Corpus ANCOR_Centre

Présentation générale

Jean-Yves Antoine[1], Emmanuel Schang[2], Judith Muzerelle[2], Anaïs Lefeuvre[1], Aurore Pelletier[2], Adèle Désoyer[4], Frédéric Landragin[4], Isabelle Tellier[4], Jeanne Villaneau[3], Iris Eshkol[2], Denis Maurel[1],

[1]**LI – Université François Rabelais de Tours** (EA 6300)
[2]**LLL – CNRS** (UMR 7270)
[3]**IRISA** (UMR 6074)
[4]**LATTICE** (UMR 8094)

Université François Rabelais Tours & Université d'Orléans

http://www.info.univ-tours.fr/~antoine/parole_publique/

Introduction

Ce document présente en détail le corpus ANCOR_Centre (par la suite : ANCOR, en abrégé), un corpus de parole spontanée transcrite annoté en relations de coréférence et relations anaphoriques. Il a été réalisé par les laboratoires LI (Université François Rabelais de Tours) et LLL (CNRS – Universités d'Orléans et François Rabelais de Tours) tout d'abord dans le cadre d'un projet interne au PRES Centre Val de Loire Université (Projet CO2) puis, pour sa plus grande partie, dans le cadre du projet ANCOR financé par la Région Centre (APR-IA 2012).

Dans le cadre de l'ANR ORFEO (Outils et Recherches sur le Français Ecrit et Oral), le laboratoire LATTICE (UMR 8094) a par ailleurs réalisé une version du corpus avec annotation intégrée et non plus déportée, plus utilisable par des algorithmes d'apprentissage automatique.

L'ensemble du corpus est diffusé librement sous licence Creative Commons et est directement récupérable sur Internet (voir §2 : distribution).

Plus précisément, ce rapport présente :

- le contenu du corpus distribué ainsi que les conditions dans lesquelles il a été réalisé,
- les modes de distributions du corpus,
- la convention à laquelle est liée l'utilisation d ce corpus à toutes fins scientifiques ou industrielles,
- les références bibliographiques associées à ce corpus.
- les conventions d'encodage et d'annotation suivies lors de la réalisation du corpus,

1 Présentation du corpus : contenu et conditions d'enregistrement

Le corpus ANCOR est un corpus de langue parlé dans lequel ont été annotées les anaphores relations de coréférence nominales ou pronominales. Afin d'atteindre une certaine diversité dans les situations de parole spontanée concernées, cette annotation a concerné quatre corpus oraux préalablement transcrits.

La présente distribution se limite à l'annotation du corpus, c'est-à-dire qu'elle se limite aux fichiers de transcription et d'annotation en coréférence. Les personnes désireuses d'obtenir en sus les fichiers audio correspondant à ces corpus devront se tourner vers les sites de distribution des corpus oraux originaux (cf infra § 1.2). Ces corpus sont également accessibles librement sous licence Creative Commons.

1.1 Fiche signalétique

Corpus	ANCOR_Centre (en abrégé : ANCOR)
Version	1.1 (novembre 2014)
Type d'oral	Parole spontanée : interview ou dialogue oral homme-homme finalisé
Taille	488 000 mots – 30,5 heures d'enregistrement
Locuteurs	Adultes hommes ou femmes
Enregistrement	Conditions réelles
Contenu	Transcription orthographique + annotation en anaphore et coréférence
Concepteur(s)	Laboratoires LI (Université François Rabelais de Tours) et LLL (CNRS)
	Jean-Yves Antoine (LI), Emmanuel Schang (LLL)
Annotation	Judith Muzerelle (LLL), Aurore Pelletier (LLL)
Révision	Judith Muzerelle (LLL), Anaïs Lefeuvre (LI)
Format intégré	Adèle Désoyer (LATTICE), Frédéric Landragin (LATTICE), Isabelle TELLIER (LATTICE)
Evaluation fiabilité	Jeanne Villaneau (IRISA), Iris Eshkol (LLL), Denis Maurel (LI), Judith Muzerelle (LLL), Anaïs Lefeuvre (LI), Jean-Yves Antoine (LI), Emmanuel Schang (LLL)
Diffusion	Licence Creative Commons CC-BY-NC-SA

1.2 Contenu : corpus audio

Le corpus ANCOR résulte de l'annotation de trois corpus oraux transcrits sous Transcriber (Barras et al., 2001) qui étaient disponibles au sein des laboratoires LI et LLL et sont également diffusés librement sous licence Creative Commons CC-BY-NC-SA

1) ESLO, qui correspond à des entretiens sociolinguistiques (Baude et Dugua 2011, Eshkol-Taravella et al. 2012). Dans la présente distribution, le corpus ESLO sera divisé en deux sous-corpus annotés :

   - ESLO_CO2, réalisé dans le cadre du projet CO2 et qui correspond à l'annotation de 3 entretiens complets.
   - ESLO_ANCOR, réalisé dans le cadre du projet ANCOR (région Centre) et pour lequel les entretiens ont été divisés en sous-dialogues thématiquement cohérents.
   - OTG, qui correspond à des dialogues interactifs entre des individus et le personnel d'accueil de l'Office du Tourisme de Grenoble. Ces corpus ont été annotés dans le cadre du projet ANCOR.
   - Accueil_UBS, qui correspond à des dialogues interactifs par téléphone recueillis auprès du standard téléphonique d'une université (Nicolas et al., 2002). Ces corpus ont également été annotés dans le cadre du projet ANCOR.

Notre objectif a été de représenter une certaine diversité de genres en termes de degré d'interactivité du dialogue. Le corpus ESLO, qui correspond à des entretiens, a une interactivité limitée à la différence des deux autres : le plus souvent, l'enquêteur pose en effet une question à laquelle s'ensuit un assez long monologue de réponse. Le tableau 1 présente la distribution des corpus oraux dans la ressource annotée.

Corpus Oral	Type dialogue	Diffusion	Taille	Durée	Financement annotation
ESLO_ANCOR	Interview	CC-BY-NC-SA	417 kMots	25 h	Projet CO2 (CVL Université)
ESLO_CO2	Intervieuw	CC-BY-NC-SA	35 kMots	2,5 h	Projet ANCOR (Région Centre)
OTG	Dialogue H-H	CC-BY-SA	26 kMots	2 h	Projet ANCOR (Région Centre)
Accueil_UBS	Dial. téléphone	CC-BY-SA	10 kMots	1 h	Projet ANCOR (Région Centre)
TOTAL			**488 kMots**	**30,5 h**	

Tableau 1 – Répartition des corpus oraux annotés dans ANCOR

1.3 Méthodologie d'annotation et estimation de la fiabilité des données

L'annotation a été réalisée sur le logiciel *GLOZZ* (Mathet et Widlöcher, 2009). La présente version (1.0) du corpus correspond donc à une annotation sous format *GLOZZ*. A terme, le corpus ANCOR sera toutefois également diffusé sous format *MMAX2* (Müller et Strube, 2006) du fait de la grande diffusion de ce dernier.

Le corpus ANCOR a fait l'objet d'un codage par plusieurs annotateurs suivi d'une révision, selon une procédure en quatre phases successives :

1) Repérage et caractérisation des entités nommées et autres mentions par un annotateur,
2) Révision croisée du repérage par l'autre annotateur et recherche de consensus,
3) Repérage et caractérisation des relations anaphoriques par un annotateur,
4) Révision finale des relations caractérisées par un superviseur.

Cette démarche séquentielle évite une surcharge cognitive aux codeurs et favorise la cohérence des annotations sur la durée. Le schéma détaillé d'annotation est décrit dans la section suivante et en annexe.

1.4 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, puis si elle réfère à une entité précédemment mentionnée (coréférence) ou si la référence a 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. Pour une vue détaillée et exhaustive de ce schéma, on consultera en annexe le guide d'annotation qui a été suivi par nos experts.

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. Il s'agit d'un choix fort qui peut effectivement induire l'oubli par l'annotation certaines coréférences, particulièrement dans le cas d'une référence temporelle. Nous avons fait ce choix afin de nous assurer d'une fiabilité maximale des données. Notre expérience a en effet montré que les codeurs éprouvaient de fortes difficultés à savoir ce qui devait être ou non considéré comme une mention ou

pas si l'on ne se limitait pas à une définition purement syntaxique (noms et pronoms) des mentions. Le corpus ANCOR propose ainsi une annotation fiable (cf § 1.5) définie sur des critères précis. Si celle-ci ne suffit pas à vos besoins, il vous suffit de la compléter sous GLOZZ.

L'annotation considère 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 est prise en compte sous forme d'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, qui dépassent largement les objectifs d'annotation du corpus ANCOR.

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. Dans le corpus ANCOR, il a été décidé 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 alors beaucoup moins évidente à caractériser pour les annotateurs. Par ailleurs, le codage en première mention rend compte des changements de genre grammatical lors de reprises successives comme dans la séquence "*j'ai une personne qui* (...) *elle téléphone* (...) *c'est un étudiant de L1 ... elle... il...*" où toutes les entités sont coréférentes.

Des arguments d'ordre linguistique ou computationnel peuvent toutefois être trouvés en faveur de chaque représentation. Dans l'immédiat, le corpus distribué (version 1.0) n'est codé qu'en première mention. Mais ses évolutions futures offriront également un codage en chaîne et un codage en clusters de coréférents.

