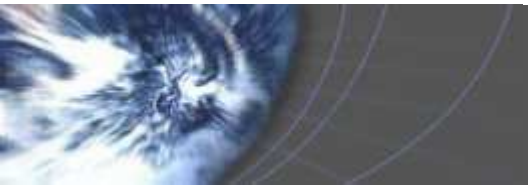
Corpus CO2

Présentation générale

Jean-Yves Antoine[1], Judith Muzerelle[2], Aurore Pelletier[1], Emmanuel Schang[2]

[1]Université François Rabelais de Tours
[2]Université d'Orléans

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



Introduction

Ce document décrit le corpus CO2, un corpus de français parlé spontané annoté en coréférence dans le cadre du projet CO2 financé par le PRES Centre-Val de Loire. Ce corpus a été réalisé par le laboratoire LI de l'université François Rabelais de Tours et le laboratoire LLL-CNRS des universités d'Orléans et de Tours. L'annotation du corpus en coréférence est été menée sur le corpus ESLO (Enquête SocioLinguistique d'Orléans) constitué par le laboratoire LLL et diffusé sous licence Creative Commons CC-BY-NC-SA [Baude et Dugua 2011, Eshkol-Taravella et al. 2012].

De même, le corpus CO2 est diffusé librement sous licence Creative Commons CC-BY-NC-SA par le laboratoire LI. Il est récupérable sur le site WWW du programme PAROLE_PUBLIQUE[1].

Plus précisément, ce rapport présente :

- le contenu du corpus distribué ainsi que le corpus ESLO sur lequel a porté l'annotation,
- les modes de distributions du corpus,
- la convention à laquelle est liée l'utilisation de ce corpus à toutes fins scientifiques ou industrielles,
- les références bibliographiques associées à ce corpus.
- les conventions d'annotation suivies lors de la réalisation du corpus,

1. Présentation du corpus : contenu

Le corpus CO2 correspond à l'annotation en co-référence de trois dialogues oraux (interviews) extraits du corpus ESLO (Enquête SocioLinguistique d'Orléans) du laboratoire LLL d'Orléans. L'annotation concerne aussi bien les relations de coréférences que les anaphores associatives, mais se limite uniquement aux relations concernant des groupes nominaux ou pronominaux. Il s'agit d'une annotation fine, au sens où le corpus est enrichi par une caractérisation linguistique des mentions concernées par la relation ainsi que de la relation elle-même (cf. infra).

L'annotation est déportée, c'est-à-dire qu'elle donne lieu à un fichier séparé du corpus originel mais synchronisé avec celui-ci. Elle a été réalisée avec le logiciel d'annotation GLOZZ [REF]. Le corpus distribué comprend ainsi la transcription au format interne GLOZZ ainsi que l'annotation dans deux fichiers séparés. Les fichiers audio source ainsi que la transcription orthographique d'origine du corpus ESLO ne font pas l'objet de cette distribution : les personnes intéressées par ces données se référeront à la distribution du corpus ESLO [Baude et Dugua 2011, Eshkol-Taravella et al. 2012].

1.1 Fiche signalétique

Corpus	CO2
Version	1.0 (juin 2013)
Type de dialogue	Dialogue oral peu interactif (interview socio-linguistique)
Locuteurs	Adultes hommes ou femmes francophones
Enregistrement	Voir distribution corpus ESLO
Contenu	Transcription orthographique + annotation en coréférence
Superviseurs	Jean-Yves Antoine (LI, Université de Tours), Emmanuel Schang (LLL, U. Orléans)
Annotateurs	Judith Muzerelle (LLL, U. Tours) et Aurore Pelletier (LLL, U. Tours)
Diffusion	libre sous réserve du respect de la licence Creative Commons CC-BY-NC-SA

1.2 Corpus source : ESLO

La constitution du corpus CO2 a consisté à annoter 3 fichiers extraits de la distribution du corpus ESLO [Baude et Dugua 2011, Eshkol-Taravella et al. 2012]. ESLO, *Enquête Sociolinguistique à Orléans*, est un projet du laboratoire LLL (Laboratoire Ligérien de Linguistique de l'université d'Orléans. Il a pour objectif de constituer un corpus oral (collection ordonnée d'enregistrements de la parole) qui soit prototype à toutes les étapes de sa réalisation et qui puisse se situer au même niveau, qualitatif et quantitatif, y compris par sa dimension patrimoniale, que les grands corpus oraux fabriqués, ou en cours de fabrication, en Europe et dans le monde. Il se compose de deux sous-corpus : ESLO1 et ESLO 2.

