotation en question. ANCORQI s'adapte au contenu de cette DTD, qui varie d'un schéma d'annotation à l'autre. Concrètement, la DTD comporte les trois objets URS qui ont tous un type et des traits spécifiques à la tâche d'annotation (voir § 2 .3 l'exemple notre schéma d'annotation) : ce sont ces éléments qui se retrouveront dynamiquement dans l'interface de requêtage. Une autre caractéristique du format Glozz est qu'il repose sur une notation déportée. Ceci se traduit dans ANCORQI par l'accès au seul fichier comportant l'annotation afin d'obtenir en premier lieu le résultat de la requête et la position des éléments répondant aux contraintes formulées. L'accès au fichier contenant les données sources n'est utile que lors de l'appel au concordancier. Cette séparation donne l'opportunité d'obtenir rapidement des données quantitatives et de prendre le temps de construire le concordancier lorsque le phénomène requiert une réelle immersion qualitative dans le corpus.

3.4 État des lieux et perspectives

En l'état actuel des choses, le moteur de requête d'ANCORQI est opérationnel mais n'est pas encore couplé à une interface graphique : ANCORQI est exécuté en ligne de commande. Nous développons actuellement une interface conviviale permettant de garder à l'écran le descriptif général du corpus afin de guider l'utilisateur dans le choix de la sélections des contraintes sur les différents objets (relations et/ou mentions). L'interface intégrera la visualisation du concordancier, ainsi qu'un utilitaire de sauvegarde des résultats de la requête aux formats `.txt` et `.csv` (ce dernier étant importable directement dans un tableur).

4 Contenu du corpus ANCOR : premières analyses distributionnelles

Ce paragraphe présente les premiers résultats réalisés avec ANCORQI afin d'illustrer la richesse du corpus ANCOR. Le tableau 4 nous donne le décompte des objets qui peuplent le corpus. Intégrant 115 672 mentions et 51 494 relations, le corpus ANCOR offre la possibilité de conduire des analyses représentatives même sur des aspects assez rares de la coréférence. Si la taille des sous-corpus n'est pas équilibrée du fait de la rareté des corpus oraux très interactifs en français, *OTG* et *Accueil_UBS* recensent tout de même un nombre d'observations sans commune mesure avec l'existant. On observe par ailleurs

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801359

que la proportion des nouvelles mentions (environ 1/3 des mentions) et le ratio mentions/relations restent stables sur tous les sous-corpus. Cette stabilité suggère que la coréférence est un processus autant guidé par les nécessités de la programmation discursive que par des considérations pragmatiques, puisque le degré d'interactivité n'a pas d'influence marquée sur ces ratios.

Tableau 4 : Décompte des entités présentes dans ANCOR et ses sous-corpus

Corpus	ESLO	CO2	OTG	UBS	TOTAL ANCOR
Mentions (tous types confondus)	97939	8399	7462	1872	**115672**
dont nouvelles mentions (NEW)	*26,8%*	*32,2%*	*38,4%*	*33,7%*	***28,0%***
dont mentions coréférentes	*73,2%*	*67,8%*	*61,6%*	*66,3%*	***72,2%***
Relations (tous types confondus)	44597	3670	2572	655	**51494**
Ratio Mentions / Relations	*2,19*	*2,29*	*2,90*	*2,86*	***2,25***

4.1 Mentions présentes dans le corpus

Les résultats donnés dans le tableau 5, qui concernent la distribution des mentions référentielles, dénotent également une stabilité assez remarquable entre des sous-corpus qui représentent pourtant des genres oraux différents (interview vs. dialogue oral finalisé). On constate tout d'abord qu'entités nominales et pronominales s'équilibrent toujours fortement. Il semble que la reprise pronominale réponde là encore avant tout à une logique de programmation discursive. Cette observation est à rapprocher des travaux de (Kenny & Huyck, 2011), qui suggèrent que l'usage des pronoms est plus lié à la saillance discursive qu'à la saillance situationnelle. Dans un autre registre, le système de la langue doit être convoqué pour expliquer la quasi-stabilité de la proportion de mentions inclues dans un groupe prépositionnel (GP). Une telle observation, qui porte sur la langue générale, serait sans doute différente en langue de spécialité.

