HAL open science

Codage en chaîne ou en première mention de la coréférence : approcher la structure des chaînes de référence par comparaison des deux annotations

Jean-Yves Antoine, Anaïs Lefeuvre, Emmanuel Schang

▶ **To cite this version:**

Jean-Yves Antoine, Anaïs Lefeuvre, Emmanuel Schang. Codage en chaîne ou en première mention de la coréférence : approcher la structure des chaînes de référence par comparaison des deux annotations . 5ème Congrès Mondial de Linguistique Française (CMLF'2016), Laboratoire Ligérien de Linguistique, Jul 2016, Tours, France. hal-01344976

HAL Id: hal-01344976

https://hal.science/hal-01344976v1

Submitted on 13 Jul 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NoDerivatives 4.0 International License

Codage en chaîne ou en première mention de la coréférence : approcher la structure des chaînes de référence par comparaison des deux annotations

Jean-Yves Antoine (1), Anaïs Lefeuvre (3,1), Emmanuel Schang (2)

Affiliation : 1 Université François Rabelais de Tours, LI, E.A. 6300
2 Univ.Orléans, CNRS, LLL, UMR 7270
3 Université Paris Sorbonne, Paris 4, STIH
anais.lefeuvre@paris-sorbonne.fr, jean-yves.antoine@univ-tours.fr, emmanuel.schang@univ-orleans.fr

1 Introduction

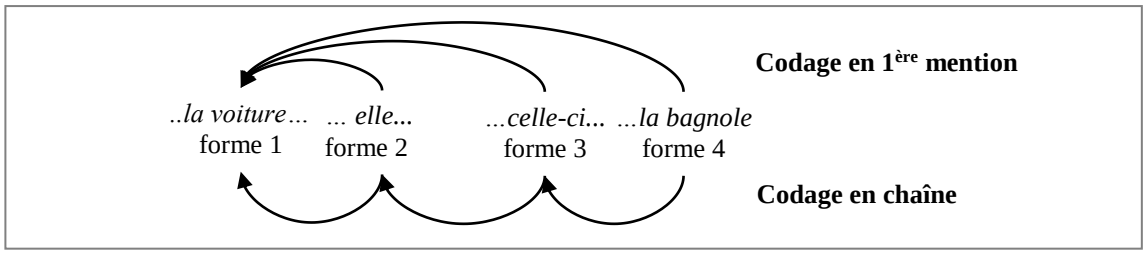
La résolution des anaphores est un sujet ancien et central dans le traitement automatique du discours (v. Mitkov 2010 notamment pour une présentation et un état de l'art). Depuis Karttunen (1976)[1], on envisage essentiellement le problème de la manière suivante :

> « Consider a device designed to read a text in some natural language, interpret it, and store the content in some manner, say, for the purpose of being able to answer questions about it. To accomplish this task, the machine will have to fulfil at least the following basic requirement. It has to be able to build a file that consists of records of all the individuals, that is, events, objects, etc., mentioned in the text and, for each individual, record whatever is said about it. »

Mais déterminer comment il est fait mention des entités (*individuals*) dans le discours n'est pas une tâche simple. A la différence des langages mathématiques (par exemple), les langues naturelles n'assignent pas un identifiant unique aux objets auxquels il est fait mention par les locuteurs. On appelle chaîne de références la suite des expressions d'un texte qui ont la même identité référentielle (Schnedecker & Landragin 2014). La tâche qui consiste à identifier ces chaînes est difficile à automatiser et demande des corpus annotés en coréférence afin de dégager des heuristiques de traitement ou pour l'entraînement des approches par apprentissage automatique. Le corpus ANCOR est précisément un de ces corpus (v. Lefeuvre & al. 2014 pour une revue des corpus disponibles).

Cette étude porte sur la comparaison des résultats de l'annotation de la coréférence en chaîne et en première mention du corpus ANCOR (Muzerelle et al. 2013). Par annotation en chaîne, nous entendons l'identification des liens entre expressions linguistiques (nominales ou pronominales) qui ont un même référent[2]. Dans l'annotation en première mention, les liens sont faits entre la première mention d'une entité et les expressions suivantes qui ont le même référent (voir figure 1).

Figure 1 – Codage en chaîne et en première mention d'une chaîne de référence.



[1] L'article date de 1976 mais sa présentation à l'*International Conference on Computational Linguistics* à Sånga-Säby en Suède a eu lieu en 1969.

[2] Pour une définition plus précise et une discussion détaillée de cette notion, nous renvoyons à Schnedecker (1997).

Dans la mesure où ANCOR a été annoté, pour des raisons pratiques (Muzerelle et al. 2013) en première mention, nous nous interrogeons sur l'impact de ce choix, en particulier du point de vue des résultats expérimentaux qui peuvent être obtenus par une analyse linguistique du corpus :

- Existe-t-il des différences significatives entre les observations faites avec les deux différents modes d'annotation et quel est leur impact sur les conclusions théoriques que l'on peut en tirer ? A terme, nous espérons que cette interrogation permettra de déterminer quelle annotation se révèle la plus efficace pour identifier des heuristiques de résolution et modélisation des coréférences.

- Disposer à la fois d'une annotation en chaîne et en première mention permet-il, par comparaison, de tirer des enseignements complémentaires sur la structure interne des chaînes de référence ? L'idée sous-jacente à cette question est assez simple : si deux expressions référentielles X et Y ont le même antécédent A dans le codage en première mention (coréférence) et que X précède Y dans le codage en chaîne, on peut déduire une chaîne A→X→Y. Mais que peut-on déduire de plus de cette nouvelle annotation ? Pour apporter un début de réponse à ces questions, une partie du corpus ANCOR_Centre a été annotée manuellement en chaîne à partir des annotations en première mention[3].

Dans un premier temps, nous allons décrire le corpus ANCOR_Centre sur lequel a été menée cette étude expérimentale, ainsi que l'outil de requêtage ANCORQI qui nous a permis d'obtenir les résultats quantitatifs ainsi que les exemples présentés dans l'article. Ensuite, nous présenterons un ensemble de résultats expérimentaux qui montrent en quoi la comparaison de deux types de codage nous permet d'obtenir des indications intéressantes sur la réalisation des coréférences. Nous verrons également que l'impact du codage sur les conclusions que l'on peut tirer d'une analyse du corpus reste limité.

