EXPRESSIONS RÉFÉRENTIELLES ET CHAÎNES DE RÉFÉRENCE EN FRANÇAIS : LE PROJET DEMOCRAT ET SON EXPLORATION DES RAPPORTS ENTRE LINGUISTIQUE TEXTUELLE ET LINGUISTIQUE DE CORPUS

Frédéric LANDRAGIN
CNRS, Laboratoire Lattice

Abstract (En): We present the Democrat project, "Description, modelling and automatic detection of coreference chains in French," and its four objectives, that is to provide: (i) an integrated, discursive, diachronic and inter-genre description of coreference chains; (ii) a corpus of written French texts with annotated coreference chains; (iii) several tools for visualizing and exploring the coreference chains; (iv) two NLP systems that are able to process raw text written in French and to extract referring expressions as well as coreference chains – which have also brought innovations to the field of deep learning. We present the main results of Democrat and we describe the work steps that made it possible to obtain them, in particular the corpus, which was manually annotated by forty members of the project.

Keywords (En): referring expressions; coreference; coreference chains; annotated corpus

Mots-clés (Fr) : expressions référentielles ; coréférence ; chaînes de référence ; corpus annoté

DOI : 10.32725/eer.2022.004

Introduction

Dans cet article, nous présentons de manière synthétique les résultats du projet ANR Democrat, en orientant leur description sur les interactions entre linguistique textuelle, linguistique de corpus, informatique et intelligence artificielle. Pour les autres résultats et pour une présentation générale du projet, nous renvoyons à divers articles et numéros thématiques de revues déjà parus (LANDRAGIN & SCHNEDECKER 2014 ; SCHNEDECKER *et al*. 2017 ; QUIGNARD *et al*. 2018 ; LANDRAGIN 2021). Ce projet de recherche, que nous avons coordonné de fin 2016 à juin 2020, a porté sur la référence et les chaînes de référence en langue française, en suivant une approche pluridisciplinaire, comme l'indique la signification du sigle Democrat : « Description et modélisation des chaînes de référence : outils pour l'annotation de corpus (en diachronie et en langues comparées) et le traitement automatique ». Il a réuni une quarantaine de chercheurs et doctorants, spécialistes de linguistique, d'outils et de traitement automatique des langues, répartis sur une dizaine de laboratoires de recherche français, notamment les laboratoires partenaires Lattice (Paris), LiLPa (Strasbourg), ICAR et IHRIM (Lyon).

Pendant une bonne partie du projet, de fin 2016 à début 2019, les participants ont constitué et annoté manuellement un corpus, qui est disponible depuis mi-2019 en accès libre et gratuit sous le nom de « corpus Democrat », sur la plate-forme Ortolang (LANDRAGIN 2019). Ce corpus regroupe 58 textes de divers genres, littéraires ou non, narratifs ou non, pour un total de 689.000 mots, ce qui représente – chiffre plus intéressant – 198.000 expressions référentielles annotées. Tous les siècles sont représentés, du XII^{e} au XXI^{e}, ce qui rend possibles des analyses

diachroniques. La quantité des annotations a été fixée afin de permettre des analyses statistiques du corpus, et d'en autoriser des exploitations par des outils de traitement automatique des langues (TAL), plus précisément par des outils faisant appel à de l'apprentissage profond, domaine clé de l'intelligence artificielle actuelle.

De nouvelles méthodes d'analyse qualitative et quantitative ont été imaginées et expérimentées dans le cadre du projet Democrat, avec des mesures adaptées aux chaînes de référence et des visualisations dédiées, qui sont à l'origine d'une évolution majeure de l'interface graphique et de la bibliothèque de macros de la plateforme de textométrie TXM (HEIDEN 2010). Deux systèmes de TAL ont pu être développés pour détecter automatiquement les chaînes de référence dans du texte brut tout-venant. Ces interfaces et outils offrent de nouvelles possibilités d'analyse de textes écrits en français, et permettent de développer une méthodologie *a priori* utilisable pour d'autres thématiques de recherche à la croisée de la linguistique textuelle et de la linguistique de corpus.

