Les chaînes de référence : annotation, application et questions théoriques

Catherine Schnedecker, Julie Glikman, Frédéric Landragin

Les chaînes de référence : annotation, application et questions théoriques

Draft auteurs (avant corrections et mise en forme de l'éditeur)

Catherine Schnedecker
LiLPa, Université de Strasbourg

Julie Glikman
LiLPa, Université de Strasbourg

Frédéric Landragin
Lattice, CNRS, ENS, Université de Paris 3, Université Sorbonne Paris Cité, PSL Research University

Résumé

Dans cet article de présentation du numéro, nous commençons par définir la notion de *chaîne de référence*, puis nous proposons une revue des méthodes et des approches de l'étude des chaînes de référence. Nous montrons que les approches linguistiques prédominantes ne suffisent pas à prendre en considération les spécificités discursives et les liens étroits entre chaînes et genres textuels. Nous montrons également que les approches de la linguistique de corpus et du traitement automatique des langues suivent des voies opérationnelles qui s'éloignent de la complexité du phénomène linguistique initial et se focalisent sur d'autres langues que le français. Nous mettons alors en avant la démarche qui a réuni les auteurs des articles de ce numéro et qui est en partie matérialisée dans le projet ANR DEMOCRAT.

Mots-clés : référence, chaîne de référence, corpus, méthodologies, applications

Abstract: Reference Chains: Annotation, Application and Theoretical Questions

This volume introductory article starts with a definition of the notion of *reference chain*, and continues with an overview of the methods and approaches for the study of reference chains. We show that the predominant linguistic approaches are not sufficient to take into account the discursive specificities of reference chains, as well as the links between them and text genres. We also show that corpus linguistics and mostly natural language processing approaches are more and more operational and sometimes forget the complexity of the initial linguistic phenomena. Moreover, they mainly focus on languages other than French. Then, we highlight the demarche that brought together the authors of these volume articles, which is partly implemented in the DEMOCRAT ANR project.

Keywords: reference, reference chain, corpus, methodologies, applications

L'étude des chaînes de référence est une question qui a été soulevée tant par les approches de type strictement linguistique, sur une ou plusieurs langues, que par les approches de traitement automatique des langues (TAL). L'intérêt de ce volume, et du groupe de recherche dont il est en partie issu, est la confrontation et le travail commun sur ces deux aspects. Nous commencerons ici par définir la notion de *chaîne de référence*, avant de faire un état de la recherche dans ce domaine, tant en linguistique qu'en TAL. Cet état des lieux sera l'occasion de mettre en avant les problèmes, les enjeux et les lacunes sur le sujet, que ce numéro cherche en partie à combler. Nous montrons ainsi les aspects communs entre les domaines, notamment la question des corpus. Cet article de présentation rappelle également le cadre dans lequel ce numéro s'est élaboré, avant de présenter les contributions qui s'inscrivent dans trois axes. Le premier a trait aux questions d'annotation dont il aborde les enjeux et présente des applications ; le second porte sur le français contemporain et propose des analyses linguistiques dans des cadres théoriques différents ; le troisième est également dédié à des analyses qui opèrent sur le français plus ancien, qui pose des problèmes d'annotation et d'analyse distincts de ceux du français contemporain.

1. Définition

Une chaîne de référence (désormais CR) est la suite des expressions coréférentielles d'un texte, en gras dans l'exemple (1).