Caractérisation des relations et de leurs entités

A fins d'études linguistiques ou d'apprentissage automatique, l'annotation associe plusieurs propriétés aux mentions et à leurs éventuelles relations. Les traits linguistiques suivants servent à décrire les mentions :

- **G** : Genre et **N** : Nombre
- **POS : partie du discours** – Ce trait peut prendre les valeurs P (pronom), N (Nom) ou NULL (artefact lié à certaines disfluences)
- **GP : inclusion dans un GP** – Valeur YES (si l'entité est un GP) ou NO (si c'est un GN)
- **EN : entité nommée** – Types retenus dans la campagne d'évaluation ESTER2 (Galliano et al., 2009), à savoir FONC, LOC, PERS, ORG, PROD, TIME, AMOUNT et EVENT. 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.
- **NEW : nouvelle entité du discours** : YES (première mention), NO (entité coréférente avec une autre). Une mention isolée recevra donc toujours le type YES.

Les relations sont caractérisées par un type (trait **TYPE**). Nous distinguons les types de relations suivantes :

- **DIR : direct**, dans le cas d'une coréférence entre mentions de même tête nominale,

  Exemple : *le bus rouge…. ce grand bus*

- **IND : indirect**, si les entités coréférentes ont des têtes nominales différentes,

  Exemple : *le cabriolet… cette décapotable*

- **PR : pronominal**, dans le cas particulier de l'anaphore indirecte où la reprise est un pronom,

  Exemple : *le cabriolet … il roulait…*

- **ASSOC** : associatif (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** : associatif pronominal, dans le cas où la reprise associative est portée par un pronom comme dans l'exemple de métonymie ci-dessous :

  Exemple : *Le Café Jeanne d'Arc, ils sont tous désagréables*.

1.5 Estimation de la fiabilité des données

La fiabilité du corpus a été estimée sur une expérience pilote 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 reste une question ouverte dans le cas de la coréférence, du fait des problèmes d'alignement entre annotations (Passoneau, 2004 ; Artstein et Poesio, 2008 ; Matthet et Widlöcher, 2011). Nous avons contourné ce problème par le calcul de mesures d'accords successifs sur la délimitation des paires coréférentes (ou en relation anaphorique) et seulement ensuite sur le typage de ces relations. Cette fiabilité a été estimée par trois mesures d'accord inter-annotateur : κ (Cohen, 1960), π (Scott, 1955) et α (Krippendorff, 2004).

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

Tableau 2 – Mesures de fiabilité sur le corpus ANCOR.

Comme le montre le tableau 2, l'accord inter-annotateur traduit une excellente fiabilité (mesure de 0,80 sur toutes les métriques) sur la tâche de typage des relations. A l'opposé, cet accord est bien plus faible sur la tâche de délimitation, où nous avons effectivement observé des choix différents entre les annotateurs, en dépit d'un apprentissage préalable du guide d'annotation. Cette estimation de la fiabilité (0,45 sur toutes les métriques) est en dessous du seuil de 0,64 généralement accepté. Il faut toutefois comprendre que cette métrique est pénalisée par notre annotation en première mention : une divergence sur la délimitation de la première mention pourra en effet entraîner une mesure de désaccord sur tous les termes coréférents qui la suivent dans la chaîne. Une fois que le corpus sera transformé également en annotation en chaîne, il nous sera possible d'estimer l'accord inter-annotateur de manière non biaisée.

Dans l'immédiat, nous avons procédé à une expérimentation avec la superviseuse principale de l'annotation, en lui demandant de reproduire l'annotation d'un extrait du corpus, et de comparer celle-ci avec l'annotation réalisée précédemment 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 qui a été conduite est très fortement cohérente.

1.6 Contenu du corpus annoté : quelques données

Le tableau 3 ci-dessous donne quelques information sur le contenu annoté présent dans le corpus. On notera que l'annotation porte sur plus d'un millier de mentions, et plus de 50,000 relations. Pour plus de renseignement sur la distribution de ces différents éléments (par type de mention, de relation, etc…), on se référera aux publications citées en référence bibliographique.

Corpus	ESLO_ANCOR	ESLO_CO2	OTG	Accueil_UBS	TOTAL
Nb. de mentions	97,939	8,798	7,462	1,872	**116,071**
Nb. de relations	44,597	3,513	2,572	655	**51,337**
Mention/relation ratio	*2.19*	*2.50*	*2.90*	*2.86*	***2.26***

Tableau 3 – Contenu des différents sous-corpus annotés.

1.7 Organisation du corpus distribué

Comme précisé précédemment, le corpus ANCOR a été réalisé à l'aide de la plate-forme logicielle GLOZZ. Le format d'encodage correspondant à la version actuelle (1.0) du corpus est donc celui adopté par GLOZZ.

Le principe d'encodage retenu par GLOZZ est celui d'une annotation déportée (*stand-off annotation*), c'est-à-dire que fichiers d'annotation et fichier corpus sources sont séparés, les fichiers d'annotation pointant sur des éléments des fichiers corpus source. L'intérêt de ce type d'annotation déportée est de faciliter l'évolutivité du corpus en permettant l'ajout de nouvelles couches d'annotation sous forme de fichiers déportés additionnels. Il rend par contre plus difficile l'utilisation directe du corpus en TAL (création de modèles par apprentissage automatique, utilisation du corpus comme ressource d'évaluation). C'est pourquoi le corpus est diffusé à la fois sous la forme d'annotation déportée *Glozz*, et d'annotation intégrée.

Le corpus ANCOR se compose ainsi de trois types de fichiers correspondant aux deux types d'annotation dsitribuées :

Annotation déportée (format Glozz)

- Les fichiers corpus (transcriptions), qui reçoivent l'extension `.ac` suivant la recommandation de GLOZZ.
- Les fichiers d'annotation, qui reçoivent l'extension `.aa` suivant la recommandation de GLOZZ.

Ces fichiers tous encodés au format XML. L'annotation repose sur une DTD spécifique (fichier avec extension spécifique `.aam`) qui est adaptée à notre schéma d'annotation. S'il n'est pas nécessaire de disposer de cette DTD pour lire les fichiers d'annotation sous GLOZZ, celle-ci devient nécessaire si vous désirez modifier ou enrichir l'annotation. Ces annotations sont facilement consultables et analysables sous GLOZZ, mais vous pouvez bien entendu développer vos propres outils de parsage XML pour travailler sur le corpus. Pour plus de renseignement sur l'utilisation et le fonctionnement du logiciel Glozz, on se référera directement à la distribution de cet outil disponible librement à l'adresse suivante : http://glozz.free.fr/

Annotation intégrée (annotation ancrée)

- Fichiers intégrant à la fois le corpus source et l'annotation en entités référentielles et en relations de coréférence ou anaphoriques. Cette conversion est décrire dans (Désoyer 2014)

Dans ce format intégré, le corpus correspond à un ensemble de fichiers xml dont la racine a deux éléments fils : un <Trans> où sont intégrées les données du fichier corpus source correspondant à la transcription, puis un <annotation> sous lequel sont intégrées les données du fichier d'annotation. Le principe retenu est d'ancrer dans la partie <Trans> les différentes unités référentielles présentes dans la transcription à l'aide d'une balise <anchor> qui attribue un identifiant unique à ces unités. Puis de décrire ensuite ces unités avec leurs attributs, ainsi que leurs éventuelles relations dans la partie <annotation>. Pour ce faire, le nœud <annotation> couvre un ensemble d'éléments fils balisés <unit> ou <relation> qui reprennent les identifiants d'unités caractérisées dans la partie <Trans>.

La figure 2 page suivantes illustre l'organisation des fichiers résultant de cette annotation ancrée. Pour plus de précision, on consultera l'annexe C (*Passage de l'annotation Glozz à l'annotation ancrée*).

Corpus complet distribué

La figure 1 ci-dessus décrit l'arborescence des fichiers telle qu'elle se présente dans la distribution du corpus. A un premier niveau, on trouve une séparation entre les 4 sous-corpus composant la ressource. Au second niveau, les sous-corpus sont répartis entre fichiers source et fichiers d'annotation Glozz d'une part, et fichiers d'annotation intégrée d'autre part. Les textes de présentation de la ressource, ainsi que le DTD d'annotation (fichier `DTD_GLOZZ_ANCOR_DEFAULT.aam`) se trouve au premier niveau de l'arborescence de distribution, dans le répertoire DISTRIBUTON_ANCOR.