Tableau 5 : Étude distributionnelle sur les mentions suivant différentes caractéristiques

Corpus	ESLO	CO2	OTG	UBS	TOTAL ANCOR
Entités nominales	48,4%	51,7 %	52,5 %	51,5 %	**48,9%**
Entités pronominales	51,6%	48,3 %	47,5 %	48,5 %	**51,1%**
% de mentions dans un GP	28,0%	29,9 %	27,8 %	25,7 %	**28,1%**
Genre mentions : % masculin	52,8%	56,7%	50,5%	49,9%	**52,9%**
Genre mentions : % féminin	43,9%	40,2%	39,3%	44,4%	**43,3%**
Genre mentions : % inconnus	3,2%	3,2%	10,2%	5,7%	**3,7%**
Nombre mentions : % singulier	65,0%	68,1%	66,0%	83,0%	**65,6%**
Nombre mentions : % pluriel	31,8%	28,8%	24,3%	14,2%	**30,8%**
Nombre mentions : % inconnus	3,2%	3,2%	9,6%	2,8%	**3,6%**
% d'indéfinis	25,1%	27,3 %	17,2 %	11,9 %	**24,5%**
% de définis simples	65,9%	66,0 %	74,0 %	80,3 %	**66,7%**
% de définis démonstratifs	6,9%	5,2 %	6,2%	6,5 %	**6,7%**
% d'explétifs	2,0%	1,5 %	2,6 %	1,3 %	**2,0%**

L'analyse du genre des mentions révèle là encore une forte stabilité, avec une prédominance sensible du genre masculin. L'existence de mentions de genre inconnu est due à la présence d'entités nommées de type toponyme ou de sociétés, pour lesquelles la notion de genre n'est pas opérante :

(1) *La Gacilly est un village charmant du Morbihan*
(2) *Le Grand Lemps est une petite ville qui peine a maintenir une activité touristique*

Le fort taux de mentions de genre inconnu dans le corpus OTG est précisément lié à la prédominance des toponymes dans un corpus recueilli en office de tourisme. Il est difficile de comparer les études portant

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801359

sur le genre. Sjöblom (2002) observe une prédominance des substantifs masculins (56,5%) sur les féminins dans l'œuvre de Le Clezio, sans que cette prédominance soit statistiquement significative. A l'opposé, Brunet (1981) observe une prédominance du féminin dans l'œuvre de Hugo (53,8 %). Une recherche sur le corpus Frantext donne enfin une prédominance de féminin (56%) selon Sjöblom.

La situation est plus tranchée du côté du nombre, où le singulier prédomine de manière significative. Là encore, les mentions de nombre inconnu relèvent le plus souvent des toponymes pour qui la notion ne fait généralement pas sens. On observe un taux encore plus fort de mentions au singulier dans le corpus *Accueil_UBS*, observation qu'une analyse qualitative plus fine devra expliquer.

Le corpus ANCOR distingue deux types de définis syntaxiques : les définis simples (introduits par l'article défini) et les définis démonstratifs. La définitude « sémantique » — critères de familiarité ou d'identifiabilité (Lyons 1999) —a de son côté été codée par les traits générique/spécifique. Les cas où les SN définis réfèrent à un type sont donc codés par le trait « générique ». Le dénombrement des génériques par type syntaxique est donc réalisable et fait l'objet d'une étude en cours. Dans l'immédiat, le tableau 5 donne la répartition des mentions en fonction de la définitude syntaxique. On observe là encore des régularités assez notables entre les différents sous-corpus, ce qui traduit l'absence d'influence sensible du degré d'interactivité sur ce facteur : les définis simples représentent ainsi toujours une très forte majorité des mentions. Moins nombreux, les indéfinis restent toutefois très fréquents dans tous les corpus. A l'opposé, définis démonstratifs et explétifs représentent dans cet ordre des catégories très marginales.

Sans pouvoir entrer dans les détails ici, nous pouvons cependant faire quelques observations concernant l'usage des définis dans le corpus ANCOR. Alors qu'il est communément admis que les articles définis en français servent à introduire un nom déjà identifié (entité déjà mentionnée dans le discours), facilement identifiable (emplois situationnels) ou bien des types (catégories générales d'êtres ou de choses), les travaux en TAL ont montré qu'il n'est pas aisé de décider si une description définie a ou non un antécédent (cf. Manuélian (2003:25)). Les travaux de Poesio & Vieira (1998) ont montré que la définition des catégories d'annotation a un impact sur l'accord inter-annotateur et sur la reconnaissance d'une description définie comme coréférentielle ou non. Recasens (2009) s'interroge sur les définis qui démarrent une chaîne référentielle et indique que les définis en début de chaîne (chain-starting) représentent plus de 50% des cas. Pour l'espagnol, Recasens trouve 73% de définis en initiale de chaîne, pour le français, Vieira & alii (2002) obtiennent 49,6% des définis classés en nouvelle entité du discours. Une requête rapide avec ANCORQI nous permet d'observer que les SN définis introduisent une nouvelle entité du discours dans 53,2 % des cas. Ces observations convergent avec les études citées précédemment.