2 Le corpus ANCOR_Centre et son outil de requêtage ANCORQI

Le corpus ANCOR_Centre est un corpus annoté en référence qui a été réalisé dans le cadre du projet régional ANCOR qui a réuni le Laboratoire Ligérien de Linguistique (CNRS & Université d'Orléans) et de Laboratoire d'Informatique (LI) de l'Université François Rabelais de Tours. Il s'agit d'un corpus francophone de dialogue oral qui ambitionne de représenter une réelle diversité de situations discursives dans cette modalité. Il regroupe ainsi quatre sous-corpus de parole spontanée (Lefeuvre et al. 2014). Deux d'entre eux ont été extraits du corpus ESLO, qui regroupe des entretiens sociolinguistiques présentant un degré d'interactivité faible. A l'opposé, les deux autres corpus, OTG et Accueil_UBS concernent des dialogues présentant une plus forte interactivité. Ces deux derniers corpus diffèrent par le média utilisé : le corpus OTG regroupe des conversations de visu au sein d'un office de tourisme pour OTG, tandis qu'Accueil_UBS a été enregistré dans un standard téléphonique. Au total, le corpus regroupe 488 000 mots pour une durée d'enregistrement de 30,5 heures (Tableau 1 page suivante).

Les formes annotées ont été définies par un critère syntaxique : seuls les syntagmes nominaux (N) et les pronoms (PR) ont été considérés par l'annotation.

Un des objectifs du projet ANCOR était de proposer une annotation fine permettant une interrogation du corpus suivant un grand nombre de propriétés linguistiques. Outre leur catégorie syntaxique, les mentions référentielles ont ainsi été décrites par les propriétés linguistiques suivantes :

- **G : genre** (masculin, féminin, inconnu) et **N : nombre** (singulier, pluriel, inconnu)
- **GP : inclusion dans un GP** – Valeur binaire : YES si l'entité référentielle est incluse dans un GP ou NO sinon.

[3] La procédure de passage en chaîne est automatisable au niveau de la délimitation des maillons (antécédent et reprise), mais pas au niveau des attributs qui permettent une caractérisation des relations annotées. Le corpus ANCOR_Centre repose en effet sur une annotation riche de la coréférence et de l'anaphore associative, afin d'autoriser des études linguistiques relativement fines. Le passage semi-automatique au codage en chaîne du corpus est en cours de réalisation dans le cadre d'un projet financé par l'ANR DEMOCRAT (ANR-15-CE38-00XX).

- **EN : types d'entités nommées** – Les types retenus sont ceux dans la campagne d'évaluation ESTER2 (Galliano et al., 2009) à savoir PERS (personne réelle ou fictive), LOC (localisation, à savoir tout géonyme), FONC (fonction politique, militaire, administrative, religieuse etc. d'une personne), ORG (organisations : états, entreprises, institutions…), PROD (production humaine), TIME (expressions temporelles : date et heure), AMOUNT (quantités telles que température, longueur, aire, volume, poids, vitesse, valeur monétaire, âge etc. mais aussi durées temporelles) et EVENT (évènement). On utilise le type NO si l'entité n'est pas une entité nommée.
- **DEF : définitude** (*definitiveness* dans la littérature anglo-saxonne) – Cet attribut sert à distinguer le caractère défini ou indéfini des entités référentielles. On distingue plus précisément les définis simples (DEF_SIMPLE), les définis démonstratifs (DEF_DEM) et les indéfinis (INDEF). Les mentions non référentielles sont caractérisées par un trait explétif (EXP) spécifique.
- **GENE : généricité** – Permet de décrire si l'entité considérée dénote un référent générique (*l'homme est un loup pour l'homme*) ou spécifique (*l'homme s'est enfuit par la sortie de secours*)
- **NEW : nouvelle mention** – Attribut binaire qui précise si la mention constitue (YES) ou non (NO) une nouvelle entité du discours. L'ancre d'une chaîne de référence aura donc le type NEW, contrairement aux maillons suivants, mais des références isolées seront également annotés NEW.

Tableau 1 – *Contenu du corpus ANCOR : corpus audio sources*

Corpus	Situation discursive	Finalisation[4]	Interactivité	Taille & Durée
ESLO_ANCOR	Interview	Modérée	Faible	417 kMots – 25h
ESLO_CO2	Interview	Modérée	Faible	35 kMots – 2,5 h
OTG	Dialogue oral	Très forte	Forte	26 kMots – 2h
Accueil_UBS	Dialogue téléphonique	Assez forte	Forte	10 kMots – 1 h

Les relations, qui relient les mentions coréférentielles, portent quant à elles des traits d'annotation d'accord en genre, d'accord en nombre et d'identité du locuteur (on cherche à savoir si la reprise est réalisée par le même locuteur que celui qui a introduit l'antécédent). Le corpus distingue par ailleurs cinq types de relations. Les trois premiers types concernent la coréférence, c'est-à-dire les situations où il y a identité de référence entre les mentions reliées :

- DIR : coréférence directe, dans le cas d'une coréférence entre mentions de même tête nominale (exemple : *le bus rouge… ce grand bus*),
- IND : coréférence indirecte[5], dans le cas d'une coréférence où les mentions ont des têtes nominales différentes (exemple : *le car… ce bus*),
- PR : coréférence pronominale[6], dans le cas particulier de la coréférence indirecte où la reprise est un pronom, (exemple : *le bus… il roulait*).

Ces trois relations constituent le cœur de notre étude sur la coréférence, à l'opposé des deux dernières qui relèvent de relations anaphoriques sans coréférence (*bridging anaphora*)

- ASSOC : anaphore associative (*bridging anaphora*) si les mentions ne sont pas coréférentes mais que l'interprétation de l'une dépend de l'autre (exemple : *le bus … son chauffeur*),

[4] Par finalisation, on entend le degré de focalisation du dialogue sur une tâche donnée liée au déroulement du dialogue

[5] Ou encore « anaphore infidèle ». Nous empruntons ici le terme à la littérature anglo-saxonne (cf. Vieira et al. 2004).

[6] Ou encore « anaphore pronominale » (*pronominal anaphora* dans la littérature anglo-saxonne). Nous préférons insister ici encore sur la nature coréférentielle de la relation plutôt que sur le processus anaphorique sous-jacent à la résolution de la référence.

- ASSOC_PR : anaphore associative pronominale, dans le cas où la reprise associative est portée par un pronom comme dans le cas de métonymies : *le café Jeanne d'Arc, ils sont tous désagréables*.

Nous avons réalisé l'exploration de ce corpus grâce à ANCORQI, un outil implémenté en langage Python et permettant d'exprimer des requêtes complexes tirant partie de la richesse d'annotation du corpus. ANCORQI est assorti d'un concordancier qui offre au chercheur une visualisation de chacune des relations répondant à la requête.

Les études que nous allons présenter dans cet article ont été réalisées sur le sous-corpus Accueil_UBS, seule partie d'ANCOR_Centre traduite en codage en chaîne à l'heure actuelle. Accueil_UBS est le plus petit élément du corpus, puisqu'il ne regroupe que 10 000 mots. Toutefois, des études distributionnelles ont montré qu'il était représentatif de l'ensemble du corpus (Lefeuvre et al. 2014). Par ailleurs, Accueil_UBS regroupe 691 relations anaphoriques, ce qui donne déjà une certaine pertinence aux résultats qui vont être présentés dans ces lignes.