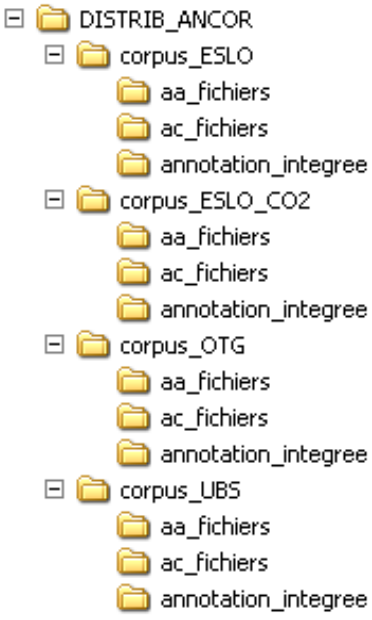


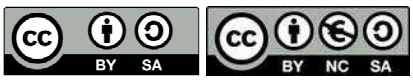
Figure 1 – Organisation des répertoires du corpus ANOCR

Figure 2 – Schéma illustratif d'organisation du format d'annotation ancré

2 Distribution du corpus et convention d'utilisation

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

- Site consacrée au projet : http://tln.li.univ-tours.fr/Tln_Corpus_Ancor.html
- Site Parole_Publique de diffusion de corpus oraux : www.info.univ-tours.fr/~antoine/parole_publique/
- Speech and Language Data Repository (Ortolang Resources) : http://sldr.org/ (resource 000903)



Le corpus est distribué gratuitement sous licence *Creative Commons* CC-BY-SA pour ce qui concerne les annotations sur les corpus OTG, Accueil_UBS et ESLO_CO2, et CC-BY-SA-NC pour le corpus ESLO_ANCOR. Cela signifie que vous devez respecter le contrat d'utilisation suivant :

- *BY : paternité* - Vous devez citer les auteurs de ce corpus pour toute utilisation du corpus. Dans le cas d'une publication s'appuyant sur ces travaux, nous vous demandons ainsi de citer les articles référencés dans la description de la ressource jointe à la distribution ou dans la liste ci-dessous.

- *SA : partage des conditions initiales à l'identique* - Vous ne pouvez créer une nouvelle ressource à partir de la ressource existante et en faire ensuite un usage différent de celui imposé par ce contrat. Là encore, nous sommes ouverts à toute utilisation du corpus pour création de nouvelles ressources, mais nous vous demandons de nous contacter pour discuter de ces nouveaux usages.

- *NC : pas d'usage commercial* sans l'accord du détenteur de la ressource.

Important - Par ailleurs, cette ressource intègre des échanges dont la communication porte atteinte à la protection de la vie privée ou portant appréciation ou jugement de valeur sur une personne physique nommément désignée, ou facilement identifiable, ou qui font apparaître le comportement d'une personne dans des conditions susceptibles de lui porter préjudice. (Code du Patrimoine, art. L. 213-2, I, 3) . A ce titre, ce corpus peut être utilisé à des fins d'analyse, mais en aucun cas ne peut être diffusés publiquement.

La distribution de ces corpus est **libre** quel que soit l'usage de ce corpus.

Par ailleurs, nous vous serions extrêmement reconnaissants de nous signaler toute utilisation du corpus à des fins de recherche ou industrielle, ainsi que de nous communiquer tout article reposant sur des données extraites du corpus. Ceci afin de nous permettre d'identifier les usages faits avec la ressource, pour son amélioration éventuelle à l'avenir.

3 Références bibliographiques

Liste des publications à la date de l'émission de ce rapport technique. Consultez le site Internet du projet Parole Publique pour une bibliographie à jour.

3.1 Publications concernant le corpus ANCOR ou ses sous-corpus

- MUZERELLE J., PELLETIER-BOYER A., ANTOINE J.Y., SCHANG E., ESKHOL I., MAUREL D., NOUVEL D. (2012). Annotation en relations anaphoriques d'un corpus de discours oral spontané en français. Proc. *Congrès Mondial de Linguistique Française*, *CMLF'2012*, Lyon.

- MUZERELLE J., LEFEUVRE A., ANTOINE J.-Y, SCHANG E,, MAUREL D., VILLANEAU J., ESHKOL I. (2013) ANCOR, premier corpus de français parlé d'envergure annoté en coréférence et distribué librement. Actes *TALN'2013*, Les Sables d'Olonnes.

- MUZERELLE J., LEFEUVRE A., SCHANG E., ANTOINE J.-Y, PELLETIER A., MAUREL D., ESHKOL I., VILLANEAU J. (2013) ANCOR_Centre, a large free spoken French coreference corpus: description of the resource and reliability measures. Proc. *LREC'2014* (submitted).

3.2 Publications citées dans ce document

- AMSILI, P., LANDRAGIN, F., ACOSTA, A., BITTAR, A. (2007). Résolution anaphorique : Etat d'une réflexion collective, *Actes Journées d'Etudes de l'ATALA 2007*, pages 1–4.

- ARTSTEIN, R., POESIO, M. (2008) Inter-Coder agreement for Computational Linguistics, *Computational Linguistics*, 34, pages 555-596

- 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), pages 5–22.

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

- DESOYER A. (2014) *Apprentissage d'un modèle de résolution automatique de la coréférence à partir d'un corpus de français oral*. Mémoire de recherche Master Documents Electroniques et Flux d'Informations,

Université Paris Ouest - Nanterre La Défense. Consultable sur la toile à l'adresse suivante : http://www.tal.univ-paris3.fr/plurital/memoires/Adele-Desoyer-memoire-TAL-RD-1314.pdf

- 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.
- GALLIANO, S., GEOFFROIS, E., MOSTEFA, D., CHOUKRI, K., BONASTRE, J., GRAVIER, G. (2005) The ESTER phase II evaluation campaign for the rich transcription of French broadcast news. Proc. *9th European Conference on Speech Communication and Technology*, *Eurospeech'2005*, Lisboa, Portugal
- GARDENT, C. et MANUELIAN, H. (2005). Création d'un corpus annoté de traitement des descriptions définies. *Traitement Automatique des Langues, TAL* , 46(1).
- MATHET, Y., WIDLÖCHER, A. (2009). La plate-forme GLOZZ : environnement d'annotation et d'exploration de corpus. *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. *Actes TALN 2011*, Montpellier, France.
- MÜLLER, C., STRUBE, M. (2006). Multi-level annotation of linguistic data with MMAX2. In: Braun, S., Kohn, K., Mukherjee, J., ed., *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*. Peter Lang, Francfort, Allemagne, pages 197–214.
- NICOLAS, P., LETELLIER-ZARSHENAS, S., SCHADLE, I., ANTOINE, J.-Y., CAELEN, J. (2002). Towards a large corpus of spoken dialogue in French that will be freely available: the "Parole Publique" project and its first realisations. *LREC'2002*. Las Palmas, Espagne. pp. 649-655.
- PASSONEAU, R. (2004) Computing reliability for Co-Reference Annotation. *LREC'2004*.

4 Financement

La réalisation de ce corpus a été financée dans le cadre de plusieurs financements distincts :

- projet ANCOR de la région Centre (APR-IA 2012), dotation de 90000 €
- projet CO2 du PRES Centre Val de Loire Université (2010), dotation de 3 000 €
- soutien de l'IRCOM Corpus Écrits à la finalisation du corpus, dotation de 1 500 €
- projet ANR ORFEO Outils et Recherches sur le Français Ecrit et Oral), dotation de 2 500 € destinée au financement du stage d'Adèle Désoyer.

ANNEXE A — Conventions d'annotation du corpus ANCOR

Guide d'annotation

Version *2.2*
Auteurs *Jean-Yves Antoine, Judith Muzerelle et Emmanuel Schang*
Date *6 juin 2013*

1 Annotation : principes généraux

L'objectif du projet est de décrire toutes les reprises référentielles existantes dans les corpus étudiés. Pour cela, il nous faut procéder à une annotation à deux niveaux :

- Annotation des entités nommées et plus généralement de tout élément susceptible d'intervenir dans une chaîne co-référentielle, à savoir tout groupe nominal et tout pronom ;
- Annotation des relations de co-référence entre les éléments annotés au niveau précédent.

Ces deux niveaux sont décrits conjointement dans les fichiers d'annotation GLOZZ. Il est toutefois conseillé de procéder à une annotation en deux passes correspondant à ces deux niveaux successifs.

Remarque – Les éléments spatio-temporels comme « *ici* », « *demain* », « *aujourd'hui* »… ancrent effectivement des anaphores. Cependant, ce type d'anaphores spatio-temporelles relève d'un autre problème et nous demanderait d'inclure dans l'annotation des adverbes alors que ce projet se limite à l'étude des groupes nominaux et des pronoms. En conséquence, nous avons choisi d'exclure ces éléments afin de garantir la cohérence de nos propos.

2 Annotation : délimitation des groupes nominaux et des pronoms

On annotera donc le corpus en déterminant tout d'abord l'ensemble des groupes nominaux et des pronoms (pronoms personnels, explétifs et relatifs).

2.1 Groupes nominaux simples

On délimitera comme groupe nominal non seulement la tête lexicale du groupe, mais également ses déterminants et adjectifs qualificatifs (voir plus loin le cas des compléments du nom ou des relatives). Donc par exemple :

- *Il regarde la maison bleue*	on délimite	*la maison bleue* et *il*
- *Il a désormais une tout autre maison*	on délimite	*une tout autre maison* et *il*
- *Je vois le Panthéon*	on délimite	*le Panthéon* et *je*
- *Il la regarde*	on délimite	*il* et *la*

Dans le cas d'un groupe nominal intégré à un groupe prépositionnel (GP), la préposition sera exclue de la segmentation :

- *Je vais à la recherche de Jean*	on délimite	*la recherche* et *Jean*
	et non pas	*à la recherche* et *de Jean*

Toutefois, plusieurs études ont suggéré qu'une entité présente dans un groupe prépositionnel avait moins de chance se servir d'antécédent à une chaîne de référence qu'un groupe nominal propre. Afin d'étudier ce type d'hypothèse, cette intégration dans un GP sera précisée lors de la phase d'annotation.

