Les chaînes coréférentielles en créole de la Guadeloupe

Emmanuel Schang, Jean-Yves Antoine, Anaïs Lefeuvre-Halftermeyer

Les chaînes coréférentielles en créole de la Guadeloupe

Emmanuel Schang [1] Jean-Yves Antoine [2] Anaïs Lefeuvre-Halftermeyer [3]
(1) LLL (UMR 7270), Université d'Orléans/CNRS - 10 Rue de Tours 46527 Orléans cedex 2, France
(2) LI, Université François Rabelais Tours, 3 place Jean Jaurès, 41000 Blois, France
(3) LIFO, Université d'Orléans - 6, rue Léonard de Vinci 45067 Orléans Cedex 2, France
emmanuel.schang@univ-orleans.fr,
jean-yves.antoine@univ-tours.fr,anais.halftermeyer@univ-orleans.fr

RÉSUMÉ

Cet article présente une étude des chaînes coréférentielles en créole de la Guadeloupe à partir d'un corpus oral annoté en relations coréférentielles (1096 relations coréférentielles). Nous contrastons les résultats de notre expérimentation avec les données observées pour le français sur le corpus ANCOR (Muzerelle *et al.*, 2013). Bien que ces deux langues partagent largement leur lexique, elles diffèrent significativement sur le plan grammatical (absence de genre, SN sans déterminants en créole), ce qui donne toute sa valeur à leur comparaison.

ABSTRACT

Coreference Chains in Guadeloupean Creole

This paper presents a study on coreference chains in Guadeloupean Creole. A corpus of spoken Guadeloupean Creole was annotated in coreference relations (1096 relations) and we have compared the results with those found in the ANCOR corpus (French). While French and Gwadloupéyen share most of their lexicon, they differ greatly in their grammars (no gender, bare NPs in Creole *inter alia*), which makes their comparison valuable.

MOTS-CLÉS : Coréférence, créole, gwadloupéyen, corpus, annotation.

KEYWORDS: Coreference, Guadeloupean Creole, annotation, corpus.

1 Introduction

Le créole français de Guadeloupe (gwadloupéyen, ou CG) tire l'essentiel de son lexique du français [1] mais sa syntaxe diffère du français, notamment sur les points suivants :

— l'absence de genre grammatical,
— la forte utilisation des SN sans déterminants.

Le genre et le nombre de l'antécédent ainsi que le type d'article sont habituellement considérés comme des traits importants pour les systèmes de résolution des coréférences par aprentissage automatique développés par le TAL (Recasens *et al.*, 2011; Recasens Potau, 2010; Désoyer *et al.*, 2015). La description des stratégies coréférentielles en créole est intéressante car nous disposons d'un point de comparaison avec le français qui ne met pas en question les relations lexicales, mais uniquement la grammaire. Dans ce papier, nous aborderons ces questions au travers de deux points :

1. Par des apports consécutifs depuis le 17ème siècle et sous l'influence d'un important bilinguisme, v. (Bernabé, 1983) entre autres.

— la mise en évidence de chaînes coréférentielles prévalentes en créole,
— l'utilisation des syntagmes nominaux sans déterminants (SNSD).

Dans un premier temps, nous présenterons succinctement les principaux points de la grammaire du gwadloupéyen qui seront pertinents pour la suite de l'exposé. Puis nous présenterons la méthode suivie pour l'annotation du corpus et enfin, nous détaillerons quelques résultats.

2 Quelques éléments de la grammaire du gwadloupéyen

Le gwadloupéyen se distingue du français par l'utilisation de marqueurs de temps, mode et aspects (TMA) antéposés au verbe (Damoiseau, 2012; Bernabé, 1983; Vaillant, 2008) et qui se combinent avec la négation et certains adverbiaux à la place de la flexion verbale du français. Le verbe ne s'accorde pas avec son sujet. Il n'y a pas de passif.

Dans le domaine nominal, qui touche directement à la question des expressions référentielles, on trouve principalement les différences suivantes :

Absence de genre : Il n'y a pas de genre grammatical en CG, comme l'attestent les exemples en (1).