ESLO 1 est la suite d'une initiative prise en 1968. A cette date, un groupe d'universitaires anglais avait entrepris de collecter des documents sonores à Orléans avec une visée didactique : l'enseignement du

[1] http://www.info.blois.univ-tours.fr/~antoine/parole_publique

français langue étrangère dans le système public d'éducation anglais. L'enquête comprend environ 200 interviews avec les propriétés sociolinguistiques des locuteurs et des situations, soit au total plus de 300 heures de parole incluant pour moitié des interviews en face à face et pour moitié une gamme d'enregistrements variés (conversations téléphoniques, réunions publiques, transactions commerciales, repas de famille, entretiens médico-pédagogiques, etc.). L'exploitation de ces matériaux se poursuit au sein du laboratoire LLL qui diffuse désormais le corpus sous licence Creative Commons CC-BY-NC-SA.

Le corpus CO2 qui est diffusé résulte de l'annotation de 3 interviews du corpus ESLO1. Au total, l'extrait annoté du corpus représente 3 heures 28 d'enregistrement et représente 35 192 mots.

1.3 Corpus annoté en coréférence : CO2

Le corpus CO2 correspond au corpus diffusé ici.

Procédure d'annotation – L'annotation a été réalisée sur le logiciel *Glozz* (Mathet et Widlöcher, 2009) et le corpus distribué suit le format d'annotation de ce logiciel qui est distribué librement. *Glozz* produit une annotation au format XML reposant sur une DTD que nous avons adaptée à notre schéma d'annotation (cf. infra). Les annotations sont séparées du corpus source avec lequel elles sont synchronisées. Cette annotation déportée (*stand-off annotation*) permet un enrichissement multi-niveaux du corpus, ce qui est intéressant en termes d'évolutivité si vous désirez rajouter d'autres couches d'annotation au corpus CO2, ce que permet la licence Creative Commons CC-BY-NC-SA.

Le corpus CO2 a fait l'objet d'un codage par deux 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.

Procédure d'annotation – Le schéma d'annotation que nous avons proposé 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 à une entité précédemment mentionnée dans le texte est nécessaire pour son interprétation (anaphore associative).

Les relations qui ont fait l'objet d'une annotation ne concernent que des entités nominales ou pronominales. Nous avons annoté l'ensemble du groupe nominal et pas uniquement sa tête. L'annotation a également concerné 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). 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 (Mazur et Dale, 2007) 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.

La délimitation des relations consiste à relier les éléments anaphoriques. Certains travaux privilégient une annotation en chaînes (Gardent et Manuélian, 2005 ; Amsili et al, 2007) c'est-à-dire en « *séquences d'expressions singulières apparaissant dans un contexte telles que si l'une de ces expressions réfère à quelque chose, toutes les autres y réfèrent également* » (Corblin, 2005). Pour le corpus CO2, il a été décidé au contraire de relier toutes les relations à la première mention de l'entité référentielle trouvée dans le texte (annotation en première mention). Il est en effet apparu que l'annotation en chaîne posait des problèmes délicats pour le dialogue, les annotateurs se trouvant devant des changements de locuteurs pour lesquels la notion de chaîne, pertinente dans la linéarité de l'écrit, devient beaucoup moins évidente à caractériser. 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.

L'annotation consiste enfin à décrire par différents traits les entités référentielles et leurs éventuelles relations. Pour les entités nous avons retenu les traits linguistiques suivants :

- 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 :

- *directe (DIR)* dans le cas d'une coréférence entre mentions de même tête nominale (*le bus rouge*…. *ce grand bus*) ;
- *indirecte (IND)* si les entités coréférentes ont des têtes nominales différentes (*le cabriolet… cette décapotable*) ;
- *pronominale (PR)* dans le cas particulier de l'anaphore indirecte où la reprise est un pronom (*le cabriolet … il roulait*…)
- *associative (ASSOC)* si les mentions ne sont pas coréférentes mais que l'interprétation de l'une dépend de l'autre (*le village … son clocher*).
- *Associative pronominale* (*ASSOC_PR*).

La description détaillée des conventions d'annotation est donnée en annexe A. On pourra consulter (Schang et al. 2012) pour une analyse quantitative et distributionnelle des différents types de mentions et de relations dans le corpus.