Attention – Dans le cas d'un déterminant contracté (*du* = *de la*, *au* = *à le*), cette dernière sera intégrée dans le groupe nominal délimité :

- *Je reviens du supermarché*	on délimite	*du supermarché*
	et non pas	*supermarché*
- *Je vais au Panthéon*	on délimite	*au Panthéon*
	et non pas	*Panthéon*

Notons que cette règle reste *a fortiori* valide si le déterminant contracté n'introduit pas un groupe prépositionnel mais est utilisé comme article partitif, comme dans *je mange du pain*.

Remarque – Les entités « *bonjour* » et « *merci* » sont exclues de l'annotation. En effet, la première est une formule ritualisée de salutation, la seconde est une interjection de remerciement. Il convient cependant de distinguer ces usages figés des usages réels tels que « *Tu as le bonjour d'Alfred* » ou encore « *Il est à sa merci* ». Dans ce type de cas, « *bonjour* » et « *merci* » sont à annoter.

2.2 Groupe nominaux récursifs : complément du nom

Il convient de faire attention aux groupes nominaux récursifs comportant des groupes prépositionnels imbriqués, comme par exemple *le président de l'université de Tours*.

Ici, il faut définir trois groupes imbriqués, ce que permet GLOZZ : [*le président de l'université de Tours*] [*l'université de Tours*] [*Tours*]. Chaque élément est en effet susceptible d'amorcer une chaîne anaphorique comme le montre cet exemple :

Le président de l'université de Tours est désormais Loïc Vaillant. Ce dernier a déclaré qu'il était fier de prendre la responsabilité de cet établissement. Cette nouvelle a été chaleureusement accueillie par le maire de la ville qui, on le sait, soutenait fortement la candidature du nouvel élu.

Ce type de structure récursive se retrouve avec tout groupe nominal ne donnant pas nécessairement lieu à une entité nommée. Par cohérence, on suivra la même règle d'annotation dans ce cas. A savoir, si on prend l'exemple *la sœur du voisin de mon père*, on caractérisera trois unités :

- *père*
- *voisin de mon père*
- *sœur du voisin de mon père*

et non pas trois entités *père*, *voisin* et *sœur*.

Cas particulier des entités nommées à référents multiples – Dans certains cas, une entité nommée réfère à plusieurs éléments qui peuvent donner lieu globalement ou individuellement à une reprise anaphorique comme dans l'exemple suivant :

- *Pierre et Marie Curie furent de célèbres physiciens. Ils travaillèrent longtemps ensemble et Marie reçut deux fois le Prix Nobel.*

Dans ce cas, il est demandé de définir trois groupes : le groupe englobant *Pierre et Marie Curie,* puis *Pierre* et enfin *Marie Curie* ; *Curie* sera également annoté mais recevra la valeur NULL. L'annotateur devra ensuite relier l'entité *Pierre* à l'entité NULL *Curie* en utilisant la notion de schéma proposée par Glozz. Pour cela :

1) On définit tout d'abord les unités *Pierre*, *Marie Curie* et *Curie* (bouton ) puis
2) On crée un schéma (bouton ) et enfin
3) On crée un lien entre *Pierre* et *Curie* par insertion d'unités dans le schéma (bouton )

2.3 Groupes nominaux récursifs : propositions relatives

Les pronoms relatifs ont un rôle très prévisible d'un point de vue syntaxique et sont le plus souvent connectés à leur antécédent. On pourrait donc choisir de s'affranchir de leur annotation. Dans certains cas, le pronom relatif peut être toutefois éloigné :

- *Il a acheté une maison près de Kercado qui est totalement à rénover*
- *La maison bleue que j'aimais tant et dans laquelle j'ai passé les plus belles années de ma vie*

Dans ce cas, la caractérisation de la chaîne anaphorique n'est pas triviale sans l'annotation du pronom relatif. Par souci de cohérence, tous les pronoms relatifs seront donc considérés comme des entités. Sur l'exemple *J'ai une voiture qui est très rapide* on délimitera donc deux groupes nominaux :

- *une voiture* - *qui*

Et non pas un seul groupe nominal récursif : *une voiture qui est très rapide*.

2.4 Les pronoms

Les pronoms personnels, explétifs et relatifs recevront toujours la valeur NO pour l'attribut NEW. L'antécédent, en revanche, recevra la valeur YES (s'ils sont en première mention). Ces pronoms ont les mêmes valeurs attributives que leur antécédent, excepté pour les attributs NEW, DEF et GP, qui dépendent du contexte environnant le pronom.

Les pronoms des verbes pronominaux ne seront pas annotés :

- *Jeanne se douche, puis s'habille et ensuite elle se coiffe les cheveux.*

Nous délimitons seulement deux groupes nominaux : « *Jeanne* » et « *elle* ». Les pronoms des verbes pronominaux « *se* » et « *s'* » ne font pas partie de l'annotation.

Cas des pronoms explétifs – Les pronoms explétifs (« *il pleut* ; « *ça le fait* », par exemple) doivent être annotés car ceux-ci peuvent tromper les systèmes de résolution des anaphores. Ils recevront la valeur GEN pour l'attribut GEN_REF. Cependant, les « *il* » et « *vous* » explétifs des tournures « *s'il vous plaît* » et « *il y a* » ne seront pas annotés car tous deux sont dans une formule figée, donc peu référentielle.

Cas des pronoms avec ellipses en « en » – Dans les situations telles que « *J'en ai un* », « *Tu peux m'en donner deux* », seul le pronom « *en* » sera annoté. Les éléments « *un* » et « *deux* » ne sont donc pas considérés comme des pronoms et ne seront pas annotés, mais ils servent à qualifier, ici numériquement, le pronom. A l'inverse dans les situations telles que « *Tu peux m'en donner un autre* » ou « *Il m'en faut d'autres* », « *un autre* » et « *d'autres* » sont des pronoms à part entière et devront donc être annotés, au même titre que « *en* ». Si besoin et selon le contexte, c'est le pronom de type « *un autre* » qui prendra la valeur YES à l'attribut NEW (cf. ci-dessous) car nous le jugeons plus fort et autonome grammaticalement que « *en* ». Nous rappelons qu'en tant que pronom, « *en* », « *un autre* » et leurs déclinaisons recevront une relation de type pronominale (ou associative pronominale) ; cette relation pointera sur l'antécédent, c'est-à-dire la première mention.

Cas des pronoms avec ellipses autres que « en » - Ces situations sont de type *pronom+adjectif*. Le schéma d'annotation suivi ne nous permet pas d'annoter ces situations, par exemple :

(1) *- Vous avez encore des robes ?*

- Oui mais je n'ai plus que la jaune.

- Je vais prendre celle-ci.

(2) *- Vous avez un plan de Grenoble ?*

- Je n'en ai plus que des payants.

- Et bien je vais prendre ceux-là.

« *La jaune* » et « *des gratuits* » ne seront pas annotés en raison de l'ellipse du nom. Ce ne sont pas des GN complètement réalisés ni syntaxiquement des pronoms, contrairement au pronom « *en* » de l'exemple (2). En revanche, les pronoms « *celle-ci* » et « *ceux-là* » seront annotés car ils désignent une nouvelle entité du discours : respectivement nous pouvons les comprendre comme « la robe jaune » et « les plans payants ». Ils recevront donc la valeur YES à l'attribut NEW (cf.ci-dessous). Nous rappelons qu'en tant que pronom, « *en* », « *celle-ci* » et « *ceux-là* » recevront une relation de type pronominale (ou associative pronominale) ; cette relation pointera sur l'antécédent, c'est-à-dire la première mention.

Remarque – Certains pronoms pourront recevoir la valeur YES à l'attribut NEW lorsque ceux-ci font partie d'une relation associative :

- *Les journaux français sont tous nuls. L'un dit quelque chose et l'autre dit tout son contraire.*

Dans la mesure où les pronoms « *l'un* » et « *l'autre* » font partie du GN « *les journaux français* » (relation ensemble/élément) et qu'ils peuvent chacun donner lieu à une reprise anaphorique indépendante de l'antécédent « *les journaux français* », alors les pronoms « *l'un* » et « *l'autre* » devront être codés NEW_YES. La relation anaphorique sera de type ASSOC_PRONOM (cf infra).

2.5 Parole spontanée : chevauchements

Il peut arriver que les interlocuteurs se chevauchent. Dans ce cas, les conventions de transcriptions conduisent à une segmentation du dialogue en tours de parole quelque peu artificiels. Par exemple :

U1 *Oui alors je voudrais maintenant de la*
U2 *Oui*
U1 *margarine et des œufs*

On constate ici que le groupe nominal *la margarine* est artificiellement partagé entre deux tours de parole. Cette situation est modélisée en caractérisant deux entités (*la* d'une part et *margarine* d'autre part) reliées par un schéma comme dans le paragraphe 4.2.

Toutefois, pour que deux entités ne soient pas comptabilisées, la partie qui ne contient pas la tête lexicale du groupe (ici, le déterminant *la*) sera typée comme artefact (TYPE = NULL, cf infra) et ne recevra aucune caractérisation lors de l'annotation.