(1) a. on nonm "un homme" b. on fanm "une femme"
c. ti gason-la bèl "le petit garçon est beau" d. tifi-la bèl "la petite-fille est belle"

Défini et démonstratif postposés : Contrairement au français, l'article défini et le démonstratif (2-a) se trouvent en marge droite du SN [2] et le nombre, lorsqu'il est marqué, est toujours accompagné du défini (2-b) :

(2) a. vwati wouj -la/lasa
voiture rouge DEF/DEM
'La/cette voiture rouge'
b. sé timoun-la (*sé timoun)
PL personne-DEF
'les enfants'

Un pluriel marqué renvoie donc à des entités 'identifiables' dans le contexte (non génériques) car l'article défini *-la* véhicule la notion de 'spécificité'. Nous continuerons cependant à nommer celui-ci 'défini', conformément à l'usage établi.

SN sans déterminants : On trouve dans la grande majorité des langues créoles des syntagmes nominaux sans déterminants (SNSD) avec un usage référentiel (Baptista & Guéron, 2007). Les SNSD véhiculent plusieurs valeurs sémantiques différentes, décrites, pour le GC dans (Gadelii, 2007).

2. Le défini et le démonstratif sont habituellement graphiés précédés d'un tiret.

(3) a. fo ou achté épis a kolombo
il.faut 2SG acheter épice à Colombo
'il faut que tu achètes des épices à Colombo.'
b. i ka fè kolombo ban nou
3SG IPFV faire Colombo pour 1PL
'il fait le/un colombo pour nous'

Nous ne pouvons pas détailler, faute de place, les exemples sans entrer dans une longue discussion pour chacun d'eux en exposant quels contextes rendent leur usage possible. On retiendra que les SNSD ont un usage beaucoup plus étendu que les SN sans déterminants en français (pour lesquels on pourra lire (Bouchard, 2003)). L'usage des SNSD pour des expressions singulier ou pluriel (identifiables par des pronoms de reprise au singulier ou au pluriel : *i* ou *li* vs *yo*) est un trait majeur à prendre en compte.

On retiendra de cette section que le gwadloupéyen ne permet pas d'utiliser l'accord en nombre [3] et en genre comme trait pour la résolution des coréférences, bien que le lexique soit semblable à celui du français. Nous avons conduit une première étude comparative entre le français et le créole en corpus pour appréhender l'impact de ces particularités.

3 Méthodologie

Le corpus ANCOR-Centre (Antoine *et al.*, 2016; Lefeuvre *et al.*, 2014; Muzerelle *et al.*, 2013) est un corpus annoté en relations anaphoriques et coréférentielles sur du français oral. Il est composé principalement d'entretiens et comprend environ 115 000 mentions et 51 500 relations anaphoriques. La méthodologie qui a présidé à la constitution de ce corpus est exposée en détails dans (Antoine *et al.*, 2013).

Le corpus ANCOR-971 est un corpus constitué à partir de la même méthodologie que le corpus ANCOR, mais sur des enregistrements de créole guadeloupéen (entretiens libres) disponibles en ligne (Glaude, 2013). Les points de divergence sont les suivants :

— les catégories annotées pour les mentions (SNSD (noté Bare), Indéfini, Défini, Démonstratif),
— annotation des principales fonctions syntaxiques (sujet, objet, attribut notamment).

Les mentions (ainsi que les relations) ont été annotées par un locuteur natif du gwadloupéyen et vérifiées par un linguiste spécialiste de cette langue. La plateforme d'annotation utilisée était GLOZZ (Widlöcher & Mathet, 2012) et les résultats ont été exploités avec les outils d'interrogation GLOZZ-QL, ANCOR-QI (Lefeuvre *et al.*, 2014) notamment. Concernant les liens coréférentiels uniquement [4], les relations annotées étaient les suivantes :

— directe : la reprise et l'antécédent ont la même tête lexicale,
— indirecte [5] : la reprise et l'antécédent ont la même tête lexicale,
— anaphore : reprise par un pronom.

3. Tout au moins de façon similaire à ce qui se fait en français.

4. Les autres relations, qui correspondent à des anaphores associatives, ne seront pas prises en compte dans cet article car elles ne portent pas sur la coréférence.