3 Méthodologie : comparaison des annotations suivant le codage

Cet article décrit les observations que nous avons pu tirer de la comparaison des annotations en chaînes et en première mention menées sur le même sous-corpus (Accueil_UBS). Les études expérimentales que nous présentons ont tiré profit de cette double annotation parallèle pour répondre à trois questionnements différents :

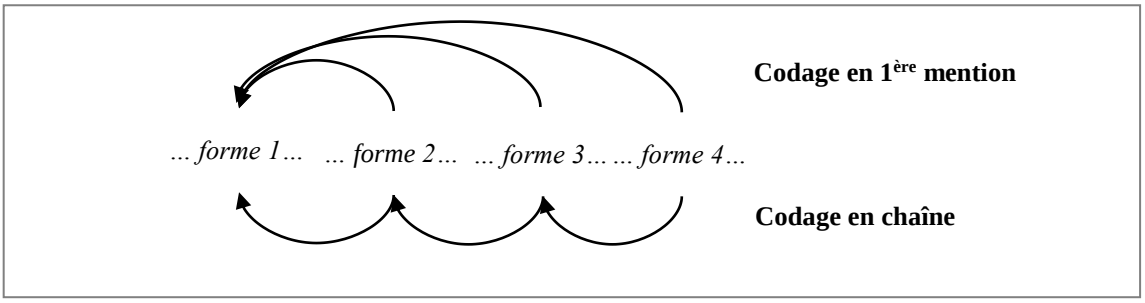
- Quels sont les éléments qui introduisent les chaînes de référence de manière privilégiée (§ 4) ?
- Que peut-on découvrir de la structure interne des chaînes de références à partir de la comparaison des deux codages (cf. § 5).
- Le type de codage a-t-il une influence sur les conclusions que l'on peut tirer en matière d'accord en genre et en nombre au fil des chaînes de références ? Ce point sera décrit en section 6.

Pour plus de clarté, nous adopterons les définitions ci-dessous dans la suite de l'article :

- **Chaîne** – On appelle chaîne de référence une suite de mentions qui réfèrent à une même entité du discours dans une énonciation donnée.
- **Ancre** – On appelle ancre la première mention qui introduit une chaîne de référence. L'ancre constituera toujours une nouvelle entité référentielle dans le discours (trait NEW).
- **Antécédent** – Pour une relation donnée, on appelle antécédent la mention qui est à son origine, i.e. la première dans le fil du discours des deux mentions liées par la relation. L'ancre d'une chaîne constitue donc un antécédent particulier (celui du premier maillon de la chaîne).
- **Reprise** – La reprise est au contraire la seconde mention d'une relation au fil du discours.

Considérons à nouveau la figure 1, dans laquelle les mentions qui constituent la chaîne de référence sont cette fois numérotées (figure 2)

Figure 2 – Codage en chaîne et en première mention d'une chaîne de référence.



La *forme1* constitue l'ancre de la chaîne de référence quel que soit le codage. Elle constitue l'unique antécédent de toutes les relations dans un codage en chaîne, toutes les autres mentions faisant office de reprises. A l'opposé, dans un codage en chaînes, *forme1*, *forme2* et *forme3* sont des antécédents, tandis que *forme2*, *forme3* et *forme4* sont des reprises. Il est donc à remarquer que dans ce codage, toutes les mentions internes à la chaîne sont à la fois reprise d'une relation et antécédent de la suivante.

4 Quelles ancres pour les chaînes de référence ?

Le codage en première mention met en exergue l'ancrage des chaînes de référence, puisque toutes les relations de la chaîne pointent sur cette mention introductive. En conséquence, l'étude comparative de la distribution des relations suivant le type de codage (chaîne ou première mention) peut être révélatrice de la nature des éléments qui ancrent les chaînes de référence. Par exemple, si un type de mention prédomine dans le codage en première mention, c'est qu'il est plus fréquemment utilisé pour ancrer une chaîne. Pour illustrer notre propos, considérons l'exemple artificiel de chaîne de référence ci-dessous, où l'on a associé à chaque forme son type de définitude.

Exemple (1)	*forme 1 …*	*forme 2 …*	*forme 3…*	*forme 3…*	*forme 2 …*	*forme 4*
	INDEF	DEF_DEM	DEF	DEF	DEF_DEM	DEF

On remarque immédiatement que l'ancre de la chaîne (*forme1*), et surtout son type (INDEF), prédominent comme antécédent dans le codage en première mention (tableau 2). C'est en partant de cette idée que nous avons cherché à caractériser de manière expérimentale le comportement de certains types de mentions comme ancres des chaînes de référence.

Tableau 2 – *Répartition des relations de coréférence en fonction de leur antécédent dans l'exemple (1)*

Codage	**Forme 1**	**Forme 2**	**Forme 3**	**Forme 4**	**Total**
1ère mention	5	0	0	0	**5**
Chaîne	1	2	2	0	**5**
Codage	**INDEF**	**DEF_DEM**	**DEF**		**Total**
1ère mention	5	0	0		**5**
Chaîne	1	2	2		**5**

On remarque que le nombre total de formes ou de types de forme reprises ne varie pas d'un codage à l'autre, tandis que le passage de première mention à chaîne fait apparaître une différence notoire dans la répartition des antécédents, pivot principal pour notre analyse. Ce transfert de répartition laisse entrevoir plus clairement la structuration des chaînes référentielles comme nous le montrerons plus loin. Sur cet exemple précis, la *forme 1*, qui est l'ancre quel que soit le codage n'apparait pas comme forme prévalente pour l'antécédence dans le codage en chaîne. Il en va de même pour son type INDEF. Les *forme 2* et *forme 3* et leurs types (DEF_DEM et DEF) n'apparaissent comme antécédents que dans le codage en chaîne. La *forme 4* est la dernière reprise de la chaîne. Elle n'apparait donc comme antécédent dans aucun des deux codages. Son statut est donc repérable directement dans le codage en chaîne puisque, contrairement au codage en première mention, elle est la seule forme pour laquelle aucune relation n'est comptabilisée.

Cette étude va concerner différents types de mention : les pronoms, les syntagmes nominaux, les démonstratifs, les indéfinis et les noms de personnes. Nous nous poserons également la question de l'ancrage d'une chaîne de référence dans un groupe prépositionnel et situerons à chaque fois nos observations avec quelques prédictions issues de la littérature sur les anaphores.