1. **Le prodige** (= Brad Mehldau) nous enchante depuis une vingtaine d'années. A 45 ans, le compteur dépasse les 30 albums. Combien de festival ai-je couverts, dont le concert qui m'a le plus renversé était le sien? Un confrère me rapportait récemment celui de Tourcoing (automne 2015), où, en hommage, **Brad** céda la vedette à Lee Konitz pour jouer **lui-même** la première partie : prestation à tomber trois fois à genoux. **L'explorateur cérébral** (je pique la formule au *TIME*), projette nos sens sur 32 compositions allant de Paul Mc Cartney (sublime *Blackbird*), Nirvana, Brahms, Pink Floyd, Beach Boys, Radiohead, à Monk. Sur scène, **l'artiste** déborde de générosité. Humainement, dans le contact, **il** émeut. Chaque fois que je **l'**ai approché, **sa** gentillesse m'a touché. Personne ne mérite de se retrouver privé de pareils sommets. Un des disques de la décennie. (http://jazz.blogs.liberation.fr/2015/12/23/sous-le-sapin-les-jazz/)

Comme on parle, métaphoriquement, de « chaîne », chacune de ces expressions référentielles constitue un « maillon » de la chaîne. Selon l'approche choisie, certains auteurs utilisent le terme de *chaîne de référence*, d'autres celui de *chaîne de coréférences* (avec également des variations quant au nombre de « références » ou de « coréférences »). Le premier terme s'impose en linguistique, le second est surtout utilisé en traitement automatique des langues, dans la mesure où ce domaine considère souvent une coréférence comme un couple, donc un ensemble de deux expressions coréférentielles, et une chaîne comme le regroupement deux à deux de plusieurs couples ayant un maillon commun. L'objet linguistique étant le même, nous utiliserons indifféremment les deux termes.

2. Méthodes et approches pour l'étude des chaînes de référence

2.1. Linguistique

Parmi les nombreux travaux traitant de l'expression de la référence en linguistique, trois types d'approches prédominent actuellement :

- des approches, souvent mais pas exclusivement monographiques, en sémantique grammaticale référentielle visant à décrire le contenu descriptif et/ou les instructions délivrées par les différentes catégories d'expressions référentielles prenant en compte la catégorie grammaticale de la tête du syntagme (nom ou pronom) et de son déterminant (défini, démonstratif, etc.), et qui portent aussi bien sur le français contemporain ou sur le français ancien, sur d'autres langues que le français (p. e. : l'anglais, le néerlandais,…) ou sur la comparaison de paires de langues dans une optique contrastive (Vanderbauwhede, 2012) ;
- des approches à caractère discursivo-fonctionnel visant à prédire la catégorie grammaticale d'une expression référentielle donnée en fonction de l'accessibilité cognitive du référent (Ariel, 1990) ou de certains types de transitions référentielles (la Théorie du Centrage ; Walker *et al.*, 1996 ; Cornish, 2000 ; la Hiérarchie du donné, cf. Gundel, Hedberg & Zacharsky, 1993) ;
- des approches qu'on pourrait dire configurationnelles ou relationnelles explicitant les liens entre une expression référentielle donnée et sa source, et qui ont mis au jour différents types d'anaphores, directe *vs* indirecte (Erkü & Gundel, 1987 ; Schwarz-Frisel, 2007).

Or, ces diverses approches ne prennent pas en considération les CR. Par nature et par méthode, les première et troisième méthodes, indispensables pour prédire les conditions d'apparition et d'emploi d'une expression référentielle, limitent le plus souvent l'analyse à des paires d'unité pour déterminer les contraintes pesant sur l'emploi de l'une d'elles ou caractériser une relation duelle. Quant aux approches du second type, qui ont mis au jour des paramètres importants conditionnant l'emploi des expressions référentielles (saillance, unité, distance, compétition, notamment, cf. Ariel, 1990), elles achoppent sur la réalité des textes qui prend souvent à contrepied leurs prédictions et restent aveugles aux surdéterminations des genres (Toole, 1996 ; Ariel, 2007).