Tableau 6 : Distribution des entités nommées par type sur l'ensemble du corpus ANCOR

PERS	LOC	ORG	AMOUNT	TIME	PROD	PROD	PROD
28856 (72,3%)	4121 (10,3%)	1832 (4,6%)	1649 (4,1%)	1465 (3,7%)	1334 (3,4%)	438 (1,1%)	201 (0,5%)

Enfin, le tableau 6 donne la répartition des entités nommées en fonction de leur type. On observe comme attendu une prédominance des personnes (PERS) et des géonymes (LOC). Par son envergure, ANCOR regroupe plus de mille entités nommées pour la plupart des types. Il s'agit donc d'une ressource potentiellement utile à des travaux spécifiques à la problématique des entités nommées.

4.2 Relations présentes dans le corpus

L'étude des relations conduit également à l'observation de régularités fortes entre corpus : elles sont autant d'indicateurs d'une représentativité de la ressource, qui va certainement au-delà des seules genres d'oral représentés. La distribution des relations par type (tableau 7) est de ce point de vue éclairante : en dépit de situations discursives différentes, on retrouve toujours la même répartition de procédés, avec une prédominance des coréférences pronominales et directes (environ 80% des observations). Cette hiérarchisation est assez naturelle : en ne mobilisant pas de processus d'accès lexical complexe, la reprise pronominale et la coréférence directe sont les plus aisées à mobiliser d'un point de vue cognitif. La réalisation des coréférences semble donc là encore répondre à une logique discursive plus que

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801359

situationnelle, ce qui explique sans doute la stabilité de nos observations. Une étude menée sur de l'écrit ou de l'oral préparé ne donnerait ainsi peut-être pas les mêmes résultats, ce que nous ne pouvons vérifier en l'absence de corpus de référence sur ces genres discursifs. Par ailleurs il semble intéressant de noter que le degré d'interactivité du discours n'est pas statistiquement signifiant dans les variations d'utilisation des diverses relations[4].

Tableau 7 : Étude distributionnelle sur les mentions suivant différentes caractéristiques

Corpus	ESLO	CO2	OTG	UBS	TOTAL ANCOR
Direct	41,1%	35,2 %	39,7 %	40,5 %	**39,9%**
Indirect	7,3%	11,2 %	6,1 %	7,5 %	**7,3%**
Pronominale	43,9%	38,2 %	46,4 %	46,0 %	**41,6%**
Associative	10,4%	14,4 %	13,5 %	11,0 %	**10,2%**
Assoc. Pronominale	0,9%	1,0 %	3,3 %	0,6 %	**1,0%**

Enfin, le tableau 8 présente la répartition des premières mentions des chaînes de coréférence. Sans surprise, l'écrasante majorité des relations anaphoriques ou de coréférence s'ancrent sur une entité nominale : les cataphores, introduites par un pronom, sont très minoritaires.

Tableau 8 : Répartition des relations en fonction de la catégorie de leur premier référent (ancre).

Corpus	ESLO	CO2	OTG	UBS	TOTAL ANCOR
Entité nominale	97,5%	97,7 %	96,9 %	97,8 %	**97,4%**
Pronom (cataphore)	2,5%	2,3 %	3,1 %	2,2 %	**2,6%**

5 Etude expérimentale sur l'accord en genre et en nombre

Ce paragraphe présente la première étude expérimentale à grande échelle menée sur l'intégralité du corpus ANCOR à l'aide de l'outil ANCORQI. Elle porte sur le respect de l'accord en genre et en nombre dans la coréférence et l'anaphore. Il a déjà fait l'objet d'études antérieures de notre part (Antoine, 2004 ; Muzerelle et al., 2012) qui n'avaient ni la profondeur ni la représentativité de la présente étude. L'accord en genre et en nombre est une contrainte très commune, toujours prise en compte par les systèmes de résolution des coréférences. 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 prises en compte par les techniques basées sur un apprentissage automatique sur des corpus annotés (Recasens, 2010). Si ces contraintes ont prouvé leur utilité sur la langue écrite, elles n'ont guère été questionnées sur la parole spontanée. Pourtant, la présence de disfluences (reprises, corrections, incises…) et l'usage fréquent de métonymies peuvent prêter lieu, par exemple, à des anaphores associatives où le respect de l'accord n'est pas garanti. La mise en œuvre d'une étude distributionnelle fine sur l'ensemble du corpus ANCOR est ainsi à même de mieux quantifier les éventuelles infractions à ces règles d'accord.