1.4 Corpus distribué

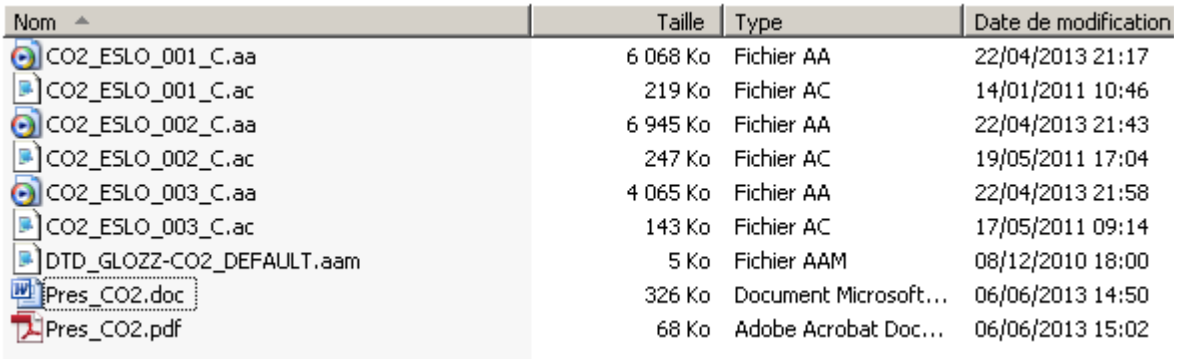
La distribution qui est diffusée ici ne concerne pas le corpus ESLO1 en lui-même : celui-ci, c'est-à-dire les fichiers audio et leur transcription orthographique au format Transcriber (Barras *et al*. 1998), est diffusé par le LLL à part. La distribution CO2 comprend au contraire :

- les transcriptions orthographiques du corpus format interne GLOZZ : fichier `.ac`,
- les annotations en coréférences (synchronisées sur les transcriptions) au format GLOZZ : fichiers `.aa`,

Le corpus en lui-même se compose donc de 3 fichiers `.ac` et 3 fichiers `.aa` (un par interview) nommés `CO2_ESLO_001, CO2_ESLO_002, CO2_ESLO_003`.

L'annotation correspond à un fichier XML qui suit une DTD spécifique définie pour le projet. La DTD est également distribuée avec le corpus : fichier `DTD_GLOZZ_CO2_DEFAULT.aam`. Notons qu'il est parfaitement possible de lire les annotations sous GLOZZ sans utiliser la DTD. Celle-ci ne vous sera utile que si vous souhaitez modifier les annotations.

S'ajoute enfin le fichier que vous avez sous les yeux, en version Word et PDF. Au final, tous les fichiers de la distribution se trouvent sous le même répertoire (figure 1).



Nom	Taille	Type	Date de modification
CO2_ESLO_001_C.aa	6 068 Ko	Fichier AA	22/04/2013 21:17
CO2_ESLO_001_C.ac	219 Ko	Fichier AC	14/01/2011 10:46
CO2_ESLO_002_C.aa	6 945 Ko	Fichier AA	22/04/2013 21:43
CO2_ESLO_002_C.ac	247 Ko	Fichier AC	19/05/2011 17:04
CO2_ESLO_003_C.aa	4 065 Ko	Fichier AA	22/04/2013 21:58
CO2_ESLO_003_C.ac	143 Ko	Fichier AC	17/05/2011 09:14
DTD_GLOZZ-CO2_DEFAULT.aam	5 Ko	Fichier AAM	08/12/2010 18:00
Pres_CO2.doc	326 Ko	Document Microsoft...	06/06/2013 14:50
Pres_CO2.pdf	68 Ko	Adobe Acrobat Doc...	06/06/2013 15:02

Figure 1 : Distribution du corpus CO2

2 Distribution du corpus

Le corpus CO2 est téléchargeable directement sur le site PAROLE PUBLIQUE où on peut le récupérer sous la forme d'une archive .ZIP : http://www.info.univ-tours.fr/~antoine/parole_publique/index.html

Il est distribué sous licence de libre diffusion *Creative Commons* (cf infra).

3 Convention d'utilisation du corpus

Le corpus CO2 est distribué sous licence *Creative Commons* CC-BY-NC-SA. 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 ci-dessous.
- *NC : non commercial* - Vous ne pouvez pas faire une utilisation commerciale de cette ressource. Nous ne sommes pas opposés sur le principe à de tels usages, mais il vous est demandé de nous contacter pour étudier ces modalités d'usage.
- *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.