Nous commençons comme il se doit par des définitions et par une présentation des réflexions approfondies dans le projet Democrat sur les notions de référence et de chaîne de référence (section 1). Nous enchaînons avec quelques détails sur le corpus Democrat, qui constitue l'une des principales matérialisations de ces réflexions (section 2). Annoter et explorer ce corpus nécessite des outils informatiques dédiés, que nous présentons alors (section 3). Quant à l'exploitation du corpus par des techniques d'apprentissage profond, elle fait l'objet de la section suivante, qui présente les deux systèmes produits par le projet, nommés COFR et DeCOFR (section 4). De fait, ce découpage en quatre sections correspond aux quatre volets du projet Democrat et à ses quatre livrables. Afin de dresser un bilan de ces résultats, nous mettons ensuite en perspective les aspects textuels et les aspects corpus (section 5), puis nous concluons.

1. Description des chaînes de référence

1.1. Les expressions référentielles et les chaînes de référence

Une expression référentielle est une forme linguistique qui donne accès à un référent du discours (KARTTUNEN 1976 ; CHAROLLES 2002). Entre autres exemples, il peut s'agir d'un nom propre, d'un groupe nominal, d'un pronom anaphorique. Une chaîne de référence regroupe les différentes expressions référentielles qui réfèrent au même référent (CHASTAIN 1975 ; CORBLIN 1995 ; SCHNEDECKER 1997). Afin d'entrer directement dans l'étude de l'objet complexe qu'est la chaîne de référence, prenons l'exemple suivant, écrit par Robert Antelme et constituant le dernier paragraphe de l'avant-propos de son essai *L'Espèce humaine* (1947) :

> Dire que l'on se sentait alors contesté comme homme, comme membre de l'espèce, peut apparaître comme un sentiment rétrospectif, une explication après coup. C'est cela cependant qui fut le plus immédiatement et constamment sensible et vécu, et c'est cela d'ailleurs, exactement cela, qui fut voulu par les autres. La mise en question de la qualité d'homme provoque une revendication presque biologique d'appartenance à l'espèce humaine. Elle sert ensuite à méditer sur les limites de cette espèce, sur sa distance à la « nature » et sa relation avec elle, sur une certaine solitude de l'espèce donc, et pour finir, surtout à concevoir une vue claire de son unité indivisible. (page 11 dans l'édition Gallimard de 1957)

Commençons par identifier quelques-uns des référents de cet extrait. La première phrase contient le pronom « on », expression référentielle dont la référence est connue pour être parfois floue, ni spécifique ni vraiment générique ; qui plus est, identifier ici qui est « on » nécessite de connaître le contexte antérieur, comme si le pronom était doté d'une valeur anaphorique (DELABORDE & LANDRAGIN 2019). La forme « on » est donc clairement repérée comme expression référentielle – ou « mention », ou « maillon » de chaîne de référence –, mais sa référence exacte reste imprécise, ce qui ne pose *a priori* pas de problème pour identifier les autres maillons qui lui sont coréférentiels. Cette première remarque est importante dans la mesure où elle souligne les deux priorités du projet Democrat : identifier les expressions référentielles, et les relier entre elles par des liens de coréférence, sans forcément approfondir la nature et l'identité exacte du référent lui-même.

La première phrase contient également deux expressions nominales : « un sentiment rétrospectif » et « une explication après coup ». Un doute porte alors sur le nombre de référents : un seul, ou deux ? La virgule séparant les deux expressions ne permet pas de décider ; il est au contraire nécessaire d'entrer dans des considérations sémantiques, et donc de lire attentivement le texte. Soit le deuxième contenu informationnel vient préciser le premier, et dans ce cas on peut considérer que le référent est unique et qu'il est désigné par l'expression référentielle unique « un sentiment rétrospectif, une explication après coup » ; soit les deux informations juxtaposées sont distinctes et nous avons alors deux référents, désignés par deux expressions référentielles. Ce qui nous amène à une deuxième remarque qui vient nuancer celle du paragraphe précédent : même sans avoir nécessairement besoin d'identifier les référents précis, il est parfois utile de le faire, ne serait-ce que pour délimiter correctement les expressions référentielles. L'analyse linguistique comme l'annotation ne peuvent pas se faire de manière « robotique », c'est-à-dire sans concentration (en mode « lecture rapide »), ou sans chercher à comprendre le texte.