2.2. Traitement automatique des langues

Le domaine du TAL peut être considéré comme une application privilégiée des théories linguistiques. Les régularités et les propriétés observées pour les expressions référentielles et les CR permettent d'envisager des systèmes informatiques, à la fois pour la reconnaissance automatique des chaînes – donc en compréhension – et pour la génération automatique de chaînes – donc en production. À notre époque où l'on voit se multiplier les agents conversationnels animés et autres « intelligences artificielles » capables de dialoguer avec des humains, il est intéressant – pour augmenter les performances très médiocres de ces systèmes – de les doter de la capacité à gérer des chaînes de référence. C'est même un enjeu essentiel pour aboutir à des dialogues cohérents et il est devenu courant de citer le défi de la résolution de la coréférence pour évaluer les performances d'un système dialoguant. Ce défi, que certains appellent « schémas de Winograd » (Levesque e*t al.*, 2012) suite à cette proposition de l'une des grandes personnalités du TAL, consiste à proposer au sujet – humain ou machine – deux phrases qui ne diffèrent que d'un seul mot, par exemple :

2. La poussette n'entre pas dans la valise car elle est trop grande.
3. La poussette n'entre pas dans la valise car elle est trop petite.

Les humains résolvent ce défi avec une très grande facilité : les expressions *elle* et *poussette* sont coréférentielles dans le premier cas ; *elle* et *valise* dans le second cas. Les systèmes TAL ont beaucoup de mal à résoudre ce type de défi, qui reste un enjeu important de l'intelligence artificielle et du TAL.

Dans ce domaine TAL de l'identification automatique des chaînes de référence, il existe, à l'heure actuelle, une dizaine de systèmes qui parviennent à identifier plus ou moins bien les

relations de coréférence dans un texte tout venant[1] (dans une langue autre que le français). Les performances sont intéressantes compte tenu de la difficulté de la tâche, mais les erreurs faites par les systèmes peuvent sembler grossières du point de vue du linguiste : des pronoms sont affectés à des référents non pertinents, des expressions référentielles ne sont même pas repérées, alors qu'une lecture (humaine) rapide donne immédiatement les solutions. Pour améliorer les performances de tels systèmes, les efforts actuels portent sur la sémantique lexicale en tant que critère supplémentaire (Ng, 2007) et, d'une manière générale, sur la liste nécessaire et suffisante de critères pour résoudre les coréférences (Bengtson & Roth, 2008), sur la hiérarchisation des critères (par ordre d'importance) et la combinaison de plusieurs algorithmes, avec, par exemple, plusieurs passes (Raghunathan *et al*., 2010), sur le développement de plateformes permettant à chacun de paramétrer les critères de son propre système de résolution des coréférences (Stoyanov *et al.*, 2010), sur l'exploitation d'informations liées aux entités nommées, après apprentissage automatique (Haghighi & Klein, 2010), sur l'exploitation d'informations permettant de relier automatiquement deux termes pleins pour résoudre les anaphores infidèles (Recasens *et al*., 2013), ou encore sur l'exploitation d'informations spécifiques au domaine (Gilbert & Riloff, 2013). Les techniques exploitées sont multiples, et l'article de F. Landragin (ce volume) donnera quelques repères sur celles qui sont le plus utilisées.

Enfin, des efforts sont faits également sur des tâches un peu plus spécifiques ou marginales, comme la résolution des coréférences dans un contexte biomédical (domaine relevant de la bio-informatique, qui apporte ses propres contraintes compte tenu de la nature des référents en présence), la résolution des coréférences événementielles (Cybulska & Vossen, 2013) ou des coréférences dans des textes multilingues (Zhekova & Kübler, 2013). Ce seront des aspects que nous n'aborderons pas dans ce volume, mais qui montrent à quel point les applications de l'identification automatique des CR sont nombreuses et peuvent avoir des répercussions sociétales importantes.

2.3. Corpus et analyse statistique de données textuelles

Un point commun des approches linguistiques (en particulier des approches « configurationnelles ») et des approches du TAL est leur recours à des corpus, et en particulier à des corpus annotés. En linguistique, les corpus servent, dans certains cas, de réservoir d'exemples et permettent de confronter à des productions réelles hypothèses et théories sur la référence. L'annotation à la main par des linguistes de phénomènes référentiels et coréférentiels permet, d'une part, de retrouver facilement un exemple illustrant une propriété précise (car la recherche se fait *via* les annotations), et, d'autre part, d'identifier des tendances, voire de commencer à quantifier, par exemple les proportions relatives des catégories grammaticales comme les déterminants ou pronoms définis, démonstratifs et indéfinis pour les expressions référentielles, ou encore les longueurs en nombre de maillons des CR.