4.1 Ancrage des chaînes de référence par un pronom ou un syntagme nominal

Le tableau 3 ci-dessous compare, suivant le type de codage, la distribution des antécédents des relations de coréférence (anaphores associatives exclues) dont les anaphores pronominales. Les observations sont édifiantes : les pronoms ne représentent que 7,6% des antécédents dans le cas du codage en première mention, contre plus du tiers dans le cas du codage en chaîne.

Tableau 3 – *Répartition des relations de coréférence suivant le type d'antécédent dans les deux codages*

Codage	**Ensemble des relations**		*dont anaphores Pronominales*	
	GN antécédent	**PR antécédent**	*GN antécédent*	*PR antécédent*
1ère mention	92,4%	7,6%	*86,3%*	*13,7%*
chaîne	65,6%	34,4%	*46,2%*	*53,8%*

Il est très largement admis que les pronoms reprennent une entité saillante et qu'ils sont de mauvais candidats en début de chaîne (Huang 2000:308ff ; Walker, Joshi & Prince 1998 ; Chiarcos 2009). Comme on pouvait donc s'y attendre, on observe que les pronoms ancrent rarement les chaînes de coréférence. Ceci était prévisible, puisqu'une recherche antérieure avait déjà montré que les pronoms ne représentent que 2,2% des premières mentions (*discourse new mentions*) dans le corpus Accueil_UBS (Lefeuvre et al. 2014). On remarque par ailleurs que le taux de 7,6% d'ancrage pronominal est avant tout dû aux anaphores pronominales, c'est-à-dire suivant un schéma où la chaîne commence par une séquence PR→PR, alors qu'à l'opposé, les chaînes commençant par une cataphore (premier chaînon PR→N) ne représentent que 2,0% cas (cf § 5.1). L'exemple (2) ci-dessous illustre l'introduction d'une nouvelle entité référentielle par une chaîne purement pronominale (*il*/*celui*/*qui*/*il*/*il*/*il*) : la standardiste passe ici d'un discours centré sur une employée à son collègue masculin sensé la remplacer, et ceci sans jamais préciser son identité ni mentionner dans un premier temps ce remplacement. On remarque au passage que cette chaîne purement pronominale concerne une personne : nous reviendrons précisément plus loin sur la spécificité de ce type d'entité référentielle ultérieurement (cf § 4.5).

Exemple (2) – Dialogue Accueil_UBS_ 054_00000037

Loc1	*oui alors je vais vous passer une personne qui se trouve* [pf] *# alors attendez elle s'est absentée ici mais je vais vous la passer ailleurs parce qu'elle partie à la repro donc elle s'occupe de ça conservez hein*
Loc2	*c'est gentil madame*
Loc1	*alors* ***il*** *ne répond pas pour le moment*
(…)	[3 tours de parole sans aucune autre reprise]
Loc 1	*e oui rappelez ici # autrement je vais vous donner* ***celui*** ***qui*** *est à la repro mais* ***il*** *va pas rester tout le temps à la repro c'est une collègue qui est absente donc* ***il*** *la remplace pour e faire des papiers e des dossiers que les profs ont besoin [pf] donc mais* ***il*** *va pas être là tout le temps donc rappelez ici*

Notons enfin que les observations faites sur les pronoms esquissent en creux une présence privilégiée des SN comme introducteurs (ancres) de chaîne.

4.2 Ancrage des chaînes de référence par un démonstratif

Les démonstratifs se prêtent à différents emplois, de déictiques à anaphoriques en passant par anadéictiques (Cornish 2011). Ils sont relativement rares dans le corpus puisqu'ils ne représentent que 6,5% des mentions (Lefeuvre et al. 2014). Comme l'explique Charolles (2002, chap.5), les SN démonstratifs sont substituables aux SN définis dans de nombreux contextes même si le mode de donation du référent qui leur est attaché diffère de celui des définis. On considère souvent que les 'vrais' démonstratifs s'accompagnent d'un signe d'ostension (regard, geste) en direction d'un référent identifiable

par le locuteur et ceux à qui il s'adresse. Dans un corpus téléphonique (pas de présentiel visuel) tel qu'UBS, il est naturel de ne pas trouver beaucoup de démonstratifs illustrant ces emplois.

Tableau 4 – *Nombre de relations de coréférences dont l'antécédent est un démonstratif*

Codage	**Toute coréférence**	*dont directe*	*dont indirecte*	*dont pronominale*
1ère mention	17	5	*0*	*12*
Chaîne	69	6	*16*	*47*

De par leur relative rareté, les démonstratifs seraient noyés au milieu des autres catégories de mention dans une étude en répartition statistique. Aussi avons-nous choisi de simplement comparer, suivant le type de codage, le nombre de relations de coréférences dont l'antécédent est un démonstratif. Les résultats présentés dans le tableau 4 montrent un quadruplement, entre le codage en première mention et le codage en chaîne, du nombre de relations dont l'antécédent est un démonstratif. Cette augmentation très significative indique que les démonstratifs apparaissent rarement comme ancres d'une chaîne de référence. Des exemples peuvent toutefois être trouvés dans le corpus, tels celui-ci qui présente un emploi situationnel – *immediate situation use* chez Hawkins (1978) :

Exemple (3) – Dialogue Accueil_UBS_048_00000031

Loc 1	*U B S Bonjour*
Loc 2	*oui bonjour madame e j'aurais souhaité parler à madame Nom s'il vous plait*
Loc 1	*madame Nom conservez je vais voir si e # ah ben elle est pas elle est sur Lorient madame Nom*
Loc 2	*ah bon parce qu'on m'a donné* ***ce numéro***
(…)	[plusieurs tours de parole sans reprise]
Loc 2	*j'ai le 88 56 09 # on m'a dit d'essayer soit l'****un*** *soit l'autre*

L'analyse de la répartition du type de relations dont l'antécédent est un démonstratif montre par ailleurs que ces derniers se trouvent dans 68% des cas (47 cas sur 69 dans le tableau 4) en tête d'une relation pronominale N→PR (*ce numéro/il* dans l'exemple ci-dessus). Les cas de reprise directe N→N du démonstratif sont au contraire bien plus rares : on n'a recensé que 6 exemples dans le corpus.

Enfin, les 23% de relations indirectes observées dans le codage en chaîne (16 observations sur 69 dans le tableau 4) correspondent toutes à des reprises nominales de pronoms (PR→N) et non pas des maillons homogènes N→N avec changement de tête lexicale. Comme nous le verrons plus loin (cf § 5.1), ces transitions PR→N se retrouvent au sein milieu de chaîne.

4.3 Ancrage des chaînes de référence et définitude

De nombreux travaux ont intégré la définitude dans leur modèle de réalisation des chaînes de référence. Ici, nous avons cherché à étudier comment les indéfinis intervenaient dans l'ancrage des chaînes. Les indéfinis sont relativement peu nombreux dans le corpus Accueil_UBS, puisqu'ils ne représentent que 11,9% des mentions (Lefeuvre et al. 2014). Aussi avons-nous préféré dans un premier temps mener là encore une analyse en dénombrement plutôt qu'en répartition statistique.