La première phrase de l'extrait contient en outre « homme » et « membre de l'espèce », qui viennent qualifier le référent de « on ». Au passage, notons que « l'espèce » est elle-même une expression référentielle, qui réfère à l'espèce humaine du titre du livre. Doit-on considérer « homme » et « membre de l'espèce » comme des expressions référentielles, ou comme des expressions d'un autre type – ce qui conduirait à les ignorer dans les chaînes de référence ? On le voit : à elle seule, la première phrase de l'extrait pose déjà des problèmes complexes d'identification et de délimitation des expressions référentielles.

Or nous nous sommes arrêtés jusqu'ici aux expressions nominales et pronoms – ce qui correspond aux choix effectués dans le projet Democrat. Mais on pourrait ajouter également un référent de type événement, avec l'expression sujet « dire que l'on se sentait alors contesté comme homme, comme membre de l'espèce », qui pourrait par exemple constituer un antécédent potentiel pour le pronom « cela » présent dans la phrase suivante (de même que l'autre antécédent possible – et plus vraisemblable – à savoir « on se sentait alors contesté […] »). Conformément à la méthodologie mise en place, le projet Democrat ignore de telles formes, qui ne sont donc pas annotées, et se focalise sur les expressions nominales au sens large : groupes nominaux, noms propres, pronoms.

Notons que les chaînes de référence regroupent très souvent des expressions référentielles qui appartiennent à des phrases différentes, parfois consécutives, parfois entrecoupées de phrases contenant d'autres références. Cette propriété en fait un objet de discours, qui relève de la linguistique textuelle. Les choix effectués dans Democrat ont pour principal objectif la constitution d'un grand corpus annoté en expressions référentielles et en chaînes de référence : les questions soulevées par ces choix sont donc au cœur des rapports entre linguistique textuelle et linguistique de corpus.

1.2. Quelques cas intéressants et représentatifs des choix d'orientation de Democrat

Penchons-nous sur la deuxième phrase de l'extrait : « c'est cela » pose un problème relatif aux constructions du type « X est Y ». D'une manière générale, on considère dans de tels cas que « X » réfère et que « Y » est attributif et donc ne réfère pas, ce qui conduit ici à considérer « c' » comme expression référentielle et à ignorer « cela ». Or ce même pronom est répété deux fois, dans « c'est cela d'ailleurs » puis dans « exactement cela », et est repris ensuite par le pronom relatif « qui » (autre exemple d'expression référentielle). On est donc en présence d'une chaîne de référence, dont les maillons posent quelques problèmes d'identification délicats. Par exemple : que faire de l'adverbe « exactement » : faut-il l'inclure dans l'expression référentielle « exactement cela » ? Comme il faut trancher, nous avons décidé dans Democrat de ne pas inclure de tels adverbes, sachant que la question reste linguistiquement pertinente et qu'une autre réponse aurait été avancée dans un autre projet d'annotation de la référence.

Continuons avec le pronom de troisième personne « elle » qui initie la dernière phrase. Quel est son référent : « la mise en question de la qualité d'homme » ou « une revendication presque biologique d'appartenance à l'espèce humaine » ? Là aussi, bien comprendre le sens du texte s'avère indispensable. En effet, aussi bien une mise en question qu'une revendication peut servir à méditer : il y a donc ambiguïté sémantique. Parmi les deux alternatives, notons que « la mise en question » est le sujet grammatical de la phrase précédente, et se rapproche donc aisément – selon le principe du parallélisme syntaxique – du sujet « elle » de la phrase courante. L'annotateur de Democrat est ainsi incité à trancher et à annoter « la mise en question » comme coréférentiel avec « elle ». Mais d'autres choix auraient été possibles : annoter les alternatives en tant que telles, par exemple, ou gérer deux couches d'annotations mutuellement exclusives.