3 Annotation : propriétés des groupes nominaux et des pronoms

Nous décrivons ici le premier niveau d'annotation qui consiste à décrire les groupes nominaux, entités nommées et pronoms compris. Les groupes nominaux et pronominaux seront décrits par huit propriétés :

- TYPE catégorie morpho-syntaxique : nom (entité nommée comprise) ou pronom ;
- GENRE genre grammatical de l'entité (masculin ou féminin) ;
- NOMBRE nombre de l'entité (singulier ou pluriel) ;
- EN type d'entité nommée (toponyme, anthroponyme…) le cas échéant ;
- GEN_REF caractère générique ou spécifique de la référence de l'item considéré
- DEF groupe nominal défini, indéfini, démonstratif ou explétif ;
- GP inclusion du GN ou du pronom dans un groupe prépositionnel ;
- NEW nouvel élément du discours le cas échéant ;

Chaque propriété correspondra à l'affectation d'une valeur suivant le paradigme attribut-valeur. On décrit ici les différentes valeurs que peuvent prendre les attributs concernés et donnons des indices pour la détermination des valeurs à associer à chaque groupe nominal. Notons que les différentes valeurs admissibles seront directement disponibles sur l'interface GLOZZ, une fois la DTD chargée. Notons que la DTD qui a été définie permet même de spécialiser les valeurs proposées suivant le type de l'item considéré. C'est précisément avec cet attribut TYPE que nous allons commencer notre description détaillée des propriétés.

3.1 Type : TYPE

Type morpho-syntaxique de l'entité.

Valeur	Description	Exemple
N	GN avec ses attributs. Il peut donc s'agir d'un nom commun, d'un nom propre formant le cas échéants une entité nommée.	*un petit chat, la voiture bleue*, *le Président la République*, *le conseil général d'Indre-et-Loire*, *Renan Luce*.
P	Pronom	*le*, *lui*, *il*…
NULL	Artefact	
UNK	*Unknown* : l'annotateur n'a pu se décider sur la catégorie morpho-syntaxique de l'entité.	Cette valeur ne devrait jamais être utilisée sur cet attribut.

Cas particulier des entités nommées à référents multiples – Dans le cas des entités nommées à référents multiples comme dans *Pierre et Marie Curie*, nous avons vu précédemment que l'entité *Pierre Curie* était décrite par deux items reliés à l'aide d'un schéma. Dans ce cas, on portera les annotations suivantes sur les deux items :

- premier item *Pierre* : on annote comme doit l'être l'entité complète
- second item *Curie* : type NULL

Cela permet de faire porter l'accord en genre, par exemple, sur le prénom et non sur le patronyme qui est, dans cet exemple, commun aux deux chercheurs.

3.2 Genre de l'entité : GENRE

Détermine le genre de l'entité.

Valeur	Description	Exemple
YES	Accord en genre	*Le livre (…) il*
NO	Pas d'accord en genre	*Le cabriolet (…) cette voiture*
UNK	On ne peut se décider	*On …*

3.3 Nombre de l'entité : NOMBRE

Précise le nombre (singulier ou pluriel) de l'entité.

Valeur	Description	Exemple
YES	Accord en nombre	*Le livre (…) il*
NO	Pas d'accord en nombre	*Le Café de la Gare (…) ils sont tous sympas*
UNK	On ne peut se décider	*On …*

Trois situations ne permettent pas la caractérisation en genre et en nombre de l'entité :

- Le caractère explétif des pronoms *il* et *ça* (*il pleut* ; *ça va*) ne permet de leur donner ni un genre ni un nombre ; ils seront donc notés UNK ;
- Il en va de même dans les emplois indéfinis du pronom *cela* et ses dérivés ;
- Enfin, les noms de villes seront également notés UNK pour le genre et le nombre. En revanche, les noms de fleuves et de pays, par exemple, ont un genre et un nombre.

3.4 Type d'entité nommée : EN

Typage de l'entité nommée lorsque le groupe nominal joue un tel rôle. Rappelons qu'une entité nommée est une entité (potentiellement polylexicale) qui décrit un élément unique de l'univers du discours. Une partie de l'annotation proposée dans le projet ANCOR reprend la codification utilisée dans le cadre de la campagne d'évaluation ESTER2.

Valeur	Description	Exemple
NO	Ne correspond pas à une EN	*une voiture*
PERS	Classe ESTER « Personne »	Personne réelle ou fictive et animaux.
FONC	Classe ESTER « Fonction »	Fonction politique, militaire, administrative, religieuse…
LOC	Classe ESTER « Lieu »	Géonyme, région administrative, axe de circulation, adresse, construction humaine…
ORG	Classe ESTER « Organisation »	Organisation politique, éducative, commerciale, géo-administrative…
PROD	Classe ESTER « Production humaine »	Classe très vague : moyen de transport, œuvre artistique, film…
TIME	Classe ESTER « Date et Heure »	Date relative ou absolue, heure. Les durées sont dans la classe AMOUNT
AMOUNT	Classe ESTER « Montant »	Age, durée, température, longueur, aire, volume, poids, vitesse, valeur monétaire…
EVENT	Evènements	Exemple : *La fête nationale*
NULL	Artefact	Uniquement si TYPE = NULL
UNK	On ne peut se décider sur le type	

Il est essentiel de se référer au guide d'annotation de la campagne d'évaluation Ester 2, version 0.3 (avril 2009) pour bien saisir l'étendue des différentes classes définies par cet attribut.

3.5 Généricité du référent : GEN_REF

Permet de décrire si l'entité considérée dénote un référent générique ou spécifique.

Valeur	Description	Exemple
GENE	Référent générique	*L'homme* dans *L'homme est un loup pour l'homme.* *Une voiture* dans *Une voiture cela pollue toujours.*
SPEC	Spécifique	*L'homme* dans *L'homme a tourné au coin de la rue.* *Une voiture* dans *Une voiture arrive, écartez-vous.*
NULL	Artefact	Uniquement si TYPE = NULL
UNK	On ne peut se décider	

3.6 Définition : DEF

Permet de décrire le caractère défini ou non de l'item considéré.

Valeur	Description	Exemple
INDEF	Indéfini	*une voiture*, *il(s)/elle(s) ou cela/ça/ce/c'* non-explétifs comme dans *Les plats cela se met ici ils ne rentrent pas ailleurs* ; les pronoms *certains, on, tout*, *n'importe qui/quoi/quel*, *personne*, *rien*, *aucun(e)*, *d'aucun(e)s*, *nul(e)s*, *l'un(e)*, *l'autre*, *l'un(e) et l'autre*, *ni l'un(e) ni l'autre*, *pas un(e)*, *plus d'un(e)*, *plusieurs*, *quelqu'un(e)*, *quelque chose*, *autrui*, *autre chose*, *chacun(e)*, *tout un chacun*, *d'autres*…
EXPL	Explétif (non référentiel)	*il* (dans *il pleut*), ça (dans *ça va*)
DEF_DEM	Défini démonstratif	*cette voiture, celui-là*
DEF_SPLE	Défini non démonstratif	*la voiture, je, nous, il(s)/elle(s)* ou *cela/ça/ce/c'* non explétifs comme dans *la cargaison, le camion servira pour ça. Il est assez gros*.
NULL	Artefact	Uniquement si TYPE = NULL
UNK	On ne peut se décider	

3.7 Groupe nominal ou prépositionnel : GP

Permet de décrire si le groupe nominal considéré est intégré dans un groupe prépositionnel ou pas. On intégrera dans cette caractérisation les pronoms qui jouent le rôle de groupe prépositionnels.

Valeur	Description	Exemple
YES	Groupe nominal dans un GP	Il est rentré *dans la voiture,* il *lui* donne un jouet, il *nous* parle.
NO	Groupe nominal	Il regarde *la télévision*, il *le* donne à Jean
NULL	Artefact	Uniquement si TYPE = NULL
UNK	*O*n ne peut se décider.	A ne jamais utiliser a priori sur cet attribut.

3.8 Nouvel élément du discours : NEW

Cet attribut précise si l'élément annoté introduit ou non un nouvel élément dans le discours.

Valeur	Description	Exemple
YES	Nouvel élément du discours	
NO	Élément introduit précédemment	
NULL	Artefact	Uniquement si TYPE = NULL
UNK	*O*n ne peut se décider.	A ne jamais utiliser a priori sur cet attribut.

Cas des cataphores – Les cataphores sont un type particulier d'anaphore pour lesquelles le pronom survient avant « l'antécédent » :

- *Elle aime bien son jardin, Jeanne.*
  NEW= NO NEW = YES

Dans ce cas, c'est toujours l'antécédent, et non pas le pronom, qui sera étiqueté NEW = YES. En effet, il faut attendre le nom propre *Jeanne* pour identifier de manière univoque le référent concerné.

4 Annotation : détermination des relations anaphoriques

Les chaînes anaphoriques seront annotées par paires de relations entre une reprise anaphorique et la première mention de l'élément du discours concerné. Par exemple, si un nouvel élément du discours est introduit par l'élément A, puis est repris par l'élément B et l'élément C, on annotera deux relations anaphoriques : la relation (B → A) puis la relation (C → A). Les relations d'une même chaîne anaphorique pointeront donc toutes vers la première mention du référent concerné : celui-ci doit toujours porter la mention NEW = YES.

Ce choix d'annotation a été préféré à une description suivant le fil du discours : (A → B) et (B → C).