5.1 Accord en genre

Le Tableau 9 présente les résultats de l'accord en genre, à la fois par sous-corpus et par type de relations. Ont été considérés comme significatifs les résultats qui correspondaient au minimum à 100 relations. On remarque tout d'abord une très forte corrélation des résultats obtenus dans chaque sous-corpus, ce qui suggère que nos données sont sans doute représentatives des observations qui peuvent être faites sur le langage parlé en général. En effet, ni le degré d'interactivité (plus grand sur les corpus OTG et Accueil_UBS), ni la finalisation du dialogue (plus importante sur le corpus OTG) ne semblent induire de variation sensible dans les résultats.

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801359

Dans l'ensemble, si nous pouvons considérer que l'accord en genre reste une contrainte forte en français parlé, les cas de désaccord ne sont pas négligeables : sur l'ensemble du corpus, 7,9% des relations ne respectent pas l'accord, et cette proportion est encore de 4,4% des cas si l'on se limite aux coréférences.

Tableau 9 : Accord en genre dans les relations de coréférence et anaphoriques, par type de relation et par sous-corpus ANCOR. (n.s. = non significatif, moins de 100 observations)

Corpus	ESLO	CO2	OTG	UBS	TOTAL ANCOR
Directe	99,9%	99,8 %	99,7 %	100,0%	**99,9%**
Indirecte	57,9%	64,9, %	*48,6 % (n.s.)*	*83,7% (n.s.)*	**58,6%**
Pronominale	97,9%	98,3 %	97,9 %	98,6%	**97,9%**
Associative	63,5%	66,7 %	48,8 %	*63,3% (n.s.)*	**63,0%**
Associative pronominale	82,1 %	*89,5% (n.s.).*	*67,9% (n.s.)*	*75,0 % (n.s.)*	**80,3%**
Total toutes relations	**92,3%**	**90,5 %**	**89,1 %**	**94,8%**	**92,1%**
dont total coréférences	***95,6%***	***94,5 %***	***95,3 %***	***98,1%***	***95,6%***

Toutefois, une analyse détaillée des taux d'accord par type de relation montre bien que l'hypothèse linguistique de respect de l'accord en genre reste valide sur l'oral dans les cas où il doit être effectivement attendu. On observe ainsi que l'accord est quasi-parfait dans le cas des coréférences directes (99,7% à 100%). Les rares contre-exemples que nous avons observés sont le plus souvent liés à des erreurs sur le genre du terme employé comme dans cet exemple :

(3) Loc1 : *et tu sais le matin une petite gnôle*
Loc 2 : *vous ne connaissez pas le gnôle ah c'est bon* [ANCOR_ESLO_025]

Dans l'exemple suivant, la reprise nominale directe concerne en fait un pronom nominalisé, qui cherche à s'accorder avec un référent implicite qui peut changer de genre pendant le travail d'élaboration lexicale :

(4) *un crayon (...) une plume (...) un stylo plume (...) de la première (...) du premier* [ANCOR_ESLO]

De même, les taux d'accord restent très élevés pour l'anaphore pronominale : dans le pire des cas (corpus ESLO et OTG), on observe en effet un taux d'accord de 97,9%. Il reste toutefois 2,1% des cas où l'accord n'est pas réalisé, tel que cet exemple extrait du corpus ACCUEIL_UBS

(5) *j'ai une personne pour toi au téléphone pour les diplômes oui c'est bon je te le passe alors* [ANCOR_UBS]

Pour les autres types de relation, l'accord ne va pas de soit en principe. L'indirecte fait appel à une autre tête lexicale, ce qui n'implique pas la reprise d'un terme du même genre si l'on part du principe que le genre est arbitraire en français (exemple : *la voiture ... ce cabriolet*). De même, l'anaphore associative ne réfère pas à la même entité du discours, il n'y a aucune raison d'observer un accord en genre. C'est ce que nous observons dans le tableau 6, les pourcentages d'accord dans toutes ses situations étant comprises (pour les valeurs considérées comme significatives) entre 48,8% (OTG, anaphore associative) et 82,1% (ESLO, associative pronominale).