Notons qu'un phénomène similaire apparaît dans « sa relation avec elle », où nous identifions deux expressions référentielles, « sa » et « elle », qui réfèrent l'une à l'espèce, l'autre à la nature. Oui, mais : laquelle réfère à l'espèce, et laquelle réfère à la nature ? *A priori*, les deux possibilités sont envisageables, c'est-à-dire que les références sont interchangeables, comme avec « l'un […] l'autre ». Là encore, la syntaxe permet d'avancer une solution : on garde le même ordre des références et on attribue donc « sa » à « l'espèce », et « elle » à « la nature ». Les chaînes de référence sont ainsi clarifiées mais, vous l'aurez compris, au prix de multiples

interrogations qui font intervenir des arguments morphologiques, syntaxiques, sémantiques, pragmatiques et textuels.

Les efforts du projet Democrat ont porté sur la prise en compte de ces multiples facteurs, afin de mettre en œuvre des conventions permettant d'annoter des textes écrits en français, et donc d'aboutir à un corpus qui satisfasse au mieux les contraintes inhérentes à la méthodologie de la linguistique de corpus outillée. D'une manière générale, retenons que lorsque l'on vise un corpus de grande taille, on peut difficilement mettre en œuvre une procédure d'annotation complexe, avec plusieurs couches d'annotation et avec la prise en compte d'incertitudes, de flous, voire d'ambiguïtés. Au contraire, le travail d'annotation doit pouvoir se faire assez rapidement, quitte à devoir trancher, sans problèmes techniques potentiels ni manipulations ergonomiques trop fastidieuses, comme nous allons le voir dans les sections suivantes.

2. Constitution et annotation du corpus Democrat

2.1. Choix des textes : deux facteurs de variation

Afin d'autoriser des analyses linguistiques qui suivent différentes préoccupations de recherche, nous avons constitué le corpus Democrat de manière à ce qu'il regroupe des textes de tailles similaires. L'objectif était de multiplier les textes, en prenant des chapitres de romans, des nouvelles, des textes journalistiques, juridiques, encyclopédiques, techniques, etc. Afin de cadrer le travail, nous avons choisi la taille de 10 000 mots pour chaque extrait – taille estimée comme suffisante pour observer une grande diversité de phénomènes référentiels –, et nous avons retenu pour moitié des textes narratifs, pour moitié des textes non narratifs, avec dans ce cas une diversité maximale de genres textuels – qui constitue le premier facteur de variation dans le corpus. Cette diversité nous semblait essentielle par rapport à notre but d'articuler linguistique textuelle et linguistique de corpus.

Le deuxième facteur de variation est la période : nous avons, dans la mesure du possible, choisi des textes représentatifs écrits en ancien français, en moyen français, en français moderne et en français contemporain. Là aussi, il a fallu cadrer et faire avec la disponibilité de textes (notamment pour l'ancien français), ce qui nous a conduits à retenir pour moitié des textes en français contemporain, pour moitié des textes plus anciens. Dans cette dernière moitié, nous avons fait notre possible pour que la répartition selon les siècles soit la plus homogène possible, afin d'équilibrer les analyses diachroniques.

2.2. Conventions d'annotation et répercussions sur les analyses

Le choix des extraits nous amène à travailler avec des textes bruts, sans annotations. Les éventuelles annotations existantes, par exemple quand tel ou tel extrait a déjà été étudié et annoté dans le cadre d'un autre projet, sont temporairement mises de côté (et ne sont pas incluses dans le corpus Democrat). Nous en venons au travail d'annotation, et en premier lieu à la définition d'un schéma d'annotation. Un tel schéma a déjà été conçu dans un grand nombre de projets relatifs à la coréférence, et a déjà suscité des critiques (VAN DEEMTER & KIBBLE 2000). Dans notre cas, nous avons retenu la simplicité : les seuls marquables

sont les expressions référentielles, qui demandent deux tâches à l'annotateur : 1. la tâche de délimitation ; 2. la tâche d'attribution d'un identifiant de référent. Cette deuxième tâche est laissée au libre choix de l'annotateur. L'identifiant en tant que tel n'a aucune importance : seul compte le fait que le même identifiant serve à annoter les expressions coréférentielles. En effet, c'est en saisissant le même identifiant que l'annotateur indique à l'outil d'annotation qu'il y a coréférence. Et c'est cet outil, de lui-même, qui va construire les chaînes de référence.