5. Aussi appelée 'anaphore infidèle'.

4 Résultats

Le corpus 971 contient 2731 entités et 1225 relations (dont 1096 relations coréférentielles), ce qui représente une base de travail considérablement plus large par rapport aux études antérieures sur le créole. A titre d'exemple, (Gadelii, 2007) a travaillé sur un corpus contenant cent exemples.

4.1 Relations

(Antoine *et al.*, 2016) ont étudié les chaînes coréférentielles en français oral dans les situations propres au corpus ANCOR (interaction orale spontanée en situation de dialogue finalisé). Ils ont observé que les chaînes répondaient à un patron prototypique (c'est-à-dire au sens de la chaîne la plus probable distributionnellement) N-N-N-P-P [6]. Cela signifie que, typiquement, une chaîne s'ancre sur un nom, suivi de deux reprises nominales puis de reprises pronominales.

Comparons ces données avec celles du gwadloupéyen.

4.1.1 Ancrage des chaînes

La proportion de SN par rapport aux pronoms est plus importante en créole qu'en français, comme le montre le Tableau 1.

	N	Pr
Français	51.2%	48.8%
Créole	70% (1901)	30% (819)

TABLE 1 – Proportion de SN vs Pronoms dans les deux langues

Les chaînes coréférentielles sont principalement ancrées par des N même si la proportion de chaînes ancrées par des pronoms (20%) est plus forte qu'en français (7% dans le corpus ANCOR).

Les 59 cas d'usage de pronoms comme ancre proviennent du pronom pluriel *yo* 'ils, eux' dans un usage indéfini, ou du pronom singulier *sa* ayant une référence mal délimitée (abstraite au sens de (Dipper & Zinsmeister, 2012)) et un seul cas de cataphore.

4.1.2 Structure des chaînes

Les chaînes coréférentielles du corpus créole sont en moyenne plus courtes que celles trouvées dans le corpus français. En créole, elles sont en moyenne légèrement en dessous de 3 maillons (2,87) alors que dans le corpus français, elles sont en moyenne constituées de 4 maillons (Antoine *et al.*, 2016).

L'analyse avec ANCORQI de la distribution des relations partant d'une ancre, ou internes à la chaîne, nous fournit des graphes de transitions comme la figure 1, qui concerne les chaînes à ancre nominale.

6. De façon à simplifier l'argumentation, nous parlerons de N à place de SN.

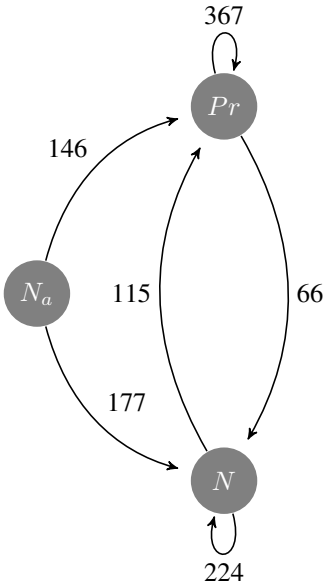

FIGURE 1 – Chaînes coréférentielles ancrées par un N (N_a) par type et en nombre de relations.

L'analyse des transitions de la figure 1, de même que la connaissance du nombre moyen de relations dans une chaîne, montre que les chaînes nominales prévalentes en créole sont du type N-P-P-P ou N-N-N-N. Quant aux chaînes ancrées par un pronom (30% des cas), elles sont constituées très majoritairement d'une séquence P-P-P-P de pronoms. Comme en français, lorsque dans une chaîne on utilise un pronom, il est rare de revenir au N par la suite.

4.2 Pronoms et SNSD

On notera (Table 2) que les SN sont en majorité des Syntagmes Nominaux Sans Déterminants (SNSD). Ceux-ci ont fait l'objet de nombreuses études dans le cadre des langues créoles (Baptista & Guéron, 2007). Dans (Gadelii, 2007), on trouve l'idée que les SNSD du CG sont principalement des sujets reprenant un topique (continued topics).

Cette hypothèse peut se vérifier facilement sur le corpus en croisant les traits annotés : nouvelle entité du discours (non), fonction (sujet) et type (SNSD). Il apparaît cependant que les SNSD apparaissent assez peu en sujet (Table 3), et en proportion, beaucoup moins que les pronoms personnels (3SG et 3PL) dont c'est la fonction privilégiée. Par ailleurs, il est à noter que contrairement aux pronoms personnels, la majorité des SNSD apparaissent comme nouvelle entité du discours (646 ancres contre 448 reprises) et que lorsqu'ils sont en reprise, ils n'apparaissent pas majoritairement en sujet (Table 4). L'hypothèse avancée dans (Gadelii, 2007) n'est donc pas confirmée par ce corpus.