Tableau 10 : Accord en genre pour les relations indirectes et associatives, comparativement à une attribution aléatoire des genres aux mentions. (n.s. = non significatif, moins de 100 observations)

Corpus	ESLO	CO2	OTG	UBS	TOTAL ANCOR
Indirecte (observation)	57,9%	64,9, %	*48,6 % (n.s)*	*83,7% (n.s.)*	**58,6%**
Aléatoire (simulation)	**47,3%**	**48,4%**	**42,0%**	**44,9%**	**46,9%**
Associative (observation)	63,5%	66,7 %	48,8 %	*63,3% (n.s.)*	**63,0%**

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801359

Pour autant, ces taux d'accord sont relativement étonnants. Nous avons en effet calculé, en prenant en considération la répartition des termes nominaux suivant leur genre (cf § 4.1), le taux d'accord qui serait attendu si l'attribution des genres entre les termes en relation était faite de manière complètement aléatoire. Le tableau 10, qui met en regard cette estimation statistique avec les résultats observés, montre que le taux d'accord en genre est notablement supérieur à celui attendu pour les relations indirectes comme pour les anaphores associatives. Des estimations menées avec un test de Student (paramétrique) et un test de Wilcoxon (non paramétrique), montrent que les accroissements que nous avons observés sont statistiquement significatifs[5].

Dans le cas de la coréférence indirecte, ce résultat semble suggérer que si le genre est arbitraire, il a tout de même une fonction classificatoire. C'est bien entendu le cas des animés, mais l'accroissement de l'accord observé semble aller au-delà de cette seule catégorie d'entité nominale. Malheureusement, le corpus ANCOR ne distingue pas à l'heure actuelle les animés des autres mentions. Dans le cas des anaphores associatives, la situation est encore plus troublante puisqu'il n'y a pas d'identité de référent dans ce cas. Il ne peut être question d'artefact ici, puisque les anaphores associatives représentent tout de même 2776 observations sur le corpus ESL0, 353 cas sur CO2 et encore 146 sur OTG. Nous sommes actuellement en train de mener une étude qualitative pour esquisser une explication à cette observation.

5.2 Accord en nombre

Nous avons mené le même type d'étude sur l'accord en nombre, toujours avec l'outil de requêtage ANCORQI. Le Tableau 11 présente les résultats obtenus, pour lesquels on observe une fois encore une stabilité remarquable des observations sur les quatre sous-corpus. A l'opposé, cette étude amène également à d'autres conclusions que précédemment. En effet, les résultats présentés dans le Tableau 11 montrent que l'accord en nombre est sensiblement moins respecté que l'accord en genre en français parlé conversationnel. Dans l'ensemble, l'accord en nombre est seulement respecté dans 86,5% des relations, et ce pourcentage de cas sans accord reste élevé (9,3%) si l'on se restreint aux seules coréférences. Ce résultat recoupe les observations de (Antoine, 2004) qui ne concernaient que l'anaphore pronominale, et donne un caractère tout relatif à la pertinence d'une hypothèse sur un accord systématique en nombre.

Tableau 11 : Accord en nombre dans les relations de coréférence et anaphoriques, par type de relation et par sous-corpus ANCOR. (n.s. = non significatif, moins de 100 observations)

Corpus	ESLO	CO2	OTG	UBS	TOTAL ANCOR
Directe	90,2%	87,0 %	92,2 %	95,7 %	**90,1%**
Indirecte	76,7%	80,7 %	66,2 % (n.s)	93,0 % (n.s.)	**76,9%**
Pronominale	93,9%	92,0 %	89,5%	98,3 %	**93,6%**
Associative	53,5%	71,0 %	56,2 %	76,7 % (n.s.)	**54,1%**
Associative pronominale	53,5%	*50,0 % (n.s.)*	*16,0% (n.s.)*	*75,0 % (n.s.)*	**47,1%**
Total toutes relations	**86,9%**	**83,2 %**	**83,0 %**	**94,8 %**	**86,5%**
dont total coréférences	***90,8%***	***88,4%***	***89,1%***	***96,8%***	***90,7%***

Fait encore plus intéressant, ce constat d'accord imparfait vaut pour chaque type de relations. En particulier, un nombre remarquable d'anaphores directes ne présente pas d'accord en nombre (taux d'accord : 90,1% uniquement sur le corpus CO2, par exemple). 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, comme le montre l'exemple suivant :

(6) *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 dépasse, sauf lorsque le malade pose des exigences ou s'il s'agit d'une urgence.* [ANCOR_ESLO]

