Typologie de chaînes de référence à la lumière de corpus annotés diversifiés

Silvia Federzoni

Typologie de chaînes de référence à la lumière de corpus annotés diversifiés

Silvia Federzoni[1]
(1) Laboratoire CLLE (CNRS-UMR 5263), Université de Toulouse 2 - Jean Jaurès, 5 allées Antonio Machado, 31058 Toulouse
silvia.federzoni@univ-tlse2.fr

RÉSUMÉ

Ce projet de thèse a pour objectif la définition d'une typologie des chaînes de référence basée sur une description systématique des enchaînements des expressions référentielles dans différents corpus annotés en chaînes de référence.

ABSTRACT

Typology of reference chains in the light of diverse annotated corpora

This thesis project aims to define a typology of reference chains based on a systematic description of the sequences of referential expressions in different corpora annotated in reference chains.

MOTS-CLÉS : chaînes de référence, corpus annotés, typologie, continuité référentielle, discours .

KEYWORDS: reference chains, annotated corpora, typology, referential continuity, discourse.

Notre thèse [1] porte sur les chaînes de référence. Ce projet s'inscrit dans un contexte de recherche dynamique pour le français, comme le montrent les numéros de revue qui ont été récemment consacrés aux chaînes de référence : numéro 195 de *Langue Française* (Schnedecker *et al.*, 2017), numéro 25 de *Discours* (Sarda & Vigier, 2019) et le numéro 72 des *Cahiers de praxématique* (Gardelle *et al.*, 2019). Les *chaînes de référence* (désormais CR) sont des structures discursives qui regroupent plusieurs propositions ayant en commun un même référent – être humain, entité abstraite ou événement – signalé dans les textes par le biais d'expressions linguistiques – noms propres, syntagmes nominaux, pronoms – appelées *expressions référentielles*, *mentions* ou *maillons* (Corblin (1995) ; Charolles (1988) ; Schnedecker (1997), inter alia). Les maillons sont liés par une relation dite de *coréférence*, et leur succession dans les textes contribue à créer des liens de cohésion entre différents segments de discours. De ce fait, les CR constituent un mécanisme fondamental dans l'organisation et l'interprétation du discours. Pour cette raison, elles ont fait l'objet de nombreuses études. En TAL, les efforts ont principalement porté sur le développement et l'amélioration des systèmes de détection et résolution automatique de l'anaphore et de la coréférence (Recasens & Hovy (2010); Mitkov (2014); Oberle (2019); Désoyer *et al.* (2015); Brassier *et al.* (2018), inter alia).

En linguistique comme en TAL, les travaux portant sur les CR se fondent sur l'exploration de corpus annotés. Bien que des ressources de grande taille soient disponibles, aussi bien pour l'anglais (cf. Poesio *et al.* (2016)), comme OntoNote (Pradhan *et al.*, 2012) ou WikiCoref (Ghaddar & Langlais, 2016), que pour le français écrit, comme ANNODIS (Péry-Woodley *et al.*, 2011), Democrat (Lattice *et al.*, 2019), elles n'ont pas permis, jusqu'à présent, de mettre au jour une définition complète et systématique des CR. En effet, ces ressources ont été conçues pour répondre à des objectifs différents

1. Encadrée par Cécile Fabre et Lydia-Mai Ho-Dac.

et rassemblent donc des annotations différentes. Si les auteurs concordent sur la definition de CR comme « la suite des expressions d'un texte entre lesquelles l'interprétation construit une relation d'identité référentielle » (Corblin, 1995) ainsi que sur le nombre minimum de maillons qui doit être de trois [2] (Schnedecker, 2019), il n'y a pas de consensus sur la taille maximale d'une CR. Une autre question qui fait débat dans la littérature est de savoir quels sont les éléments aptes à constituer les maillons d'une CR et comment les délimiter, car un choix s'impose entre prendre en considération uniquement la tête lexicale ou bien inclure ses dépendants. De plus, la prise en compte des singletons (référents qui ne font pas l'objet d'une réprise référentielle) varie d'une annotation à l'autre : certains considèrent qu'il est indispensable de les annoter pour que les systèmes de résolution soient en mesure de les détecter. D'autres, ayant des objectifs différents, ne les annotent pas.
Ce manque de consensus se traduit dans les ressources existantes, conçues sur des modèles linguistiques différents, par une grande hétérogénéité en termes de choix d'annotation, ce qui rend les résultats obtenus difficilement comparables. De même, au vu de cette hétérogénéité, toutes les applications TAL ne peuvent pas exploiter n'importe quelle ressource, et les architectures des sytèmes de résolution développés sont dépendantes du type de ressource sur laquelle le modèle a été entraîné.

À cette difficulté, s'ajoute la complexité du phénomène des CR, dont l'analyse requiert la prise en compte des configuarations d'indices (Das & Taboada, 2019). Par conséquent, aucune étude à large échelle, n'a proposé une description systématique des CR dans leur complexité et leur complétude. Pour l'anglais, la plupart de travaux se focalise sur les paires coréférentielles, sans analyser les CR complètes. Pour le français, les études sur les CR ont été effectuées sur des corpus de petite taille ou échantillons de texte (Schnedecker & Longo, 2012), parfois en se focalisant sur un type de référent particulier.