Sens de la relation – La relation partira de la reprise pour pointer sur l'antécédent (donc l'élément NEW = YES). Cette convention est applicable également dans le cas des cataphores.

Anaphore *Jeanne aime bien son jardin. Elle y passe tous ces après-midi.*
NEW= YES ← NEW = NO

Cataphore *Elle aime vraiment bien son jardin Jeanne.*
NEW= NO → NEW = YES

Cas des mentions des interlocuteurs – Tous les pronoms personnels sont à annoter. Cependant, nous n'annotons pas les relations de coréférence entre entités nominales (les « *monsieur* » ou « *madame* », ou les noms des interlocuteurs) et/ou pronominales (*je*/*tu* et *nous*/*vous*) faisant référence aux personnes de l'interlocution. En effet, il y a coréférence, mais pas anaphore car nous n'avons pas besoin du contexte pour résoudre la référence :

- *Bonjour Juliette, je t'appelle pour le dossier numéro 33*
- *Merci Jean-Marc, je suis avec un client. Je te rappelle dès que j'aurai terminé avec lui.*

Ici, le pronom « *Juliette* » et ses anaphores pronominales (en vert) seront annotés mais jamais reliés à l'antécédent « *Juliette* ». Il en est de même pour le pronom « *Jean-Marc* » et ses anaphores pronominales (en violet). En revanche, le « *client* », dont il est fait référence, et le pronom « *lui* », doivent être annotés et reliés car il s'agit là d'un référent non-présent dans la situation d'énonciation.

Dans une situation où deux personnes discutent entre elles à propos d'une tierce personne, puis cette dernière entre dans la communication, tant que cette tierce personne n'entre pas dans la discussion, nous n'annotons pas les relations de coréférence la concernant. Considérons l'exemple suivant :

- Personne1 : *J'ai un client pour toi qui a un problème avec son inscription*
- Personne2 *: Et bien passe-le-moi, je vais voir ça avec lui*
- Personne3 *: Bonjour, je suis la personne qui a un problème d'inscription*

Nous rappelons que tous les pronoms personnels doivent être annotés. Cependant, dans les deux premiers tours de parole (ceux des Personne1 et 2), la Personne3 n'est pas présente dans la discussion. En conséquence, les pronoms personnels qui font référence à lui (« *qui* », « *le* » et « *lui* ») devront être reliés à leur antécédent « *un client* ». Par contre, dès lorsque la Personne3 entre dans la discussion, les pronoms ne seront plus reliés à l'antécédent.

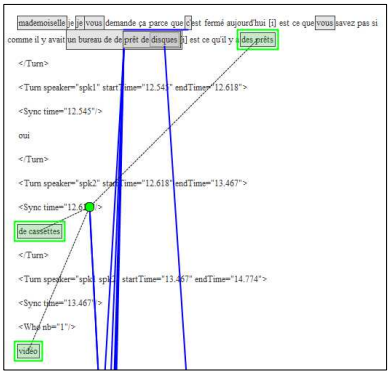
Enfin, les pronoms personnels collectifs « *on* » et « *nous* » qui comprennent le(s) locuteur(s) et d'autres locuteurs présents ou non dans la situation, seront annotés mais jamais reliés.

Cas des pronoms relatifs – Les pronoms relatifs seront reliés anaphoriquement à leur antécédent grammatical direct :

- *Alors comme ça, vous habitez à Orléans. C'est une ville charmante que vous habitez là.*
- *Oui, je l'aime beaucoup.*

Dans cet exemple, le pronom relatif « *que* » sera relié non pas à l'antécédent « *Orléans* », mais à son antécédent grammatical « *une ville charmante* » (même si celui-ci a reçu la valeur NO pour l'attribut NEW). En revanche, « *une ville charmante* » et les anaphores situées après le pronom relatif (ici, « *l'* ») seront reliées à l'antécédent de la chaîne, « *Orléans* ».

Cas des pronoms déictiques – Ils sont à annoter, mais pas à relier dans une relation anaphorique. En effet, résoudre la référence nous demanderait de pouvoir voir la scène, par exemple, voir ce que le locuteur montre sans le nommer.



Cas des relations impliquant un schéma – Lorsqu'une reprise anaphorique implique l'ensemble des entités d'un schéma (et non une plus petite entité imbriquée dans cet ensemble), la relation doit partir ou arriver, selon le sens de la relation, à partir du schéma lui-même et non à partir de la première unité du schéma. La relation doit donc s'ancrer sur le point liant les différentes entités du schéma. Sur la capture d'écran ci-contre, le point vert symbolise l'ensemble du schéma, soit l'entité « *des prêts de cassettes vidéo* ».

Anaphores et autres relations – On n'annotera pas les relations qui ne relèvent pas de l'anaphore mais de relations de prédication, d'attribution ou d'apposition. Seront ainsi ignorées les relations suivantes :

- Jean est un grand homme prédication/attribution
- Jean est l'homme de la situation prédication/attribution
- Jean, homme de foi, est … apposition
- Jean, le directeur de l'usine, est … apposition

5 Annotation : propriétés des relations anaphoriques

Chaque relation anaphorique sera décrite par quatre propriétés :

- TYPE Type d'anaphore : directe, indirecte, associative, pronominale et pronominale associative
- GENRE Respect de l'accord en genre par la relation anaphorique
- NOMBRE Respect de l'accord en nombre par la relation anaphorique
- ID_LOC Reprise faite par le locuteur, ou non

5.1 Type de relation : TYPE

Permet de décrire le type de relations anaphorique entre les deux éléments concernés.

Valeur	Description	Exemple
DIRECTE	Reprise par une expression de même tête	Les deux groupes nominaux réfèrent au même élément du discours. L'antécédent est un nom qui a la même tête que la reprise anaphorique. *La voiture rouge* (…) *Cette belle voiture*
INDIRECTE	Reprise par une autre expression	Les deux groupes nominaux réfèrent au même élément du discours mais n'ont pas la même tête. En particulier, les deux têtes peuvent partager une relation de synonymie, d'hyponymie/hyperonymie. *Le cabriolet (…) cette décapotable (…) la voiture* *Il vend des voitures* (…) *Ce vendeur…*
ANAPHORE	Reprise anaphorique (pronom)	On classe dans cette catégorie les reprises anaphoriques : *la voiture* (…) *elle*.

Valeur	Description	Exemple
ASSOC	Anaphore associative nominale	Cas des *bridging anaphora* : les deux groupes nominaux ne réfèrent pas au même élément du discours, mais ces deux éléments partagent une relation ontologique certaine. Par exemple : Méronymie : *j'ai rejoint la voiture* (…) *la porte était fermée à clef.*
ASSOC_PRONOM	Anaphore associative pronominale	Cas des *bridging anaphora* mais entre un GN et un pronom reprenant tout ou en partie le GN Exemple : *Les factures sont rangées dans les classeurs sur l'étagère. Tu trouveras donc les factures d'électricité dans l'un d'entre eux.*

Les paragraphes ci-dessous décrivent un peu plus précisément ces différents types.

Reprise directe – Lorsque l'élément anaphorique reprend une entité déjà présente dans le discours (ici : le texte), et que l'auteur a utilisé auparavant la même expression (ou du moins de même tête syntaxique), vous la classez comme « reprise directe » en cliquant sur la case à cocher correspondante dans l'interface.

- *La voiture jaune……. La voiture*

Bien entendu, cela ne vaut pas pour : *La voiture rouge…… La voiture jaune* car il n'y a pas coréférence.

Reprise par autre expression – Lorsque l'élément anaphorique reprend une entité déjà présente dans le discours (ici : le texte), et que l'auteur a utilisé auparavant une autre expression (synonyme, description, ...), vous classez la relation comme « reprise indirecte ».

- *le livre… l'ouvrage* (synonymie)
- *Napoléon... l'Empereur* (hyperonymie)
- *Le cabriolet... la voiture* (hyperonymie)

Parfois, la relation anaphorique n'est détectable que par le verbe dont l'antécédent est le sujet voire plus rarement l'objet :

- *L'homme achète le livre (…) l'acheteur repart* (sujet)
- *L'homme achète le livre (...) il repart avec cet achat* (objet)

Reprise anaphorique – Cas simple d'anaphore où la reprise est faite par un pronom.

- *L'homme achète le livre (…) ce dernier est en tête des ventes*
- *Il regarde Carla (...) elle sourit un peu niaisement*

Reprise associative – Si l'entité désignée par l'élément anaphorique n'était pas mentionnée auparavant dans le texte mais que son interprétation est basée sur, dépendante de ou reliée à une autre entité mentionnée par une expression nominale dans le texte, vous classerez la relation comme « reprise associative ».