Par conséquent, le manuel d'annotation, qui décrit les conventions d'annotation, se focalise pour l'essentiel sur les critères permettant à l'annotateur, d'une part de déterminer si telle expression est considérée comme référentielle ou non (voir les exemples donnés dans la section précédente), d'autre part de délimiter correctement chaque expression référentielle (*idem*).

Ce choix opéré dans Democrat présente un avantage énorme : celui de ne pas avoir à gérer de chaîne de référence en tant qu'objet à annoter selon une procédure spécifique. En effet, les objets linguistiques que sont les chaînes de référence ne sont pas adaptés aux contraintes techniques et ergonomiques des outils d'annotation, ne serait-ce que par leur caractéristique consistant à couvrir potentiellement l'intégralité du texte : on ne peut tout simplement pas visualiser une chaîne dans son intégralité, car cela nécessiterait d'afficher à l'écran l'intégralité du texte, qui en deviendrait illisible. Face à ces préoccupations ergonomiques, le projet Democrat a apporté des solutions que nous allons maintenant explorer et illustrer.

3. Outils de visualisation et d'exploration des chaînes de référence

3.1. Outil Analec et plateforme TXM

Historiquement au laboratoire Lattice, l'annotation de textes écrits se faisait à l'aide de l'outil Analec – « annotation de l'écrit » – qui a repris le modèle URS (unités, relations, schémas) issu de la thèse d'Antoine WIDLÖCHER (2008) et qui a proposé une interface humain-machine légère et rapide à utiliser (LANDRAGIN *et al*. 2012 ; LANDRAGIN 2016), que l'on voit en figure 1 avec un exemple de schéma d'annotation très complet (en trois couches) – différent de celui du projet Democrat – qui, pour des impératifs de rapidité, ne comporte qu'une seule couche d'annotation.