(7) *d'accord parce qu'en fait les dates d'incription c'est quand la clôture (...) c'est pour la date d'inscription* [ANCOR_UBS]

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801359

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, qui peuvent chuter jusqu'à 76,7% dans le cas de la coréférence indirecte (corpus ESLO) voire même 66,2% dans le cas du corpus OTG (sur 96 observations seulement). Ces résultats ne sont pas étonnants et ne semblent pas spécifiques à la modalité orale spontanée : le taux d'accord que nous observons avec les anaphores pronominales est ainsi 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 alii (2002) observent comme nous que l'utilisation de pluriels génériques ou de noms collectifs est une des principales causes de non respect de l'accord en nombre.

Enfin, l'accord en nombre chute encore plus fortement dans le cas des anaphores associatives (53,5% d'accord sur le corpus ESLO, par exemple). Cette observation est moins étonnante, puisqu'alors les deux mentions en relation associative ne partagent pas le même référent : l'antécédent peut ainsi être par exemple au singulier (*la maison*), tandis que l'anaphore associative au pluriel (*ses volets*). Notons toutefois que sur nos exemples, la présence de métonymie est une explication très fréquente de ce manque d'accord, comme le montre l'exemple suivant :

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

La comparaison avec une attribution aléatoire du nombre est ici éclairante (tableau 12). Tout d'abord, elle confirme que le nombre n'est pas arbitraire : on observe que même s'ils sont en baisse, les taux d'accord observés dans le cas de la coréférence indirecte sont nettement supérieurs aux valeurs attendues dans le cas d'une affectation au hasard du nombre. Ces différences sont stastitiquement significatives[6]. Il y a dans ce cas identité de référence, il est normal que les mentions en relation relèvent préférentiellement du même nombre, même si on observe effectivement des cas de non accord dans le cas de termes génériques (cf. précédemment). A l'opposé, il n'est pas possible d'observer d'écart statistiquement significatif dans le cas de l'anaphore associative[7] : les référents ne sont pas identiques, et il n'existe pas de raison particulière pour les deux mentions de la relation aient (*la voiture … son volant*) ou n'aient pas (*la voiture … ses freins*) le même nombre.

Tableau 12 : Accord en nombre pour les relations indirectes et associatives, comparativement à une attribution aléatoire des genres aux mentions. (n.s. = non significatif, moins de 100 observations)

Corpus	ESLO	CO2	OTG	UBS	TOTAL ANCOR
Indirecte (observation)	76,7%	80,7 %	66,2 % (n.s)	93,0 % (n.s.)	**76,9%**
Aléatoire (simulation)	52,5%	54,8%	50,4%	59,3%	**52,6%**
Associative (observation)	53,5%	71,0 %	56,2 %	76,7 % (n.s.)	**54,1%**

Pour conclure, cette étude a clairement prouvé que l'accord en nombre est modérément respecté quelque soit le type de relation anaphorique considéré. Le taux d'accord moyen en nombre de 86,5% sur l'ensemble du corpus montre en effet qu'il serait risqué que les processus de résolution le considèrent comme obligatoire 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. On pourra nous objecter que ces taux d'accord ont été mesurés avec la première mention, alors que les systèmes symboliques de résolution des coréférences considèrent généralement uniquement la mention la plus proche. Lorsque nous disposerons également du codage en chaîne des relations, il nous sera possible d'observer si l'accord en genre ou en nombre est mieux respecté. Dans l'immédiat, rappelons que les approches par apprentissage utilisant des classifieurs binaires considèrent des traits d'accord en genre et nombre qui peuvent aussi bien concerner la première que la dernière mention. Les résultats présentés dans ce paragraphe les concernent donc directement.

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801359

Nous avons vu que 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 et de l'utilité, par son envergure, du corpus ANCOR pour de telles études.

6 Conclusion

Dans cet article, nous avons présenté le corpus libre ANCOR_Centre et l'outil de requêtage ANCORQI qui lui est associé. ANCORQI, qui est appelé à être intégré dans une interface interactive d'interrogation, permettra aux chercheurs non informaticiens d'avoir à leur disposition un outil efficace pour mener des analyses sur ce qui constitue à l'heure actuelle un des corpus mondiaux les plus importants sur la coréférence, et une ressource sans aucune mesure équivalente pour le français. Nous avons présenté une étude distributionnelle des différentes entités qui peuplent le corpus, de même qu'une étude sur l'accord en genre et en nombre qui démontrent, nous l'espérons, l'intérêt de cette ressource. Nos observations quantitatives sur l'accord nous ont fourni des indications précieuses pour le linguiste aussi bien que pour le TAL (pertinence des traits d'accord pour les systèmes de résolution des coréférences), ce qui était un des objectifs initiaux du projet ANCOR qui a conduit à la réalisation de cette ressource. Nos travaux futurs vont consister bien entendu à mener des études linguistiques approfondies sur le corpus, qui sera par ailleurs enrichi par une annotation en coréférence temporelle dans le cadre du projet TEMPORAL (financement : MSH Val de Loire) qui débute en 2014.