- *Monsieur R. voulait acheter un appartement, mais le prix était trop élevé.*

Plusieurs situations peuvent se rencontrer dont nous donnons quelques exemples afin de caractériser les liens parfois ténus qui peuvent exister au sein de la chaîne dans ce cas :

- *Le voleur s'approcha de la maison (...) la porte était fermée* (meronyme)
- *Il approcha du village (...) il pouvait déjà voir le clocher* (localisation)
- *Je n'aime pas le Café Central (...) ils sont tous guindés* (metonymie)
- *Toute la famille était réunie (...) Le père l'accueillit* (élément)
- *Le TFC est en tête de la Ligue 1 (...) Son président exulte* (fonction)
- *J'adore ce foulard (...) La soie est si douce exulte* (matériau)
- *La guerre durait depuis 4 ans (...) Sa fin était proche* (temporel)
- *Le concert était génial (...) J'ai adoré le chanteur* (theta)
- *La pâte est prête (...) La farine vient juste d'être mise* (theta)
- *Le disque se vend bien (...) Ses acheteurs se comptent par centaines*

La relation associative doit également être utilisée dans le cadre d'une relation `ensemble/élément`. Considérons l'exemple ci-dessous :

- *Les enfants sont dans la cour (...) Jean joue au ballon tandis que Paul lit. Paul a toujours été le moins turbulent.*

Ici, nous sommes en présence de deux relations associatives de type `ensemble/élément` entre *Jean* et *les enfants* d'une part, et *Paul* et *les enfants* d'autre part. Nous avons donc deux chaînes anaphoriques (puisque leurs référents sont distincts) qui pointent sur le même élément NEW. Par ailleurs, nous sommes en présence d'une reprise directe entre les deux mentions de *Paul*. On observe que la première mention de *Paul* est donc à la fois reprise dans une relation associative, mais également élément NEW d'une autre relation.

Par ailleurs, nous ne considérerons pas comme anaphore associative les cas où la relation entre les deux entités référentes peut être directement identifiée par la syntaxe. Ainsi, nous ne considérerons pas qu'il y a anaphore associative dans les relations de possession ou meronymie (relation partie_de) suivantes :

- Possession — *La voiture de Jean est belle*
- Meronymie — La portière de la voiture est abimée

Alors que nous marquerons la relation dans le cas suivant : *La voiture est abimée. La portière est cabossée*. Enfin, dans certains cas, peu fréquents, une relation associative peut concerner un pronom qui reprend en totalité ou seulement une partie du GN (pronominale associative) :

- Partie — Les journaux français sont en difficulté ; pendant que l'un restructure, l'autre licencie à tout va.
- Tout — Je cherche l'agence *France Télécom* la plus proche. Je crois qu'il y en a une rue Nationale.

Dans tous les cas de figure, on a cette règle : *dans le cas d'une reprise associative, les deux éléments de la relation portent la valeur YES pour l'attribut NEW (y compris dans le cas d'une associative pronominale).*

5.2 Accord en genre : GENRE

Précise si l'anaphore se traduit ou non par un respect du genre entre l'antécédent et la reprise.

Valeur	Description	Exemple
YES	Accord en genre	*Le livre (…) il*
NO	Pas d'accord en genre	*Le cabriolet (…) cette voiture*
UNK	*Unknown* : information non renseignée	*Cet attribut doit être utilisé systématiquement dans le projet ANCOR. Il sera positionné automatiquement par GLOZZ*

5.3 Accord en nombre : NOMBRE

Précise si l'anaphore se traduit ou non par un respect du nombre entre l'antécédent et la reprise.

Valeur	Description	Exemple
YES	Accord en nombre	*Le livre (…) il*
NO	Pas d'accord en nombre	*Le Café de la Gare (…) ils sont tous sympas*
UNK	*Unknown* : information non renseignée	*Cet attribut doit être utilisé systématiquement dans le projet ANCOR. Il sera positionné automatiquement par GLOZZ*

5.4 Reprise d'un autre interlocuteur : ID_LOC

Cet attribut est optionnel et n'est utilisé que pour les annotations de corpus de parole conversationnelle fortement interactive. Plus précisément :

- si l'annotation suit la `DTD DTD_GLOZZ_ANCOR_DEFAULT.aam`, cet attribut n'a pas été considéré,
- si l'annotation suit la `DTD DTD_GLOZZ_ANCOR_DIALOGUE.aam`, l'attribut a été considéré,

Cet attribut précise si la reprise considérée est le fait du locuteur qui a fait la première mention au référent au cours du dialogue, ou si au contraire il s'agit d'un autre interlocuteur.

Valeur	Description	Exemple
YES	Même locuteur	SP1 : *Tiens j'ai terminé un livre super hier* SP2 : *Ah bon* SP1 : *Oui il traitait de la coréférence.*
NO	Locuteur différent	SP1 : *Tiens j'ai terminé un livre super hier* SP2 : *Ah bon* SP1 : *Oui j'ai vraiment adoré* SP2 : *Et il causait de quoi*
UNK	On ne sait se décider.	

6 Compléments : FAQ sur les annotations

Cette Foire Aux Questions tente de répondre aux interrogations que pourraient se poser les experts annotateurs face aux données qu'ils auront à traiter.

Explétif ou référentiel : pronom dans les tournures du type *c'est / cela / il y a*

Seront annotés, les pronoms *c'est*/*cela*/*il y a* dans les situations suivantes :

- Le pronom *ça*/*ce*/*cela* joue avant tout un rôle de remplissage syntaxique, néanmoins **lorsqu'il prête à référence**, il sera annoté.

- Le pronom *il* sera annoté, par exemple, dans les **usages explétifs** tels que « *il pleut* », qui ne sont pas automatiquement caractérisables par une forme bien identifiable. Les pronoms explétifs sont considérés comme génériques, ils ne correspondent à aucune classe d'entités nommées (EN_NO) et reçoivent la valeur UNK pour le genre et le nombre.

Ne seront pas annotés, les pronoms *c'est*/*cela*/*il y a* dans les situations suivantes :

- Les **présentatifs** *c'*/*ça*/*cela* ne seront pas annotés, par exemple dans les structures clivées « *C'est Jean qui partira en premier* » (où seuls « *Jean* » et le pronom relatif « *qui* » seront annotés) et dans les pseudo-clivées « *Celui qui partira en premier, c'est Jean* » (où « *celui* », « *qui* » et « *Jean* » seront annotés).

- Les occurrences des pronoms « *il* » et « *vous* » des expressions « ***il y a*** » et « ***s'il vous plait*** » ne seront pas annotés, ainsi que les tournures « *ce que* » et « *ce dont* » non-référentiels dans des tournures pseudo-clivées du type « ***ce que*** *je veux dire* » ou « ***ce dont*** *on a parlé* ».

Disfluences orales : cas des groupes nominaux avec reprise

L'oral spontané se caractérise par une forte fréquence de structures disfluentes, parmi lesquelles se trouvent de multiples formes de reprises ou répétitions. Celles-ci peuvent concerner des entités nommées et plus généralement des groupes nominaux. Considérons les deux exemples suivants :

- *J'ai acheté une voiture enfin une brouette plutôt vu son état*
- *C'est la fille du docteur du comment déjà de l'ostéopathe*

Le premier cas ne pose pas de problème à l'annotation : on est en présence de deux groupes nominaux et on délimitera séparément *voiture* et *brouette*.

Le second cas est plus délicat, puisque la reprise concerne un complément du nom, donc un groupe prépositionnel intégré dans un groupe nominal. Dans ce cas, on définira trois entités :

- *docteur*
- *ostéopathe*
- *fille du docteur du comment déjà de l'ostéopathe*

Répétitions

Nous distinguons trois cas de figure, communs aux GN et aux pronoms :

- Dans le cas des répétitions directes (sans élément interposé), seul le dernier élément de la répétition sera annoté :

(1) J'habite à Orléans Orléans Orléans dans le Loiret / J' j' j'habite à Orléans dans le Loiret

- Dans le cas des répétitions indirectes (avec élément interposé), tous les éléments seront annotés :

(2) J'habite à Orléans à Orléans à Orléans dans le Loiret / J'ai j'ai j'ai un appartement à Orléans dans le Loiret

- Dans le cas des répétitions directes mais dont la tête nominale ou pronominale n'est pas identique, tous les éléments seront annotés, il s'agit alors d'une anaphore de type indirecte (cf. 7.1. Type de relation : Type) :

(3) Regarde le bateau le grand bateau le navire plutôt / Moi j'habite à Orléans

Dans ce dernier exemple, « *le bateau le grand bateau* » est une répétition, la dernière occurrence est donc la seule à être prise en compte. Il y a un changement de tête nominale entre « *le grand bateau* » et « *le navire* » : les deux s'annotent et feront l'objet d'une relation anaphorique.

Structure en « *c'est X qui s'appelle Y* »

Cette structure pose problème quant au choix de l'antécédent : s'agit-il d'une anaphore, auquel cas *Y* est l'anaphore de *X*, ou s'agit-il d'une cataphore, auquel cas *X* est l'anaphore de *Y* ? Nous avons choisi cette seconde possibilité car notre corpus nous montre que c'est cette entité Y, mieux définie par un nom propre par exemple, qui semble reprise dans la suite du texte :

- *Il y avait une association qui s'appelait Pop Ski qui existait et qui se réunissant par ici*

- *C'était un club ?*

- *Je ne sais pas trop, c'était une association*

Les mentions présentes dans les deux derniers tours de parole nous semblent plutôt faire référence à « *Pop Ski* » qu'à la mention floue « *une association* », c'est pourquoi nous avons choisi « *Pop Ski* » comme antécédent.