Tableau 5 – *Nombre de relations de coréférence dont l'antécédent est un indéfini dans les deux codages*

Codage	**Toutes relations**	*dont directes*	*dont indirectes*	*dont pronominales*
1ère mention	110	37	*7*	*66*
chaîne	81	29	*7*	*45*

A l'opposé des démonstratifs, on remarque sur le tableau 5 une augmentation notable (36%) des relations introduites par un indéfini lors du passage du codage en chaîne au codage en première mention. Cet accroissement suggère que les indéfinis inclus dans une chaîne de référence le sont de manière privilégiée comme ancre (première mention). C'est le cas, très fréquent dans le corpus, de l'exemple suivant où l'indéfini *une personne* donne lieu ensuite à des reprises pronominales définies (pronom *elle*) :

Exemple (4) – Dialogue Accueil_UBS_054_00000037

Loc 1 *oui alors je vais vous passer **une personne** **qui** se trouve [pf] # alors attendez **elle** est absentée ici mais je vais vous **la** passer ailleurs parce qu'**elle** est partie à la repro donc **elle** s'occupe de ça conservez hein*

Comme nous le verrons dans la section suivante, cette observation va dans le sens d'une progression INDEFINI -> DEFINI privilégiée au cours du dialogue à une progression DEFINI -> INDEFINI pour lequel l'identité de référence pourrait être questionnée, comme sur l'exemple artificiel ci-dessous : dans l'exemple (5b) ci-après, on ne peut admettre une lecture coréférentielle entre *le navire* et *un navire*, mais y voir le passage d'un référent spécifique à un autre référent générique.

Exemple (5) – Exemple artificiel

(5a) ***Un navire** apparait à l'horizon. **Il** a cargué ses voiles. **Ce navire** sera à quai avant le jusant.*
(5b) ***Le navire** apparait à l'horizon. **Un navire** aussi rapide devrait être à quai avant le jusant.*

Comme l'explique Charolles (2002, chap.6) à la suite de nombreux autres auteurs, les SN indéfinis sont des expressions autonomes référentiellement et cette autonomie leur confère un rôle prototypique d'introduction des référents nouveaux, c'est-à-dire des ancres de chaînes.

Les observations que l'on peut faire sur les définis (simples ou démonstratifs) sont le pendant de ceux faits sur les indéfinis. Nous venons d'observer que les indéfinis étaient plus facilement en tête qu'en milieu de chaîne. De fait, d'un point de vue distributionnel, leur proportion est effectivement plus importante dans un codage en première mention (tableau 6).

Tableau 6 – *Répartition des relations de coréférence suivant leur antécédent dans les deux codages*

Codage	**Indéfinis**	**Définis**	*dont définis simples*	*dont démonstratifs*
1ère mention	18,6 %	81,4 %	*74,4 %*	*7,0 %*
chaîne	13,8 %	86,2 %	*78,5 %*	*7,7%*

Sur l'ensemble du corpus ANCOR, nous avons par ailleurs observé que 12% des mentions étaient des indéfinis. On remarque dans le tableau 6 que 13,8% de relations de coréférences sont introduites par un indéfini. Cette proximité de répartition des indéfinis dans la population totale des mentions et dans celles des antécédents tend à montrer que définis et indéfinis ont une capacité équivalente à jouer le rôle d'antécédent. A l'opposé, avec 18,6% des premières mentions, on peut faire l'hypothèse que les indéfinis jouent plus significativement un rôle d'ancre que d'antécédent en cours de chaîne.

Il n'en reste pas moins que les définis représentent une forte majorité d'antécédents de relations, en début comme en milieu de chaîne. On retrouve des résultats comparables dans la littérature. (Recasens et al., 2009) observe ainsi 73% de définis en initiale de chaîne (ancre) sur l'espagnol.

4.4 Ancrage des chaînes dans un groupe prépositionnel

Le corpus ANCOR codant ce type d'information, nous nous sommes penchés sur la situation des groupes prépositionnels (GP) comme première mention de chaînes de référence. Selon la plupart des modèles de coréférence, l'inclusion dans un GP rend la mention moins accessible pour ancrer une chaîne. Nos observations expérimentales sont plus mesurées sur cette influence syntaxique. Comme le montre le tableau 7, les antécédents de relation sont inclus dans un GP dans 22,3% (codage en chaîne) à 31,7%

(codage en première mention) des cas. Ces situations sont certes minoritaires, mais la moindre accessibilité du GP n'empêche nullement qu'une mention incluse dans un GP introduise une nouvelle chaîne. La comparaison entre les deux codages suggère par ailleurs que le GP est plus facilement antécédent en début qu'en milieu de chaîne, puisque leur proportion est significativement supérieure (accroissement de 42% entre 22,3% et 31,7%) en codage en première mention : toute choses égales par ailleurs, les mentions incluses dans un GP sont donc préférablement des ancres de chaînes.

Tableau 7 – *Répartition des relations de coréférence suivant le type syntaxique de l'antécédent (GN ou GP) dans les deux codages (première mention et chaîne).*

Codage	**GN hors GP**	**GP**	*dont pronominale*	*dont directes ou indirectes*
1ère mention	68,3 %	31,7 %	*9,2 %*	*22,5 %*
chaîne	77,7 %	22,3 %	*3,7 %*	*18,6 %*

Nous faisons l'hypothèse que cet ancrage par un GP se retrouve dans des situations de rupture thématique qui amènent la création de nouvelles chaînes de référence.

C'est le cas de l'exemple (6), où après avoir tenté durant de nombreux tours de parole à ajuster leurs emplois du temps, les interlocuteurs se focalisent sur la personne qui pourrait les suppléer dans le cas d'un impondérable. La mention de cette personne est introduite dans le GP *à mon collègue* qui ancre ensuite une chaîne d'anaphores pronominales :

Exemple (6) – Dialogue Accueil_UBS_024_00000019

Loc 2	*sur le programme de demain matin*
Loc 1	*voilà demain e et qu'on fasse cela*
Loc 2	*et je téléphone* ***à mon collègue***
Loc 1+2	*1 : demain* *2 : ouais*
Loc 2	*matin pour savoir s'****il*** *est là au cas où il y aurait un*
Loc 1	*souci*
Loc 2	*un cheese e qu'****il***
Loc 2+1	2 : *soit prêt pour me rendre service* 1 : *oui*

Le fait que ce GP soit argument du verbe lui confère une saillance qui peut expliquer cette facilité de reprise. Une étude en corpus plus approfondie permettra de vérifier cette hypothèse souvent donnée par la littérature (Chiarcos 2009).