Pour tout contact : Jean-Yves.Antoine_AT_univ-tours.fr. Enfin, nous vous serions reconnaissants de nous informer amicalement de toute utilisation du corpus, il s'agit d'une information toujours intéressante pour un concepteur de corpus.

4 Références bibliographiques

Liste des publications à la date d'émission de ce rapport technique. Consultez le site Internet du projet Parole Publique pour une bibliographie à jour. Les corpus CO2 et ESLO sont diffusés sous licence Creative Commons CC-BY-NC-SA qui impose, entre autres, la citation de ces publications en cas d'utilisation de ces ressources.

4.1 Publications concernant le corpus CO2

- SCHANG E., BOYER A., MUZERELLE J., ANTOINE J.-Y., ESKHOL I., MAUREL D. (2012). Coreference and Anaphoric Annotations for Spontaneous Speech Corpora In French. *Proceeding of the Discourse Anaphora and Anaphor Resolutionon Colloquium, DAARC'2012*, Faro, Portugal.

4.2 Publications concernant le corpus ESLO

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

4.3 Publications citées dans ce document

- GALLIANO S., GRAVIER G., CHAUBARD L. (2009). The ester 2 evaluation campaign for the rich transcription of french radio broadcasts. In Interspeech'09, p. 2583–2586.
- MATHET, Y., WIDLÖCHER, A. (2009). La plate-forme GLOZZ : environnement d'annotation et d'exploration de corpus. *Actes de TALN-2009*, pages 1–10.

ANNEXE A — Guide de transcription du corpus CO2

Version *2.2*
Auteurs *Jean-Yves Antoine (U. Tours), Judith Muzerelle (U. Tours), Emmanuel Schang (U. Orléans)*
Date *6 juin 2013*

1 Contexte

L'objectif du projet CO2 est de mener une étude systématique en corpus des différentes formes de réalisation des co-références anaphoriques. Cette étude linguistique constitue un préalable incontournable à la modélisation informatique de processus de résolution de ces anaphores complexes.
Ce travail portera plus précisément sur le langage oral spontané. La présence de multiples disfluences (répétitions, reprises, hésitations, incises) dans cette modalité pose, en effet, différents problèmes scientifiques qui n'ont été qu'imparfaitement résolus à l'heure actuelle. Ces travaux seront réalisés sur les ressources orales développées par les laboratoires partenaires du projet.

- Corpus ESLO 1 et 2 du laboratoire LLL. Ce corpus de référence, qui est en cours de finalisation, constituera le plus grand corpus oral en français diffusé auprès de la communauté scientifique. Il comporte avant tout des interviews légèrement interactives.
- Corpus (OTG, Accueil UBS) de la banque de corpus PAROLE_PUBLIQUE. Diffusés par le LI, ces ressources correspondent à du dialogue finalisé (renseignement touristique, accueil téléphonique) fortement interactif. Ils représentent, à ce jour, la plus grande base de corpus de dialogue finalisé diffusée librement en français.

Le projet CO2 a donc pour objet l'étude de toutes les formes de reprises anaphoriques en s'attachant en particulier à décrire les reprises nominales. Il consiste à étudier les reprises anaphoriques, en particulier celles qui apportent des informations nouvelles sur le référent.

Nécessaire à la conduite de nos analyses, ce projet conduira par ailleurs à la constitution d'un **corpus annoté en reprises anaphoriques**. Ce document décrit précisément les annotations qui seront apportées aux corpus. Celles-ci visent à décrire :

- la distribution entre les différents types de reprises référentielles (nominale, pronominale) ;
- l'étude des respects de critères d'accord morpho-syntaxiques (nombre, genre par exemple) dans le cas des anaphores pronominales ;
- la détermination des syntagmes nominaux ;
- les types d'entités du discours référées : personnes physiques ou morales, groupes de personnes, objets, lieux, etc. ;
- la nature des relations anaphoriques ;
- la part du "nouveau" et du "déjà connu" déterminée par l'analyse sémantique.