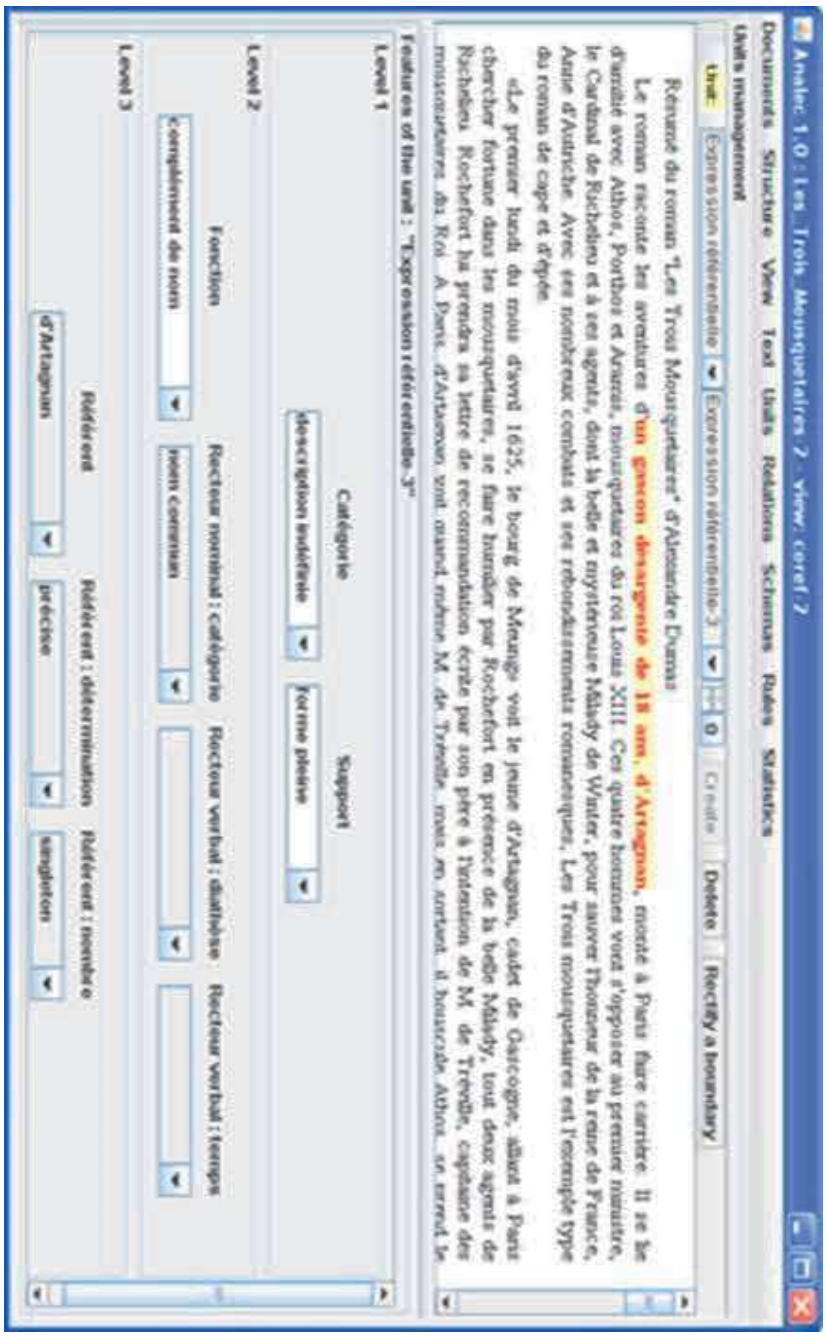

Figure 1 : exemple d'utilisation de l'outil Analec.

Analec ne propose cependant pas de fonctionnalités de gestion de corpus comportant plusieurs textes. Le projet Democrat s'est ainsi tourné vers la plateforme TXM (HEIDEN 2010), beaucoup plus complète et, qui plus est, modulaire, permettant la personnalisation à l'aide de macros, et incluant de nombreuses possibilités d'exploration de corpus. Le projet Democrat a permis à TXM d'évoluer

en intégrant le modèle URS (WIDLÖCHER 2008), selon l'approche d'Analec. En plus d'une plateforme de gestion et d'interrogation de corpus, TXM est maintenant également un outil d'annotation. La figure 2 montre une copie d'écran d'une des fenêtres de TXM, en l'occurrence celle permettant l'annotation.

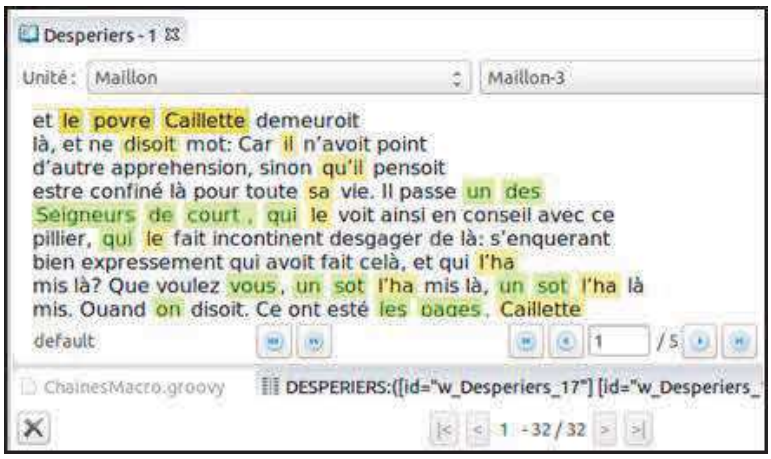

Figure 2 : annotation avec la plateforme TXM

De fait, les développements de la plateforme TXM se sont étendus sur toute la durée du projet Democrat. En plus de l'interface d'annotation inspirée de celle d'Analec, le projet a pu présenter de nouvelles macros – par exemple pour le calcul de la distance en nombre de mots séparant deux maillons, utilisée (entre autres) dans une étude de corrélation entre cette distance et le genre textuel (ROUSIER-VERCRUYSSEN & LANDRAGIN 2019) – et de nouvelles fonctionnalités de visualisation et d'exploration des chaînes de référence. Notons parmi elles l'adaptation aux chaînes d'un outil classique en linguistique de corpus outillée : le concordancier (*cf.* figure 3). Notons également le développement d'un nouveau mode de visualisation, nommé « diagramme de progression », qui permet d'apprécier d'un coup d'œil les progressions des références tout au long du texte : à chaque référent est attribuée une couleur, ainsi qu'une courbe qui s'élève d'un cran à chaque apparition d'une expression référentielle qui y réfère. La figure 4 montre ainsi les degrés de croissance de différentes courbes correspondant à autant de référents.