Cependant, la linguistique de corpus peine à s'intéresser aux CR, du moins pour ce qui concerne la langue française. De fait, il n'existe pas pour le français de corpus annoté en CR, comme le montre l'inventaire exhaustif récent de Recasens (2010, 10). Salmon-Alt (2002) faisait déjà ce constat sur les données disponibles à l'étranger (art. cit., 165) et en France (art. cit., 166) où les ressources sont, selon elle, disparates (annotations variables d'un corpus à l'autre, schémas d'annotation différents les uns des autres, ressources comprenant d'autres informations que la coréférence, désaccords fréquents entre annotateurs). *A fortiori*, il n'existe pas, pour le français, de corpus longitudinal annoté où soit prise en considération l'évolution de la composition des CR, alors qu'il existe des corpus annotés morpho-syntaxiquement de textes issus de toutes les périodes de la langue française.

Il est donc difficile d'apporter aux études linguistiques des mesures quantitatives fiables pour les CR : il faudrait pour cela que l'annotation de celles-ci soit plus répandue dans les corpus, et soit

[1] Cf. par exemple la comparaison visible sous le titre « Coreference Resolution Tools: A First Look » sur : http://www.minvolai.com/blog/2010/09/Coreference-Resolution-Tools--A-first-look/coreference-resolution-tools-a-first-look/ – consulté le 17 novembre 2016).

exploitée par la communauté de l'analyse statistique de données textuelles (ADT), dans ses efforts pour que les statistiques prennent en compte à la fois le texte et les annotations. Plusieurs articles de ce numéro franchissent un pas dans cette direction, en adaptant à la méthodologie de l'annotation des mesures proposées pour l'étude des CR[2] et en montrant comment les outils d'annotation et de gestion de corpus peuvent les intégrer[3], permettant ainsi d'étendre considérablement les possibilités d'analyse linguistique des CR.

En TAL, les corpus annotés ont deux rôles essentiels. Le premier, et ce depuis le tout début de la linguistique de corpus, est celui du « corpus de référence », c'est-à-dire de point de comparaison permettant d'évaluer les performances d'un système : on implémente un système et on le teste sur un exemple de texte existant par ailleurs dans une version annotée manuellement par des linguistes. Plus le résultat du système s'approche de cette version de référence, plus le système est considéré comme performant.

Le second rôle des corpus est de fournir des données d'apprentissage. Beaucoup de systèmes exploitent des techniques d'apprentissage artificiel, et ces techniques ne peuvent se mettre en œuvre que lorsque l'on dispose d'exemples correctement annotés, servant de modèles à un apprenant (artificiel, en l'occurrence). L'apprentissage par l'exemple présente des avantages – souplesse, efficacité – et des défauts, notamment le fait que le système peut « mal » apprendre. Toutefois, plus la taille et la qualité des corpus annotés utilisés pour apprendre est grande, plus ces défauts s'estompent.

Au final, plusieurs communautés se sont approprié les CR, pour des études qui restent pour le moment peu corrélées. Les raisons sont nombreuses, et résident en particulier dans la variété des méthodologies suivies. Mais la raison principale est que, comme dit *supra*, il n'existe pas de corpus de référence pour le français annoté en CR : en l'absence d'un tel corpus, il est difficile de faire progresser les systèmes de TAL, et il est difficile de renforcer les aspects statistiques et opérationnels des études de corpus outillées.