Références bibliographiques

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

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

Ariel M. (1990). *Accessing Noun-Phrase Antecedents*, Londres : Routledge.

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

Baude O., Dugua, C. (2011). (Re)faire le corpus d'Orléans quarante ans après : quoi de neuf, linguiste ? *Corpus*, 10, pages 99-118.

Chiarcos C. (2009). *Mental Salience and Grammatical Form: Toward a Framework for Salience Metrics in Natural Language Generation*. Doctoral dissertation. Postdam U.

Cohen J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, pages 37-46

Cornish F. (1999). *Anaphora, Discourse, and Understanding. Evidence from English and French* Oxford : Clarendon.

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*, pages 54–59.

Eshkol-Taravella I., Baude O., Maurel D., Hriba L., Dugua C., Tellier I., (2012). Un grand corpus oral « disponible » : le corpus d'Orléans 1968-2012. *TAL*. 52(3), pages 17-46.

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

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801359

Grosz B., Joshi A., Weinstein S. (1995). Centering : a framework for modelling the local coherence of discourse. U. Pennsylvania, Philadelphia, PA. IRCS Report 95-01.

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

Heindrickx I., Bouma G., Coppens F., Daelemans W., Hoste V., Kloosterman G., Mineur A.-M., Van Der Vloet J., Verschelde J.-L. (2008). A coreference corpus and resolution system for Dutch. *Proc. LREC'2008*.

Hinrichs E., Kübler S., Naumann K., Zinsmeister H. (2005). Recent developments in linguistic annotations of the TüBa-D/Z Treebank. *27th Meeting of the German Linguistic Association*, Köln.

Iida R., Mamoru K., Kentaro I., Yuji M. (2007). Annotating a Japanese text corpus with predicate-argument and coreference relations. Proc. *Linguistic Annotation Workshop*, 132-139. Stroudsburg.

Hobbs J. R. (1978). Resolving Pronoun References', *Lingua*, Vol. 44, pages 311-338

Kenny I., Huyck C. (2011). Resolution of Anaphoric and Exophoric Definite Referring Expressions in a Situated Language Environment; *Proc. DAARC'2011*. University of Lisbon, Portugal

Kleiber G. (2002). Marqueurs référentiels et théorie du centrage, *Linx*, 47.

Krippendorff K. (2004). *Content Analysis: an Introduction to its Methodology*. Sage: Thousand Oaks, CA

Lappin S., Leas H. (1994). An algorithm for pronominal anaphora resolution. *Computational Linguistics 20(4)*, pages 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*, pages 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*, pages 1–12.

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

Muzerelle J., Schang E., Antoine J.-Y., Eshkol I., Maurel D., Boyer-Pelletier A., Nouvel D. (2012). Annotation en relations anaphoriques d'un corpus de discours oral spontané en français. Actes. *1er Congrès Mondial de Linguistique Française, CMLF'2012*, Lyon

Nedoluzzhko A., Mírovský J., Ocelák R., Pergler J. (2009). Extended coreference relations and bridging anaphora in the Prague Dependancy Treebank. Proc. *DAARC'2009*, pp. 1-16. Chennai Goa, Indica.

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

Ogrodniczuk M., Kopeć M., Głowinska K., Savary A., Zawisławska, M. (2013). Polish coreference corpus, submitted to *LTC'2013*.

Passoneau R. (2004). Computing reliability for Co-Reference Annotation. *LREC'2004*.

Poesio M., Stevenson R., di Eugenio B., Hitzeman J. (2004). Centering: A Parametric theory and its instantiations. *Computational Linguistics*, 30(3), pages 309-363

Pradhan S. S., Ramshaw L., Weischedel R. MacBride J., Micciula L. (2007). Unrestricted coreference: identifying entities and events in OntoNotes. Proc. *1st IEEE Int. Conf. on Semantic Computing (ICSC'07).* pages 446-453. Washington, DC. USA. IEEE.