Une analyse plus fine montre que l'accroissement entre les deux codages de la proportion de GP antécédents est précisément le fait de ces anaphores pronominales, dont la fréquence triple (9,2% contre 3,7%) dans le codage en première mention. Étant donné que le premier maillon d'une chaîne n'est pas impacté par le type de codage, ces chaînons supplémentaires sont donc le fait d'anaphores pronominales qui ne suivent pas directement la première mention GP. Cette observation suggère donc que, contrairement aux noms, les pronoms ont plus rarement pour antécédent une mention incluse dans un GP et arrivent ultérieurement dans la chaîne. Ce résultat peut trouver une explication : ceux des GP qui ne sont pas arguments du verbe sont syntaxiquement moins saillants. Leur reprise par un pronom, qui ne partage aucun matériel lexical avec eux, est donc plus difficile.

Dans l'exemple (6) précédent, c'est pourtant cette situation qui se produit. On peut là encore se demander s'il n'existe pas une spécificité des mentions de personnes à favoriser les reprises pronominales. Considérons l'exemple (7) où la référence concerne une inscription et non une entité de personne :

Exemple (7) – Dialogue Accueil_UBS_028_0000001d

Loc 2	*e j'ai en communication là une personne qui a un souci avec e* ***son inscription*** *je te passe*
Loc 1	*un souci avec* ***une inscription***
Loc 2	*ouais*
Loc 1+2	*1 : ouais ben passe-la moi ouais* *2 : d'accord*

Ici, il semble difficile d'imaginer ici une reprise pronominale à la place d'une coréférence directe :

*Loc 1 *un souci avec elle*

L'emploi du pronom privilégierait en effet une référence à l'étudiante et non à son inscription. L'expérience suivante concerne précisément l'ancrage par les entités nommées de personnes.

4.5 Ancrage par une entité nommée de personne et reprise pronominale

Le corpus ANCOR a été annoté en entité nommées, suivant la typologie adoptée par la campagne d'évaluation ESTER2 (Galliano et al. 2009). Cela nous permet d'étudier l'influence du type d'entité nommée sur la réalisation des chaînes. En particulier, nous nous sommes demandé si les noms de personnes (type EN=PERS) ancraient plus fréquemment les chaînes de référence. Nous avons pour cela comptabilisé les relations dont l'antécédent était une entité nommée de personne et comparé une fois encore les résultats obtenus entre le codage en première mention, qui privilégie les ancres, et le codage en chaîne. Nous n'avons pas observé de prédisposition à l'ancrage par les entités de personnes dans les résultats présentés dans le tableau 8. Certes, le nombre total de relations dont l'antécédent est une entité nommée PERS est plus important dans le codage en première mention. Mais cette légère augmentation est équivalente à celle observée pour l'ensemble des GN (cf § 4.1). Au final, l'accroissement observé nous parait trop faible pour être expliqué par une influence autre que celle du type syntaxique GN.

Tableau 8 – *Distribution des relations de coréférence dont l'antécédent est un groupe nominal*

Codage	EN PERS	Autres GN
1ère mention	84,0 %	16,0 %
chaîne	79,3 %	20,7 %

En dépit de ce résultat, on sait que les animés (dont font partie les PERS), constituent une classe particulièrement saillante et sont de ce fait d'excellents topiques[7]. Il est donc particulièrement intéressant de disposer d'une classification en type d'entités nommées comme dans le corpus ANCOR pour évaluer le rôle potentiel joué par les animés (vs. non-animés). Ainsi, étant donné leur (potentielle) place privilégiée parmi les entités saillantes, on s'attend à ce que ces EN soient plus facilement reprises par un pronom. Une interrogation du corpus avec ANCORQI confirme cette prédiction d'une manière très significative (tableau 9).

Quel que soit le codage utilisé (chaîne ou première mention), on observe en effet que les pronoms représentent les 2/3 des reprises après une entité de personnes, alors qu'ils sont minoritaires dans le cas général d'une reprise après un groupe nominal. Les situations de reprise N PERS→PR sont effectivement très nombreux dans le corpus, à l'image de celles données dans l'exemple (6) vu précédemment. Cette différence est légèrement moins marquée, mais toujours aussi significative, dans le cas du codage en

[7] Cette spécificité des animés que semble indiquer ces résultats est validée par des études typologiques (Lyons 1999 : chapitre 9).

première mention : les expérimentations sur la structure interne des chaînes vont nous expliquer cette observation par la surreprésentation des maillons N-N en début de chaîne.

Tableau 9 – *Répartition des relations de coréférence suivant la nature de l'antécédent et de la reprise*

	Antécédent : N		**Antécédent : N PERS**	
Codage	**N→N**	**N→PR**	**N PERS→N**	**N PERS→PR**
1ère mention	54,6%	45,4%	35,2%	64,8%
chaîne	69,5 %	30,5 %	32,6 %	67,4 %

5 Structure interne des chaînes de référence

Comme l'ont suggéré les résultats obtenus sur les groupes prépositionnels ou les indéfinis, la comparaison des distributions entre le codage en chaîne et celui en première mention peut également nous permettre de lever le voile sur la structure interne des chaînes. Nous présentons dans cette section trois expériences qui illustrent ce propos.

5.1 Chaînes et alternance N – P

La première expérience que nous avons menée consiste à étudier dans quelle mesure les relations de coréférence donnaient lieu à une conservation de la catégorie syntaxique, ou partie du discours (POS pour *Part Of Speech*), dans le maillon (relation N-N ou P-P). Ces informations peuvent être obtenues grâce la richesse de l'annotation du corpus, en considérant le POS des mentions reliées et le type de la relation. Cela nous permet de distinguer quatre situations dont la distribution est donnée dans le tableau 10 :

- Maintien d'un codage homogène en POS : directe N-N, indirecte N-N et anaphore P-P
- Transition N-P: anaphore N-P
- Transition P-N : indirecte P-N

Tableau 10 – *Répartition des relations de coréférence suivant les couples de POS antécédent – reprise*

Codage	**N→N**	**P→P**	**N→P**	**P→N**
1ère mention	50,4 %	6,5 %	41,1 %	2,0 %
chaîne	44,8 %	25,7 %	23,6 %	5,9 %

Si l'on s'intéresse au codage en chaîne, on remarque que la recherche d'une chaîne homogène, i.e. le maintien de la partie du discours dans les chaînons, domine (70,5% = 44,8% + 25,7% si l'on regroupe les situations N-N et P-P). L'ancrage d'une chaîne par un pronom étant rare (cf. § 3.1.), celles-ci débutent donc généralement par une mention de type N qui introduit de manière privilégiée un premier chaînon N-N : ce type de séquence est en effet près de deux fois plus fréquent que les transitions N-P (44,8% contre 23,6% des observations. La fréquence non négligeable de ces transitions N-P montre qu'une reprise pronominale à un moment donné de la chaîne reste toutefois fréquente. A l'opposé, la faible proportion (5,9%) de transitions P-N montre que le retour ensuite à une mention nominale initiale n'est pas fréquente, alors que le maintien d'une séquence pronominale (P-P) représente 25,7% des relations.