3. Objectifs de ce volume

3.1. Objectifs scientifiques

C'est pour combler une partie de ces lacunes (faibles interactions entre les disciplines concernant les CR ; besoin de mesures pour renforcer les études linguistiques des CR et d'un corpus annoté en CR pour la langue française) que les contributeurs du présent volume ont confronté leurs expériences et méthodes. Dans ce contexte, ce volume réunit différentes approches, tant linguistiques qu'informatisées, et montre leurs interactions réciproques ainsi que leurs apports mutuels. Ce volume pose notamment des questions cruciales, relatives aux outils et procédures d'annotation, aux spécificités des genres discursifs soumis à l'étude et aux contraintes que font peser sur la constitution mais aussi la détection des CR les phénomènes de variation :

- générique, avec la question de savoir si, et dans quelle mesure, les genres discursifs contraignent l'expression de la coréférence,
- diachronique, avec les problèmes de l'évolution du système linguistique mais aussi de celle des genres textuels et des phénomènes transphrastiques.

Il en démontre aussi certains des enjeux et intérêts, théoriques et méthodologiques, qui vont de la constitution de corpus d'apprentissage automatique à des applications comme la simplification automatique des textes à l'attention d'apprenants de statuts divers.

---

[2] À la suite de travaux allant dans cette direction, comme Schnedecker 2005, 2014, ce volume, et Schnedecker & Longo 2012.

[3] Continuant le travail amorcé dans le projet MC4, voir *infra*.

3.2. Historique de la constitution de ce volume

Les réflexions présentées dans ce volume font suite à un précédent numéro de revue thématique (*Langages* 195, 2014) intitulé *Les Chaînes de référence*, qui présentait les premiers résultats d'un projet de recherche collaboratif[4]. Ce numéro avait pour ambition de proposer un état des acquis, des problèmes et des perspectives dans le domaine des chaînes de référence.

Dans la suite du projet MC4, une journée d'étude intitulée *Chaînes de référence et phénomènes de variation* s'est déroulée à Strasbourg en juin 2015[5]. Cette journée a permis une première confrontation des approches issues de projets différents ou de perspectives différentes, et a donné lieu, lors des discussions, à une réflexion commune et une mise en commun des méthodologies employées, des critères d'analyses pertinents, des phénomènes de variation à prendre en compte… Les contributions de ce volume sont des versions approfondies et remaniées de présentations à cette journée.

3.3. Perspectives et projet ANR DEMOCRAT

Le présent volume poursuit la réflexion, l'annotation et les développements informatiques (outil d'annotation des CR), dans le cadre d'un nouveau projet ANR, le projet DEMOCRAT, « DEscription et MOdélisation des Chaînes de Référence : outils pour l'Annotation de corpus (en diachronie et en langues comparées) et le Traitement automatique », dirigé par F. Landragin et réunissant des chercheurs des laboratoires Lattice, LiLPa, ICAR et IHRIM pour une durée de quatre ans (2016-2020).

Ce projet DEMOCRAT vise quatre objectifs. Il s'agit, premièrement, de construire un modèle théorique de la référence et de la coréférence, qui tienne compte des aspects discursifs des CR et réponde aux critiques énoncées au début de cet article sur l'état de l'art linguistique. Il s'agit deuxièmement de constituer un corpus annoté en CR, comprenant plusieurs états de la langue française, depuis le XI^{e} jusqu'au XXI^{e} siècle, intégrant des textes de plusieurs genres textuels. Le troisième objectif porte sur l'amélioration et l'adaptation des outils de linguistique de corpus pour qu'ils tiennent mieux compte d'annotations complexes comme celles du type des CR. Sur ce point, les efforts portent sur le logiciel d'annotation ANALEC (Landragin *et al.*, 2012) – qui sert dans un premier temps à produire le corpus DEMOCRAT – et surtout sur la plateforme de gestion, d'exploration et d'analyse statistique TXM (Heiden *et al.*, 2010) qui ne permet actuellement pas l'annotation manuelle de corpus. L'outil TXM devrait se doter des fonctionnalités d'annotation d'ANALEC et, au passage, adapter ces fonctionnalités aux CR. Le quatrième et dernier objectif est la réalisation d'un ou de plusieurs – en fonction des techniques utilisées – systèmes de TAL pour l'identification automatique des CR.