Pour aller un peu plus loin dans l'analyse des SNSD par rapport aux pronoms, on pourra remarquer que le nombre d'entités du discours (mentions) entre deux membres d'une chaîne dont le second terme est un pronom (N-P ou P-P) est bien inférieur à celui des chaînes dont le second terme est un SNSD (N-SNSD ou P-SNSD). La moyenne du nombre d'entités entre un pronom et son antécédent

SNSD	définis	indéfinis	démonstratifs
56,5% (1074)	21,2% (403)	20,78% (395)	1,52% (29)

TABLE 2 – Distribution des SN par type.

Type	sujet	objet	autre
SNSD	9,5% (102)	28,5% (306)	62% (666)
defini	19,35% (78)	23,82% (96)	56,82% (229)
pronom pers.	56,93% (115)	7,92% (16)	35,15% (71)

TABLE 3 – Fonctions grammaticales des SNSD, SN définis et pronoms

Type	sujet	objet	autre
SNSD reprise	14,02% (60)	21,96% (94)	64,02% (274)
SNSD ancre	6,58% (42)	33,23% (212)	60,19% (384)

TABLE 4 – Fonctions grammaticales des SNSD (reprise ou ancre)

est de 3.07 (médiane = 1) tandis que pour les SNSD, elle est de 8,83 (médiane = 2). Ce qui indique que la résolution des pronoms (toutes catégories confondues) est locale par rapport aux SNSD.

On peut faire l'hypothèse raisonnable que, en l'absence de genre grammatical, les pronoms sont utilisés pour reprendre une entité du discours immédiatement disponible et, dès lors qu'une ambiguïté est possible, la reprise par un SNSD est préférée.

On remarque par ailleurs que les SN définis [7] trouvent leur antécédent à une distance (en nombre de mentions) moyenne de 14 (médiane = 4). On le voit donc, ils reprennent en moyenne des mentions situées plus loin que les SNSD.

Ceci est compatible avec les théories de l'accessibilité des référents (Gundel, 2010; Gundel *et al.*, 2003; Gundel, 2003). On peut proposer l'échelle d'accessibilité suivante (accessibilité du référent en fonction de la distance par rapport à l'antécédent) :

++ accessible : pronom
+ accessible : SNSD
- accessible : SN défini

5 Conclusion

Bien qu'il soit impossible de détailler tous les paramètres annotés dans le corpus de créole guadeloupéen dans cet article, cette étude préliminaire, qui demande à être complétée, nous a permis de mettre en avant quelques faits intéressants :

— bien que le créole ne fasse pas usage du genre grammatical, on retrouve deux points communs avec le français : lorsque dans une chaîne coréférentielle un pronom est utilisé, il est rare de poursuivre la chaîne avec un N,
— le créole fait usage de SN sans déterminants pour reprendre des entités moins proches (distance calculée en nombre de mentions entre l'antécédent et la reprise) que les pronoms. On peut faire

7. Les SN indéfinis étant principalement des ancres, ils ne sont pas pertinents pour notre analyse.

l'hypothèse que l'absence de genre en créole favorise la reprise par un SN sans déterminant (accessibilité du référent par la tête nominale).

Les résultats de cette étude nous conduisent à envisager de développer ces mêmes investigations pour comparer d'autres langues créoles avec leur langue lexificatrice. Par exemple, la comparaison des créoles portugais du Golfe de Guinée ou de Haute-Guinée (forro et kriyol respectivement) avec le portugais serait intéressante car ceux-ci diffèrent du portugais (la langue qui a fournit l'essentiel de leur lexique) par l'absence de genre grammatical. Ces études devraient conduire à une meilleure compréhension à la fois de la créolisation [8] (v. entre autres (Mufwene, 2005)) et des processus de résolution des coréférences en général. Enfin, ce travail contribue à la prise en compte de langues dites 'peu dotées' dans les réflexions linguistiques générales et dans le traitement automatique des langues naturelles.