Prince A., Smolensky P. (1993). Optimality Theory: Constraint Interaction in Generative Grammar. Blackwell.

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

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

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801359

Schang E., Boyer-Pelletier A., Muzerelle J., Antoine J.-Y., Eshkol I., Maurel D. (2011). Coreference and anaphoric annotations for spontaneous speech corpus in French, *Proc. Discourse Anaphora and Anaphor Resolution Colloqium, DAARC'2011*., Faro, Portugal.

Scott W. (1955). Reliability of content analysis: the case of nominal scale coding. *Public Opinions Quaterly*. 19, pages 321-325.

Sjöblom M. K. (2002). L'écriture de J.M.G. Le Clezio, une approche lexicométrique. Thèse U. Nice

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

Tutin A., Trouilleux F., Clouzot C., Gaussier E., Zaenen A., Rayot S., Antoniadis G. (2000). Annotating a large corpus with anaphoric links. In *Proc. DAARC2000*.

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

Vieira R., Salmon-Alt S., Schang E. (2002). Multilingual Corpora Annotation for Processing Definite Descriptions. *Proc. Portugal for Natural language Processing, PorTAL 2002*, Faro, Portugal.

---

[1] Pour l'annotation du corpus CRISTAL cinq relations sont codées (nous reprenons les exemples de Tutin et al., 2000) :

- *Coreference* : mêle la coréférence directe et pronominale,
- *Set membership* : requière que l'antécédent réfère à un ensemble, (**des quatre locomotives** … **l'une**),
- *Description* : permet de lier toutes les propriétés d'un référent à celui-ci, et relier des reprises anaphoriques lorsqu'elle traitent de ces dites propriétés (si toutes les ressources énergétiques naturelles sont **exploitées**…l'énergie hydraulique **l'**est insuffisamment),
- *Sentencial antecedent* : lorsque l'antécédent est plus qu'un syntagme interne à la phrase, lorsqu'il est constitué d'une proposition ou d'une phrase (**Ces records se déroulent**, il faut **le** dire, **dans une période exceptionnellement favorable à l'innovation en France**),
- *Indefinite relation* : permet de récupérer toutes les autres relations qui ne peuvent être classées dans les quatre catégories énoncées, par exemple, lorsque la reprise anaphorique est sous la portée d'une négation (parmi **ces étudiants**, **aucun** n'a fait son travail).

Ces différentes catégories ne permettent pas de mettre en valeur l'impacte de la tête lexicale sur l'accord en genre et en nombre par exemple, ni le fonctionnement des relations associatives en tant que telles. Ainsi les relations associatives et associatives pronominales ne peuvent être observées.

[2] La finalisation d'un discours est définie par rapport au but de celui-ci, en d'autres termes, le locuteur initiant le dialogue poursuit un but pragmatique. Par exemple, un locuteur appelle l'office du tourisme pour obtenir le numéro de la mairie, le discours sera donc très fortement finalisé, avec pour finalité l'obtention par ce premier locuteur du numéro de téléphone en question.

[3] Un schéma d'annotation particulier était nécessaire pour ces phénomènes (Dipper & Zinmeister, 2010)

[4] A titre d'exemple, les différences d'occurrences de relations indirectes sont importantes entre le corpus CO2 et OTG, puisque l'on assiste à un presque doublement de ce procédé (6,1% contre 11,2%). Cette variation correspond toutefois à une valeur de test de Student de 0,627, bien inférieure au seuil d'erreur de 10% T(0,1) qui serait de 1,711 sur ce jeu de données.

[5] Un test de Wilcoxon donne une p-valeur de 0,00098 dans les deux cas. Celle du test de Student (qui suppose une distribution normale des données) est de 0,000299 pour la comparaison indirecte/aléatoire et de $1{,}04\ 10^{-5}$ pour la comparaison associative/aléatoire. Ces tests ont été effectués en divisant ANCOR en 11 sous-corpus de tailles comparables (UBS et OTG pris intégralement, CO2 scindé en 3 et ESLO en 6).

[6] Suivant la même méthodologie que précédemment, on obtient une p-valeur de 0,000977 avec un test de Wilcoxon, qui ne fait aucune supposition sur la distribution des données, et de $7{,}20\ 10^{-8}$ avec un test de Student.

[7] Un test de Wilcoxon donne une valeur de 0,197 supérieur au seuil de significativité de 0,01 (1% d'erreur de rejeter l'hypothèse H0) de même que le test de Student donne une p-valeur proche de 0,186.