Parmi les six contributions de ce numéro que nous allons maintenant présenter, trois font partie des efforts entrepris dans le projet DEMOCRAT : les articles de F. Landragin, de C. Schnedecker et de V. Obry *et al.* Les trois autres – articles d'A. Todirascu *et al.*, d'E. Baumer et de B. Combettes – viennent enrichir ces efforts avec des points de vue plus différents, soit qu'ils proviennent de cadres théoriques particuliers (la théorie des opérations énonciatives d'A. Culioli, dans l'approche d'E. Baumer), soit qu'ils prennent en considération des aspects linguistiques à envisager dans le cadre du projet DEMOCRAT comme la notion de *point de vue* ou celle de *cataphore*. Nous espérons ainsi donner une image complète et réaliste des efforts entrepris actuellement par la linguistique française autour de l'étude des CR.

---

[4] Le projet MC4 : *Modélisation Contrastive et Computationnelle des Chaînes de Coréférence*, dirigé par F. Landragin et réunissant des chercheurs en linguistique et en informatique des laboratoires Lattice, LiLPa et ICAR (dont les participants de l'époque – pour ce dernier laboratoire – font désormais partie du laboratoire IHRIM).

[5] Cette journée a pu être réalisée grâce au soutien de l'équipe Fonctionnement Discursif et Traduction du laboratoire LiLPA (EA 1339).

4. Présentation des contributions

Les contributions sont regroupées en trois volets. L'un porte sur les questions de corpus d'étude, d'annotation et d'outillage servant la procédure d'annotation. L'autre, aborde la question des chaînes de référence par un versant linguistique et se subdivise en deux sections selon l'état historique du français (français contemporain et moyen français)

Dans son article intitulé « Analyse, visualisation et identification automatique des chaînes de coréférences : des questions interdépendantes ? », F. Landragin passe en revue un ensemble de propositions visant l'annotation (manuelle et automatique) des chaînes de référence en montrant les avantages et inconvénient des unes et des autres, considérations dont la portée dépasse bien évidemment la question de l'annotation des chaînes de référence et vise les procédures d'annotation en général.

Dans le cadre d'un projet visant à concevoir un système de simplification automatique des textes, destiné à adapter les textes en fonction des compétences en lecture du public et donc à proposer des règles de simplification conservant la cohésion, A. Todirascu, T. François, D. Bernhard, N. Gala, A.-L. Ligozat et R. Khobzi présentent, dans leur contribution « Chaînes de référence et lisibilité des textes: le projet ALLuSIF », une expérience d'annotation et une étude des propriétés de chaînes de références et des chaînes anaphoriques dans deux corpus français, composés de textes, majoritairement de type informatif, mais aussi narratif, et destinés à différentes catégories de lecteurs (enfants/adultes pour le premier, apprenants de FLE pour le deuxième). Y sont étudiées les propriétés des chaînes annotées (la longueur, le changement des fonctions syntaxiques, la distance et le type de maillons) et passés en revue les problèmes rencontrés lors de l'annotation de ces corpus dans le but de proposer un ensemble de solutions.

Dans une perspective linguistique, et suite à l'examen des éléments de discussion collectés dans les travaux français et anglo-saxons, C. Schnedecker montre dans « Les chaînes de référence : une configuration d'indices pour identifier les genres textuels », que l'expression de la coréférence est réellement tributaire des genres textuels dans lesquels elle se manifeste. Néanmoins, pour être véritablement efficiente dans l'identification et la caractérisation des genres de discours, la dimension référentielle doit dépasser la seule prise en compte quantitative des catégories référentielles telles qu'elle apparaît dans certaines approches communément appelées « paradigmatiques » des genres discursifs. Le propos est illustré et étayé par l'étude comparée de deux genres, les faits divers et les incipit de contes de fées mettant en évidence les nombreuses divergences linguistiques, notamment au niveau de la composition des chaînes de référence. L'étude aboutit à une « grille d'analyse » des chaînes de référence susceptible de s'appliquer à d'autres genres.