Remerciements

Les auteurs tiennent à remercier : Laura Noreskal pour ses annotations, Flora Badin poour ses scripts, les membres du GDRI SEEPiCLa pour leurs retours sur des versions préliminaires du texte ainsi que les relecteurs anonymes de DILITAL.

Références

ANTOINE J.-Y., LEFEUVRE A. & SCHANG E. (2016). Codage en chaîne ou en première mention de la coréférence : Approcher la structure des chaînes de référence par comparaison des deux annotations. *SHS Web of Conferences*, **27**(nil), 02001.

ANTOINE J.-Y., SCHANG E., MUZERELLE J., LEFEUVRE A., PELLETIER A., ESHKOL I., MAUREL D. & VILLANEAU J. (2013). *Corpus ANCOR_Centre*. Rapport interne.

BAPTISTA M. & GUÉRON J. (2007). *Noun phrases in creole languages : a multi-faceted approach*, volume 31. John Benjamins Publishing.

BERNABÉ J. (1983). *Fondal-natal*. l'Harmattan Paris.

BOUCHARD D. (2003). Les sn sans déterminant en français et en anglais. *Essais sur la grammaire comparée du français et de l'anglais*, p. 55–95.

DAMOISEAU R. (2012). *Syntaxe créole comparée*. Karthala et CNDP-CRDP edition.

DÉPREZ V. (2005). Morphological number, semantic number and bare nouns. *Lingua*, **115**(6), 857–883.

DÉSOYER A., LANDRAGIN F., TELLIER I., LEFEUVRE A. & ANTOINE J.-Y. (2015). Coreference Resolution for Oral Corpus : a machine learning experiment with ANCOR corpus. *Traitement Automatique des Langues*, **55**(2), 97–121.

DIPPER S. & ZINSMEISTER H. (2012). Annotating abstract anaphora. *Language Resources and Evaluation*, **46**(1), 37–52.

GADELII K. (2007). The bare np in lesser antillean. *Creole Language Library*, **31**, 243.

GLAUDE H. (2013). *Corpus Créoloral*. oai :crdo.vjf.cnrs.fr :crdo-GCF, SFL Université Paris 8 - LLL Université Orléans.

8. On entend par *créolisation* les forces qui produisent une langue nouvelle née de contacts entre plusieurs langues.

GUNDEL J. K. (2003). Information structure and referential givenness/newness : How much belongs in the grammar. In *Proceedings of the HPSG'03 Conference*, p. 143–162.

GUNDEL J. K. (2010). Reference and accessibility from a Givenness Hierarchy perspective. *International Review of Pragmatics*, **2**(2), 148–168.

GUNDEL J. K., HEGARTY M. & BORTHEN K. (2003). Cognitive status, information structure, and pronominal reference to clausally introduced entities. *Journal of Logic, Language and Information*, **12**(3), 281–299.

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. In *SHS Web of Conferences*, volume 8, p. 2691–2706 : EDP Sciences.

MANUELIAN H. & FATTIER D. (2011). L'utilisation des déterminants en créole haïtien : Etude de quelques chaînes de référence.

MUFWENE S. S. (2005). *Créoles, écologie sociale, évolution linguistique*. Editions L'Harmattan.

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 de TALN*, p. 555–563.

RECASENS M., HOVY E. & MARTÍ M. A. (2011). Identity, non-identity, and near-identity : Addressing the complexity of coreference. *Lingua*, **121**(6), 1138–1152.

RECASENS POTAU M. (2010). Coreferència : Teoria, anotació, resolució i avaluació.

VAILLANT P. (2008). Grammaires factorisées pour des dialectes apparentés. In *TALN 2008 : Actes de la 15ème conférence annuelle sur le Traitement Automatique des Langues Naturelles*.

WIDLÖCHER A. & MATHET Y. (2012). The Glozz platform : a corpus annotation and mining tool. In *Proceedings of the 2012 ACM symposium on Document engineering*, p. 171–180 : ACM.

ZRIBI-HERTZ A. & JEAN-LOUIS L. (2013). From noun to name : definiteness marking in modern martinikè. *Crosslinguistic studies on Noun Phrase structure and reference*, p. 269–315.