2 Outil d'annotation : GLOZZ

L'outil d'annotation qui sera utilisé est GLOZZ dans sa version 1.1.0. Cet outil est disponible à l'URL suivante : http://www.glozz.org/.
GLOZZ repose sur le principe d'une annotation déportée : chaque mot typographique du document est numéroté par le logiciel. Les annotations sont portées dans un fichier XML à part du corpus. Le numéro des items typographiques dans le corpus servant lieu de synchronisation entre ce dernier et les annotations. Cette synchronisation est utilisée par l'interface de GLOZZ pour afficher de manière transparente les annotations portées sur le corpus. Les annotations sont sauvées dans un fichier d'extension `.aac`.
Le format d'annotation de GLOZZ n'est pas figé mais peut-être défini à l'aide d'une DTD (fichier avec extension `.aam`). On donne en annexe B et C les DTD correspondant aux annotations qui seront réalisées dans le cadre de ce projet. Ces fichiers s'appellent `DTD_GLOZZ_ANCOR_DEFAUT.aam` (pour les fichiers ESLO uniquement) et `DTD_GLOZZ_ANCOR_DIALOGUE.aam` (pour les fichiers Parole_Publique uniquement, cf. §7.4).

3 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.

4 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).

4.1. 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 (cf. § 5.4)

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.

4.2. 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 )

4.3. 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*.

4.4. 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. §7.1).

4.5. 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 dans le §5.1) et ne recevra aucune caractérisation lors de l'annotation.

5 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.

5.1. 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 (voir §4.4.)	(voir §4.4.)
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 (§4.2) 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.

5.2. 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 …*

5.3. 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.

5.4. 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 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.

5.5. 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	

5.6. 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	

5.7. 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	*Unknown* : l'annotateur n'a pu se décider.	Cette valeur ne devrait jamais être utilisée sur cet attribut.

5.8. 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	Elément introduit précédemment	-
NULL	Artefact	Uniquement si TYPE = NULL
UNK	*Unknown* : l'annotateur n'a pu se décider.	Cette valeur ne devrait jamais être utilisée 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.

6 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.

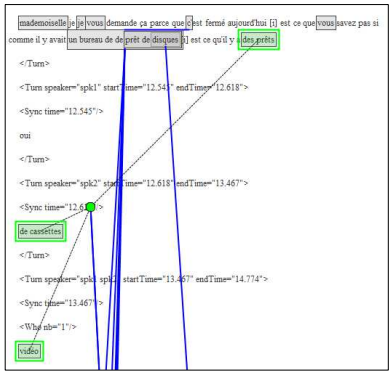
Cas des pronoms relatifs – Les pronoms relatifs seront reliés anaphoriquement à leur antécédent grammatical direct :

- *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-dessous, 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

7 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

7.1. 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 référent 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 référent 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*.
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, celui-ci reprenant tout ou en partie le GN (cf. §4.4). 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 ne s'agit pas de la même entité.

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 anaphore 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 entités référentes ne peut pas 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).

7.2. 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 CO2. Il sera positionné automatiquement par GLOZZ*

7.3. 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 CO2. Il sera positionné automatiquement par GLOZZ*

7.4. 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.	

ANNEXE B — DTD Glozz adaptée pour le corpus CO2

```
<?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>
					</possibleValues>
				</feature>
				<feature name="DEF">
					<possibleValues default="DEF_SPLE">
						<value>INDEF</value>
						<value>DEF_SPLE</value>
						<value>DEF_DEM</value>
					</possibleValues>
				</feature>
				<feature name="GEN_REF">
					<possibleValues default="SPEC">
						<value>SPEC</value>
						<value>GENE</value>
						<value>NULL</value>
						<value>UNK</value>
					</possibleValues>
				</feature>
				<feature name="GP">
					<possibleValues default="NO">
						<value>YES</value>
						<value>NO</value>
					</possibleValues>
				</feature>
				<feature name="NEW">
					<possibleValues default="NO">
						<value>YES</value>
						<value>NO</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>
			</possibleValues>
		</feature>
		<feature name="DEF">
			<possibleValues default="DEF_SPLE">
				<value>EXPL</value>
				<value>INDEF</value>
				<value>DEF_SPLE</value>
				<value>DEF_DEM</value>
			</possibleValues>
		</feature>
		<feature name="GEN_REF">
			<possibleValues default="SPEC">
				<value>SPEC</value>
				<value>GENE</value>
				<value>NULL</value>
				<value>UNK</value>
			</possibleValues>
		</feature>
		<feature name="GP">
			<possibleValues default="NO">
				<value>YES</value>
				<value>NO</value>
			</possibleValues>
		</feature>
		<feature name="NEW">
			<possibleValues default="NO">
				<value>YES</value>
				<value>NO</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>
                    </possibleValues>
                </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>
                    </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>
```