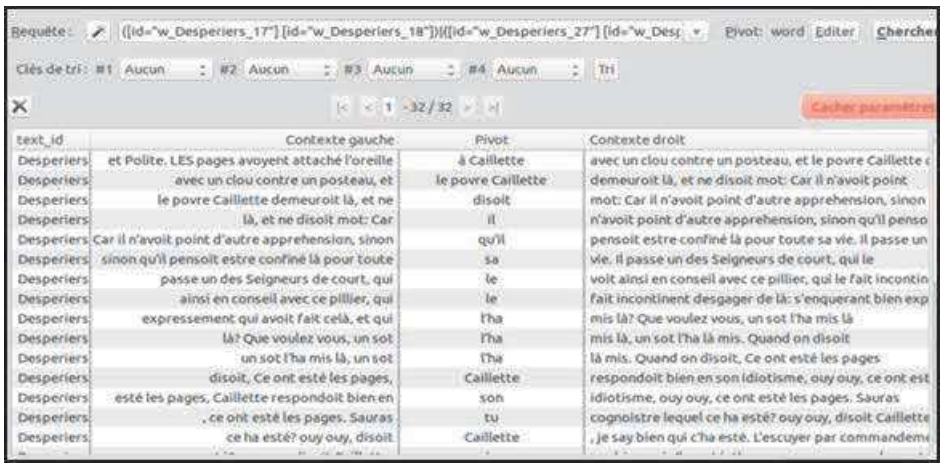

Figure 3 : concordancier appliqué aux chaînes de référence (QUIGNARD et al. 2018)

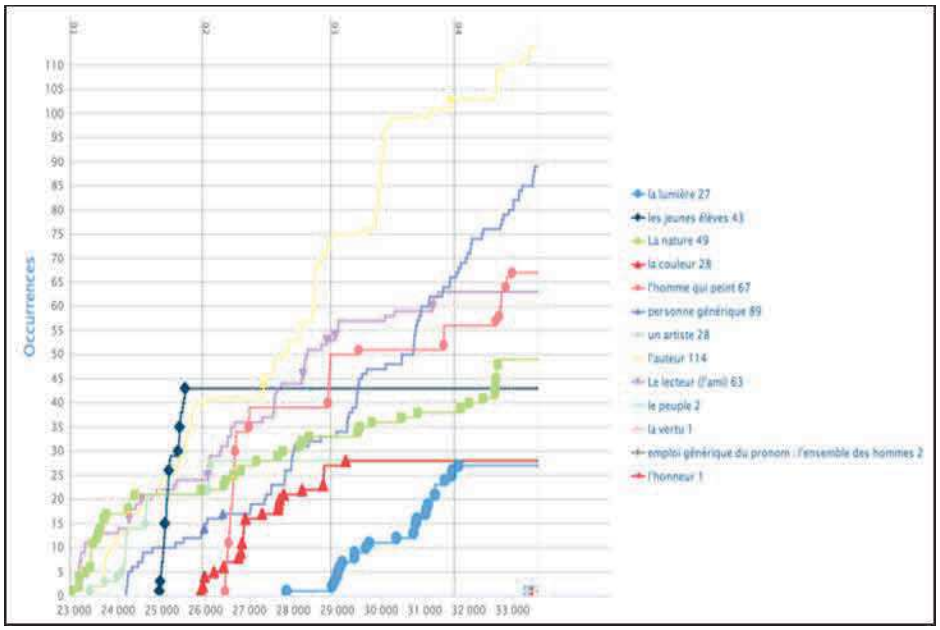

Figure 4 : diagrammes de progression (QUIGNARD et al. 2018)

3.2. SACR et CRViewer

TXM (HEIDEN 2010) constituait la plateforme choisie par le projet Democrat pour implémenter les fonctionnalités nécessaires au bon déroulement du projet. Néanmoins, comme dans tout projet, le travail des annotateurs a conduit certains d'entre eux à réfléchir à de nouvelles méthodes d'annotation – au sens ergonomique – ainsi qu'à de nouvelles métaphores graphiques pour visualiser des chaînes de référence.