L'article d'E. Baumer, « Chaînes de référence et point de vue dans la fiction littéraire : Le cas des nouvelles courtes », s'inscrit dans le cadre théorique dit « théorie des opérations énonciatives » (TOE) et nous propose de suivre le fonctionnement des chaînes de référence renvoyant aux personnages principaux (animés humains) dans un corpus constitué de onze nouvelles contemporaines courtes afin d'étudier les interactions entre ces CR et le point de vue narratif (PDV). L'influence du genre textuel sur les CR est abordée par la comparaison de cette sous-catégorie de fiction littéraire à un corpus de nouvelles « standard », plus développées.

L'article de V. Obry, J. Glikman, C. Guillot et B. Pincemin, intitulé « Les chaînes de référence dans les récits brefs en français : étude diachronique (XIII^{e} - XVI^{e} s.) » porte sur la composition des chaînes de référence dans un corpus de textes narratifs composés entre le début du XIII^{e} et le milieu du XVI^{e} siècle, constitué de récits brefs de type humoristique centrés sur un petit nombre de référents que sont les nouvelles et les fabliaux. L'objectif de l'étude est d'interroger l'importance du genre textuel dans la construction des chaînes de référence en diachronie. Les textes du corpus ont fait l'objet d'une annotation manuelle à l'aide du logiciel ANALEC (Mélanie-Becquet et Landragin, 2014) et le calcul des mesures a été fait à l'aide de la plateforme TXM (Heiden *et al.*, 2010). L'article pose en outre la question de la pertinence des différentes mesures pour l'étude des CR.

Dans son article intitulé « Facteurs discursifs et contraintes syntaxiques : aspects diachroniques de la relation de cataphore », B. Combettes se propose d'examiner l'évolution, au cours de la

diachronie du français, de certains aspects du système des relations de cataphore dans la prose narrative. Après avoir passé en revue deux types d'analyses en concurrence, l'un privilégiant le rôle des hiérarchisations syntaxiques dans le cadre de la phrase, l'autre insistant au contraire sur l'importance du « flux discursif » et des contraintes textuelles. C'est cette dernière approche qui est privilégiée par l'auteur, la prise en compte de la dimension diachronique semblant mettre en lumière sa pertinence d'autant qu'elle pose la question de la délimitation des unités pertinentes, des « paliers de traitement », qu'il convient de prendre en compte dans une étude diachronique.

Remerciements

Ce travail a bénéficié du soutien de l'ANR dans le cadre du projet DEMOCRAT (ANR-15-CE38-0008).

Références

ARIEL M. (2004), "Accessibility marking", *Discourse Processes* 37 (2), 91-116.

ARIEL M. (1990), *Accessing Noun-Phrase Antecedents*, London/New York: Routledge.

BENGTSON E. & ROTH D. (2008), "Understanding the Value of Features for Coreference Resolution", *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, Honolulu, 294-303.

CORNISH F. (1999), *Anaphora, Discourse, and Understanding. Evidence from English and French*, New York: Oxford University Press.

CORNISH F. (éd.) (2000), « Référence discursive et accessibilité cognitive », *Verbum* XXII (1), 7-30.

CORNISH F. (2006), "Discourse anaphora", In K. Brown (ed.), *Encyclopedia of Language and Linguistics (second edition)*, Oxford: Elsevier, 631-638.

CYBULSKA A. & VOSSEN P. (2013), "Semantic Relations between Events and their Time, Locations and Participants for Event Coreference Resolution", *Proceedings of Recent Advances in Natural Language Processing (RANLP 2013)*, Hissar, Bulgaria, 156-163.

GILBERT N. & RILOFF E. (2013), "Domain-Specific Coreference Resolution with Lexicalized Features", *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*, Sofia, Bulgaria.