Au final, ces observations dressent l'image d'une chaîne prototypique de la forme N – N – N – P – P avec 2 chaînons de type N-N suivis d'un chaînon N-P et d'un chaînon P-P. Le décompte des types de chaînons dans une telle chaîne (tableau 11) s'accorde ainsi avec les observations faites sur l'ensemble du corpus.

Tableau 11 – *Répartition des types de chaînons dans une chaîne prototypique N-N-N-P-P*

Codage	**N-N**	**N-P**	**P-P**	**P-N**	**Total**

1ère mention	2 (50%)	2 (50%)	0 (0%)	0 (0%)	**4 (100%)**
Chaîne	2 (50%)	1 (25%)	1 (25%)	0 (0%)	4 (100%)

Les observations faites sur le codage en première mention sont compatibles avec cette analyse : la baisse très significative de la proportion de relations introduites par un pronom rappelle que les chaînes s'ancrent de manière très privilégiées sur un groupe nominal. Et les proportions assez proches de relations de type N-N (50,4% des situations) et N-P (41,1% des cas) en codage en première mention répondent parfaitement à l'image d'une chaîne N – N – N – P – P (même nombre de N-N et de N-P en première mention dans cette chaîne).

Notons bien entendu qu'il ne s'agit que d'un schéma prototypique moyen, mais qui traduit bien l'idée dominante d'une succession de nominaux puis de pronoms, avec peu de retour aux nominaux. Dans l'exemple (8) ci-dessous, on observe un patron de transitions plus complexe avec passage du nom (*monsieur Nom*) au pronom, puis retour à une entité nominale (*Prénom Nom*) et bascule finale sur une chaîne anaphorique à nouveau. Nous retrouvons ici une alternance de noms propres et de pronoms qualifiée de chaîne homogène par (Schnedecker 2005).

Exemple (8) – Dialogue Accueil_UBS_083_00000054

Loc 2	*oui bonjour madame pourrais-je parler à monsieur Nom s'il vous plait*
Loc1	*monsieur Nom*
Loc 2	*oui*
Loc1	*# attendez je vais rechercher je vais regarder s'il est dans bon bureau # et dans quelle filière il travaille*
Loc 2	*il travaille e au niveau informatique*
Loc1	*ah c'est Prénom Nom*
Loc 2	*voilà tout à fait malade*
Loc1	*oui alors attendez lui il est en # en sciences attendez je vais vous le passer*

5.2 Chaînes et définitude

Nous avons également étudié cette tendance à l'homogénéité en matière de définitude Pour cela, nous avons distingué cette fois les situations suivantes :

- Maintien d'un codage homogène en définition : DEF-DEF ou INDEF-INDEF
- Transition INDEF-DEF
- Transition DEF-INDEF

Tableau 12 – *Répartition des relations de coréférence suivant les couples de définition antécédent – reprise, en fonction des deux codages en première mention et en chaîne*

Codage	DEF-DEF	DEF-INDEF	INDEF-INDEF	INDEF-DEF
1ère mention	76,5 %	4,3 %	8,4 %	10,8 %
chaîne	80,6 %	4,6 %	8,1 %	6,6 %

Nos observations, résumées dans le tableau 12, montrent que les trois quarts et plus des relations concernent deux mentions définies (76,5% et 80,6% des observations suivant le codage). La plupart des chaînes doivent donc le plus souvent se résumer à une succession de termes définis (DEF).

Très naturellement, les transitions DEF-INDEF, comme dans l'exemple (9) déjà rencontré plus haut, sont présentes, mais en proportion très réduite (4,6% des observations dans le codage en chaîne) et inférieure aux relations INDEF-DEF (6,6% des cas) dans le codage en chaîne. La comparaison avec le codage en

première mention montre par ailleurs une augmentation significative des relations de type INDEF-DEF (accroissement de 52% entre 6,6% et 10,8% des observations), qui ne peut s'expliquer que par la surreprésentation des indéfinis comme ancres, par opposition à leur présence en milieu de chaîne.

Exemple (9) – Dialogue Accueil_UBS_028_0000001d	
Loc 2	*e j'ai en communication là une personne qui a un souci avec e* <u>*son inscription*</u> *je te passe*
Loc 1	*un souci avec* <u>*une inscription*</u>
Loc 2	*ouais*

Ajouté au maintien du caractère indéfini de la mention dans plus de 8% des relations (8,1% et 8,8% des observations suivant le codage considéré), il ressort de ce tableau de résultats que deux types distincts de chaînes peuvent être observées plus fréquemment :

- un schéma DEF – DEF – … – DEF très majoritaire
- un schéma INDEF – INDEF – … (– DEF), bien plus rare mais présent pour les chaînes introduites par un indéfini. Ces situations mériteront une analyse plus approfondie sur l'ensemble du corpus ANCOR pour obtenir plus d'exemples représentatifs.

6 Influence du codage sur l'accord en genre et nombre

La dernière étude que nous avons menée vise à s'interroger sur l'impact du codage sur certains principes généraux qui semblent gouverner la réalisation de ces chaînes. En particulier, nous nous sommes intéressés à une propriété supposée des relations anaphoriques qui a déjà été questionnée (Antoine 2004, Lefeuvre et al. 2014) : celle de l'accord en genre et nombre entre l'antécédent et sa reprise.

6.1 Accord en genre

Nous n'allons pas ici reprendre en détail les conclusions des études menées par (Lefeuvre et al. 2014), qui servent de base à ce travail. Rappelons simplement les expériences qui ont déjà été menées sur le corpus ANCOR annoté en première mention ont montré que si le genre est arbitraire, il a tout de même une fonction classificatoire qui fait que l'accord en genre est relativement bien respecté en français oral (Lefeuvre et al. 2014).

Tableau 13 – *Taux d'accord en genre en fonction du type de relation de coréférence considérée*

Codage	**directe**	**indirecte**	**pronominale**	***TOTAL***
1ère mention	97,0 %	73,5 %	95,7 %	***94,5 %***
chaîne	96,7 %	84,3 %	95,7 %	***95,8 %***

Le tableau 13 regroupe précisément les résultats de (Lefeuvre et al. 2014), obtenus avec le codage en première mention, auxquels nous avons ajouté ceux observés désormais sur un codage en chaîne. On observe que les taux d'accord restent très proches. Le type de codage n'a donc d'influence qu'à la marge et ne change pas la conclusion principale que l'on peut tirer de ces études : l'arbitraire du genre ne se manifeste réellement qu'avec les relations indirectes où l'on trouve des exemples de désaccords tels que :

[023_00000018] *le bureau … la pièce*

[104_00000069] *une demande de dérogation … le dossier de dérogation*

Au final, il ressort d'une analyse en chaîne ou en première mention que l'accord en genre reste une contrainte opérante en français parlé spontané. Pour une analyse plus fine des situations de non accord avec les relations directes ou pronominales, on consultera (Lefeuvre et al. 2014). Il n'y a ici rien de

surprenant car le genre (ou la classe pour les langues à classes nominales) est bien un élément permettant de suivre les référents dans le discours, comme cela a déjà été remarqué (Huang 2000:8).