Dans ce contexte, un premier objectif de notre thèse est de fournir une typologie des CR. Pour y parvenir, il s'agit préalablement d'unifier les corpus annotés afin de fournir une description, la plus exhaustive possible, de la complexité et de la variété des CR. À partir de cette typologie, la thèse proposera une étude contrastive entre différents types de textes ainsi qu'une description systématique qui puisse être exploitée pour l'amélioration d'un modèle de prédiction automatique des CR.

L'idée est de concevoir un modèle qui nous permette de considérer les enchaînements des maillons et d'aller au-délà d'un traitement par paire "antécédent-anaphorique", tout en prenant en compte les traits linguistiques et les variations qui peuvent exister entre différents genres textuels ou différents niveaux d'expertise rédactionnelle. L'application de ce modèle consentira de faire émerger les traits linguistiques ayant une influence sur la présence d'une expression référentielle donnée. L'application de ce modèle nous permettra également de confronter certaines hypothèses cognitives à la réalité des usages en corpus (en particulier les théories de l'accessibilité (Ariel, 2001) et du centrage (Walker *et al.*, 1998)). Nous serons ainsi en mesure d'évaluer systématiquement les décalages entre ce qui est théoriquement attendu en matière de chaînes de référence et ce qui est effectivement attesté dans les usages réels. L'analyse de ces décalages nous permettra de mieux caractériser les traits linguistiques que les systèmes de résolution devraient/pourraient apprendre pour améliorer leurs performances.

À ce stade, nous n'avons pas encore établi la procédure nécessaire au développement de ce modèle. Notre connaissance des modèles existants doit d'abord être approfondie. Nous explorerons ensuite l'intérêt des modèles comme le CRF (Kudo, 2005 cité par Godbert & Benoit (2017)), ou bien des modèles bayésiens, en prenant en compte une diversité de traits pour évaluer ceux qui sont les plus discriminants.

2. À default on parle d'*anaphore* ou de *coréférence* (Schnedecker, 2019)

Références

ARIEL M. (2001). Accessibility theory : An overview. *Text representation : Linguistic and psycholinguistic aspects*, **8**, 29–87.

BRASSIER M., PURET A., VOISIN-MARRAS A. & GROBOL L. (2018). Classification par paires de mention pour la résolution des coréférences en français parlé interactif. *Conférence jointe CORIA-TALN-RJC 2018*.

CHAROLLES M. (1988). Les plans d'organisation textuelle : périodes, chaînes, portées et séquences. *Pratiques*, **57**(1), 3–13.

CORBLIN F. (1995). *Les formes de reprise dans le discours. Anaphores et chaînes de référence*. Presses Universitaires de Rennes.

DAS D. & TABOADA M. (2019). Multiple signals of coherence relations. *Discours. Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics*, (24).

DÉSOYER A., LANDRAGIN F. & TELLIER I. (2015). Machine Learning for Coreference Resolution of Transcribed Oral French Data : the CROC System. *Vingt-deuxième Conférence sur le Traitement Automatique des Langues Naturelles*, p. 439–445.

GARDELLE L., ROSSI C. & VINCENT-DURROUX L. (2019). La gestion de l'anaphore en discours : complexités et enjeux. *Cahiers de praxématique*, (72).

GHADDAR A. & LANGLAIS P. (2016). WikiCoref : An English Coreference-annotated Corpus of Wikipedia Articles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia : European Language Resources Association (ELRA) European Language Resources Association (ELRA).

GODBERT E. & BENOIT F. (2017). Détection de coréférences de bout en bout en français.

LATTICE, LILPA, ICAR & IHRIM (2019). Democrat. ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr.

MITKOV R. (2014). *Anaphora resolution*. Routledge.

OBERLE B. (2019). Détection automatique de chaînes de coréférence pour le français écrit. *Conférence sur le Traitement Automatique des Langues Naturelles (TALN-RECITAL) 2019*.

PÉRY-WOODLEY M.-P., AFANTENOS S., HO-DAC L.-M. & ASHER N. (2011). La ressource ANNODIS, un corpus enrichi d'annotations discursives. *Traitement Automatique des Langues*, **52**(3), 71–101.

POESIO M., PRADHAN S., RECASENS M., RODRIGUEZ K. & VERSLEY Y. (2016). Annotated corpora and annotation tools. In *Anaphora Resolution*, p. 97–140. Springer.

PRADHAN S., MOSCHITTI A., XUE N., URYUPINA O. & ZHANG Y. (2012). Conll-2012 shared task : Modeling multilingual unrestricted coreference in ontonotes. p. 1–40.

RECASENS M. & HOVY E. (2010). Coreference resolution across corpora : Languages, coding schemes, and preprocessing information. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, p. 1423–1432.

SARDA L. & VIGIER D. (2019). *Discours. Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics*, (25).

SCHNEDECKER C. (1997). *Nom propre et chaînes de référence*. Recherches linguistiques. Librairie Klincksieck.

SCHNEDECKER C. (2019). De l'intérêt de la notion de chaîne de référence par rapport à celles d'anaphore et de coréférence. *Cahiers de praxématique*, (72).

SCHNEDECKER C., GLIKMAN J. & FRÉDÉRIC L. (2017). Les chaînes de référence en corpus. *Langue française*, (195).

SCHNEDECKER C. & LONGO L. (2012). Impact des genres sur la composition des chaînes de référence : le cas des faits divers. *3ème Congrès Mondial de Linguistique Française*, p. 1957–1972.

WALKER M. A., JOSHI A. K. & PRINCE E. F. (1998). *Centering theory in discourse*. Oxford University Press.