GUNDEL J., HEDBERG N. & ZACHARSKI R. (1993), "Cognitive Status and the Form of Referring Expressions in Discourse", *Language* 69, 274-307.

HAGHIGHI A. & KLEIN D. (2010), "Coreference Resolution in a Modular, Entity-Centered Model", *Proceedings of Human Language Technologies: The 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, Los Angeles, 385-393.

HEIDEN S., MAGUE J.-P. & PINCEMIN B. (2010), « TXM : Une plateforme logicielle open-source pour la textométrie – conception et développement », *Actes des 10e Journées Internationales d'Analyse statistique des Données Textuelles (JADT 2010)*, Rome, 1021-1032.

LANDRAGIN F., POIBEAU T. & VICTORRI B. (2012), « ANALEC: a New Tool for the Dynamic Annotation of Textual Data », *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC)*, Istanbul, Turquie, 357-362.

LEVESQUE H.J., DAVIS E. & MORGENSTERN L. (2012), "The Winograd Schema Challenge", *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*, Rome, Italy, 552-561.

MELANIE-BECQUET F. & LANDRAGIN F. (2014), « Linguistique outillée pour l'étude des chaînes de référence : questions méthodologiques et solutions techniques », *Langages* 195, 117-137.

NG V. (2007), "Shallow Semantics for Coreference Resolution", *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, Hyderabad, India, 1689-1694.

RAGHUNATHAN K., LEE H., RANGARAJAN S., CHAMBERS N., SURDEANU M., JURAFSKY D. & MANNING C. (2010), "A Multi-Pass Sieve for Coreference Resolution", *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Massachusetts Institute of Technology, 492-501.

RECASENS M. (2010), Coreference: Theory, Annotation, Resolution and Evaluation, PhD thesis, Barcelona: University of Barcelona.

RECASENS M., CAN W. & JURAFSKY D. (2013), "Same Referent, Different Words: Unsupervised Mining of Opaque Coreferent Mentions", *Proceedings of Human Language Technologies: The 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, Atlanta, Georgia, 897-906.

SALMON-ALT S. (2002), « Le projet ANANAS : annotation anaphorique pour l'analyse sémantique de corpus », *Workshop sur les chaînes de référence et résolveurs d'anaphores, 9e conférence sur le traitement automatique des langues naturelles (TALN)*, Nancy.

SCHNEDECKER C. & LANDRAGIN F. (2014), « Les chaînes de référence : présentation », *Langages* 195, 3-22.

STOYANOV V., CARDIE C., GILBERT N., RILOFF E., BUTLER D. & HYSOM D. (2010), "Coreference Resolution with Reconcile", *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, Uppsala, Sweden, 156-161.

TOOLE J. (1996), "The effect of genre on referential choice", *in* T. Fretheim & J.K. Gundel (eds.), *Reference and Referent Accessibility*, Amsterdam: John Benjamins, 262-290.

VANDERBAUWHEDE G. (2012), *Le déterminant démonstratif en français et en néerlandais. Théorie, description, acquisition*, Bern: Lang.

VANDERBAUWHEDE G. (2012), « La référence démonstrative dans les corpus écrits : théorisation et analyse de données empiriques ». *Actes du 3e Congrès Mondial de Linguistique Française*, 1991-2009.

WALKER M., JOSHI A. & PRINCE E. (1998), "Centering in Naturally Occurring Discourse: An Overview", *in* M. Walker, A. Joshi & E. Prince (eds.), *Centering Theory in Discourse*, Oxford: Clarendon Press, 1-28.

WALKER M., JOSHI A. & PRINCE E. (eds.) (1998), *Centering Theory in Discourse*, Oxford: Clarendon Press.

ZHEKOVA D. & KÜBLER S. (2013), "Machine Learning for Mention Head Detection in Multilingual Coreference Resolution", *Proceedings of Recent Advances in Natural Language Processing (RANLP-2013)*, Hissar, Bulgaria, 747-754.