On remarque d'ailleurs qu'en créole de la Guadeloupe, lequel partage une vaste majorité de son lexique avec le français (soit par étymologie soit par emprunt récent) mais ne connaît pas de genre grammatical, on trouve beaucoup moins de pronoms que de noms : 70% de N pour 30% de PR (Schang, 2015). Ceci appuie le fait que le genre est un critère efficace pour l'identification du référent : ne portant ni contenu lexical ni indice de genre, le pronom créole est moins susceptible de porter une reprise anaphorique.

6.2 Accord en nombre

Les mêmes conclusions peuvent être faites quant à l'accord en nombre dans les relations coréférence : on observe en effet un taux d'accord encore plus marqué, et ce quel que soit le type de codage choisi (tableau 14). Une augmentation de l'accord en nombre est également observée pour les relations indirectes, et s'explique de la même manière.

Tableau 14 – *Taux d'accord en nombre suivant le type de relation de coréférence considéré*

Codage	directe	indirecte	pronominale	*TOTAL*
1ère mention	92,8 %	81,6 %	95,3 %	***93,2 %***
chaîne	92,1 %	87,1 %	94,7 %	***92,8 %***

Nos études sur l'accord en genre et en nombre montrent que le type de codage (première mention versus en chaîne) n'influe pas sur l'utilité de ces derniers dans les modèles qui gouvernent la résolution de la coréférence. Ces résultats est très important de notre point de vue. Nous espérons pouvoir le généraliser à d'autres heuristiques liées aux chaînes de référence et l'étendre à des analyses expérimentales sur l'ensemble du corpus ANCOR, une fois celui-ci intégralement annoté en première mention.

7 Conclusions générales sur la coréférence

Ainsi que l'ont montré les expériences présentées dans cet article, disposer à la fois d'une annotation en chaîne et en première mention sur une même ressource permet, par comparaison de résultats d'études distributionnelles, de tirer des enseignements sur la structure des chaînes de référence. Dans cet article, nos expérimentations ont permis en particulier de retrouver certains résultats connus dans la littérature, ce qui valide à nos yeux la pertinence de la démarche.

Ces recoupements avec l'état de l'art concernent avant tout les capacités de certains types de mentions (définis, démonstratifs, pronoms etc…) à ancrer (ou non) les chaînes de référence comme première mention, mais également simplement à assurer un rôle d'antécédent au milieu d'une chaîne. De telles observations nous semblent déjà intéressantes et méritent un approfondissement par des études plus fines qui pourrons être menées avec l'outil ANCORQI sur l'ensemble du corpus, une fois celui-ci intégralement annoté également en chaîne.

Plus originales sont certainement les observations que nous avons pu faire, toujours par analyse comparative, sur la configuration globale des chaînes (cf. § 5). Nous pensons en particulier ici à nos études sur les configurations de transition (ou non) entre définis et indéfinis, ou entre groupes nominaux et pronoms. Ce type de résultat peut nous amener à nous interroger sur le type d'annotation qui peut être le plus efficace, dans une perspective de suivi de mentions, pour identifier des heuristiques de résolution ou simplement de modélisation de la coréférence. Clairement, l'annotation en chaîne permet idéalement ce suivi de mentions et de leurs caractéristiques au fil de la chaîne de référence. Nous pensons toutefois avoir montré que l'annotation en première mention, permet, par comparaison, d'éliciter certaines heuristiques qui pourront ensuite être validées par une étude quantitative fine de l'annotation en chaîne.

En quelque sorte, disposer de cette double annotation est un outil très précieux pour la fouille de motifs et l'émergence d'hypothèses sur la réalisation des chaînes de référence.

Références bibliographiques

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

Charolles, M. (2002). *La référence et les expressions référentielles en français*. Editions Ophrys.

Chiarcos, C. (2009). Mental salience and grammatical form: toward a framework for salience metrics in natural language generation (Doctoral dissertation)

Cornish, F. (2011). 'Strict'anadeixis, discourse deixis and text structuring". *Language Sciences*, 33(5), 753-767.

Galliano, S., Gravier, G., Chaubard, L. (2009). The ESTER 2 Evaluation Campaign for the Rich Transcription of French Radio Broadcast. Proc. *Interspeech'2009*, Brighton, UK.

Gundel, J. K., Hedberg, N., & Zacharski, R. (2001). Definite descriptions and cognitive status in English: Why accommodation is unnecessary. *English Language and Linguistics*, 5(02), 273-295.

Hawkins, J. A. (1978). *Definiteness and indefiniteness: a study in reference and grammaticality prediction*. London: Croom Helm.

Huang, Y. (2000). *Anaphora: A cross-linguistic approach*. Oxford University Press.

Karttunen, L. (1976). Discourse referents. In Proceedings of the 1969 conference on computational linguistics (pp. 1-38).

Lefeuvre A., Antoine J.-Y., Schang E. (2014) Le corpus ANCOR_Centre et son outil de requêtage : application à l'étude de l'accord en genre et nombre dans les coréférences et anaphores en français parlé. Actes du *4ème Congrès Mondial de Linguistique Française*, *CMLF'2014*, Berlin, Allemagne.

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

Mitkov, R. (2010). Discourse processing. *The handbook of computational linguistics and natural language processing*, 599-629.

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'Olonne, juin 2013

Recasens M., Martí M.A., Taule M. (2009). First mention definites: More than exceptional cases. *The Fruits of Empirical Linguistics: Products* 2:217.

Schang, E. (2015) Anaphoric chains in Guadeloupean Creole. Présentation à la *Society for Pidgin and Creole Studies*, Graz, juillet 2015.

Schnedecker, C. (1997). *Nom propre et chaînes de référence*. Klincksieck

Schnedecker, C. (2005). Les chaînes de référence dans les portraits journalistiques : éléments de description. *Travaux de linguistique*, 51(2), 85 :133.

Schnedecker, C., & Landragin, F. (2014). Les chaînes de référence: présentation. *Langage*, n°195.

Walker, M. A., Joshi, A. K., & Prince, E. F. (1998). *Centering theory in discourse*. Oxford University Press.