ANNEXE B — DTD GLOZZ du corpus ANCOR

```
<?xml version="1.0" encoding="UTF-8"?>

<annotationModel>
      <units>
            <type name="N">
                  <featureSet>
                        <feature name="GENRE">
                              <possibleValues default="M">
                                    <value>M</value>
                                    <value>F</value>
                                    <value>UNK</value>
                              </possibleValues>
                        </feature>
                        <feature name="NB">
                              <possibleValues default="SG">
                                    <value>SG</value>
                                    <value>PL</value>
                                    <value>UNK</value>
                              </possibleValues>
                        </feature>
                        <feature name="EN">
                              <possibleValues default="PERS">
                                    <value>LOC</value>
                                    <value>PERS</value>
                                    <value>FONC</value>
                                    <value>ORG</value>
                                    <value>AMOUNT</value>
                                    <value>TIME</value>
                                    <value>PROD</value>
                                    <value>EVENT</value>
                                    <value>NO</value>
                                    <value>UNK</value>
                              </possibleValues>
                        </feature>
                        <feature name="DEF">
                              <possibleValues default="DEF_SPLE">
                                    <value>INDEF</value>
                                    <value>DEF_SPLE</value>
                                    <value>DEF_DEM</value>
                                    <value>UNK</value>
                              </possibleValues>
                        </feature>
                        <feature name="GEN_REF">
                              <possibleValues default="SPEC">
                                    <value>GENE</value>
                                    <value>SPEC</value>
                                    <value>NULL</value>
                                    <value>UNK</value>
                              </possibleValues>
                        </feature>
                        <feature name="GP">
                              <possibleValues default="NO">
                                    <value>YES</value>
                                    <value>NO</value>
                                    <value>UNK</value>
                              </possibleValues>
                        </feature>
```

```
            <feature name="NEW">
                  <possibleValues default="NO">
                        <value>YES</value>
                        <value>NO</value>
                        <value>UNK</value>
                  </possibleValues>
            </feature>
      </featureSet>
</type>
<type name="PR">
      <featureSet>
            <feature name="GENRE">
                  <possibleValues default="M">
                        <value>M</value>
                        <value>F</value>
                        <value>UNK</value>
                  </possibleValues>
            </feature>
            <feature name="NB">
                  <possibleValues default="SG">
                        <value>SG</value>
                        <value>PL</value>
                        <value>UNK</value>
                  </possibleValues>
            </feature>
            <feature name="EN">
                  <possibleValues default="PERS">
                        <value>LOC</value>
                        <value>PERS</value>
                        <value>FONC</value>
                        <value>ORG</value>
                        <value>AMOUNT</value>
                        <value>TIME</value>
                        <value>PROD</value>
                        <value>EVENT</value>
                        <value>NO</value>
                        <value>UNK</value>
                  </possibleValues>
            </feature>
            <feature name="DEF">
                  <possibleValues default="DEF_SPLE">
                        <value>EXPL</value>
                        <value>INDEF</value>
                        <value>DEF_SPLE</value>
                        <value>DEF_DEM</value>
                        <value>UNK</value>
                  </possibleValues>
            </feature>
            <feature name="GEN_REF">
                  <possibleValues default="SPEC">
                        <value>GENE</value>
                        <value>SPEC</value>
                        <value>NULL</value>
                        <value>UNK</value>
                  </possibleValues>
            </feature>
            <feature name="GP">
                  <possibleValues default="NO">
                        <value>YES</value>
                        <value>NO</value>
                        <value>UNK</value>
                  </possibleValues>
            </feature>
```

```
                  <feature name="NEW">
                        <possibleValues default="NO">
                              <value>YES</value>
                              <value>NO</value>
                              <value>UNK</value>
                        </possibleValues>
                  </feature>
            </featureSet>
      </type>
      <type name="NULL">
            <featureSet>
                  <feature name="GENRE">
                        <possibleValues default="NULL">
                              <value>NULL</value>
                  </possibleValues>
                  </feature>
                  <feature name="NB">
                        <possibleValues default="NULL">
                              <value>NULL</value>
                  </feature>
                  <feature name="EN">
                        <possibleValues default="NULL">
                              <value>NULL</value>
                  </possibleValues>
                  </feature>
                  <feature name="DEF">
                        <possibleValues default="NULL">
                              <value>NULL</value>
                        </possibleValues>
                  </feature>
                  <feature name="GEN_REF">
                        <possibleValues default="NULL">
                              <value>NULL</value>
                        </possibleValues>
                  </feature>
                  <feature name="GP">
                        <possibleValues default="NO">
                              <value>YES</value>
                              <value>NO</value>
                              <value>UNK</value>
                        </possibleValues>
                  </feature>
                  <feature name="NEW">
                        <possibleValues default="NULL">
                              <value>NULL</value>
                        </possibleValues>
                  </feature>
            </featureSet>
      </type>
</units>

<relations>
      <type name="DIRECTE">
            <featureSet>
                  <feature name="GENRE">
                        <possibleValues default="UNK">
                              <value>YES</value>
                              <value>NO</value>
                              <value>UNK</value>
                        </possibleValues>
                  </feature>
                  <feature name="NOMBRE">
                        <possibleValues default="UNK">
```

```
                        <value>YES</value>
                        <value>NO</value>
                        <value>UNK</value>
                  </possibleValues>
            </feature>
      </featureSet>
</type>
<type name="INDIRECTE">
      <featureSet>
            <feature name="GENRE">
                  <possibleValues default="UNK">
                        <value>YES</value>
                        <value>NO</value>
                        <value>UNK</value>
                  </possibleValues>
            </feature>
            <feature name="NOMBRE">
                  <possibleValues default="UNK">
                        <value>YES</value>
                        <value>NO</value>
                        <value>UNK</value>
                  </possibleValues>
            </feature>
      </featureSet>
</type>
<type name="ANAPHORE">
      <featureSet>
            <feature name="GENRE">
                  <possibleValues default="UNK">
                        <value>YES</value>
                        <value>NO</value>
                        <value>UNK</value>
                  </possibleValues>
            </feature>
            <feature name="NOMBRE">
                  <possibleValues default="UNK">
                        <value>YES</value>
                        <value>NO</value>
                        <value>UNK</value>
                  </possibleValues>
            </feature>
      </featureSet>
</type>
<type name="ASSOC">
      <featureSet>
            <feature name="GENRE">
                  <possibleValues default="UNK">
                        <value>YES</value>
                        <value>NO</value>
                        <value>UNK</value>
                  </possibleValues>
            </feature>
            <feature name="NOMBRE">
                  <possibleValues default="UNK">
                        <value>YES</value>
                        <value>NO</value>
                        <value>UNK</value>
                  </possibleValues>
            </feature>
      </featureSet>
</type>
<type name="ASSOC_PRONOM">
      <featureSet>
```

```
                        <feature name="GENRE">
                              <possibleValues default="UNK">
                                    <value>YES</value>
                                    <value>NO</value>
                                    <value>UNK</value>
                              </possibleValues>
                        </feature>
                        <feature name="NOMBRE">
                              <possibleValues default="UNK">
                                    <value>YES</value>
                                    <value>NO</value>
                                    <value>UNK</value>
                              </possibleValues>
                        </feature>
                  </featureSet>
            </type>
      </relations>

</annotationModel>
```

ANNEXE C — Passage de l'annotation GLOZZ à l'annotation ancrée

Modifications du format originel du corpus ANCOR (Glozz)

Version *1.1*
Auteurs *Adèle Désoyer*
Date *15 septembre 2014*

1) Concaténation des fichiers source (*.ac) et annotations (*.aa) dans un nouveau fichier de la forme

<ANCOR>
<Trans>
Contenu du fichier *.ac
</Trans>
<annotations>
Contenu du fichier *.aa
</annotations>
</ANCOR>

2) Suppression d'éléments insérés automatiquement par Glozz, tels que :

a) Les unités *<paragraph>* créées à chaque saut de ligne (notre corpus étant un fichier xml de transcription oral, les sauts de lignes ne correspondent pas à des paragraphes).

b) Les balises *<metadata>* des éléments *<unit>, <relation> et <schema>* qui ne contiennent aucune information linguistique.

3) **Insertion d'ancres pour repérer les unités coréférentes directement** dans le corpus. On conserve en attribut l'identifiant unique de l'élément <*unit*> correspondant, afin d'accéder à ses métadonnées.

4) **Ajout d'un identifiant numérique pour les éléments *<Section>, <Turn> et <anchor>*** utile au calcul des distances entre deux unités (cf. point 6).

5) **Ajout d'informations additionnelles sur les unités en tant qu'élément *<feature>* à l'aide de nouveaux attributs**[i]

a) Contenu de l'expression (attribut *CONTENT*)

b) Locuteur (attribut *SPEAKER*)
c) Token précédant la mention (attribut *PREVIOUS*)
d) Token suivant la mention (attribut *NEXT*)

NB : Lorsque l'unité annotée débute un nouveau tour de parole, le token précédant est représenté par le symbole ^ ; Lorsqu'elle termine un tour de parole, le token suivant est représenté par le symbole *$*.

6) Ajout d'informations additionnelles sur les relations en tant qu'élément *<feature>* à l'aide de nouveaux attributs

a) En nombre de tour de parole (attribut *Distance_turn*)
b) En nombre de mention (attribut *Distance_mention*)
c) En nombre de caractères (attribut *Distance_char*)
d) En nombre de mots (attribut *Distance_word*)

[i] Les illustrations des modifications sont extraites du fichier *1AG0141.xml* du sous corpus *Corpus_OTG*