Un premier exemple est celui de la visualisation des expressions référentielles sous la forme d'une suite de points colorés (LANDRAGIN 2016). Un second exemple

réside dans les travaux exploratoires de Bruno Oberlé : d'une part un script permettant d'annoter en ligne – script d'annotation des chaînes de référence, en abrégé SACR – d'autre part un outil dédié à la visualisation de chaînes. Les figures 5 et 6 présentent des copies d'écran de ces outils (OBERLÉ 2018). On reste ici dans le périmètre de la linguistique de corpus outillée. Il s'agit en effet d'interfaces humain-machine permettant aux linguistes d'annoter et d'analyser des annotations de manière efficace, et non de travaux relevant du domaine du traitement automatique des langues (TAL), travaux que nous allons maintenant présenter.

Figure 5 : SACR, script d'annotation des chaînes de référence (OBERLÉ 2018)

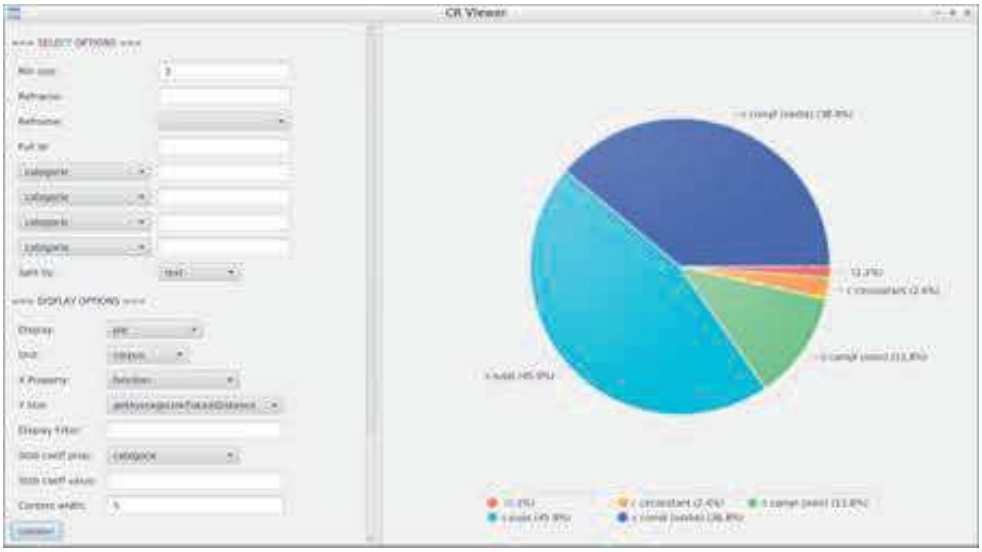

Figure 6 : outil de visualisation des chaînes de référence (CRviewer)

4. Systèmes de détection automatique des expressions référentielles et des chaînes de référence

4.1. COFR

Avant le projet Democrat, la détection automatique de chaînes de référence avait fait l'objet de travaux (trop nombreux pour être cités ici), qui peuvent se catégoriser selon deux approches : premièrement l'approche dite classique à base de systèmes de règles ; deuxièmement l'approche fondée sur de l'apprentissage artificiel. Contrairement à la première approche, la seconde nécessite de disposer d'un corpus, qui sert de corpus d'apprentissage. Un premier ensemble de techniques a été exploré avec le système CROC, « *Coreference Resolution for Oral Corpus* », développé par des membres du projet Democrat juste avant le lancement de celui-ci (DÉSOYER *et al*. 2018). Ce système présentait cependant une limitation forte : il nécessitait de disposer d'un texte déjà annoté en expressions référentielles.

Dans le cadre du projet Democrat, un nouvel élan a pu être donné à ces recherches, et les dernières techniques de l'apprentissage artificiel – à savoir le *deep learning*, ou apprentissage profond – ont pu être expérimentées. Plusieurs techniques ont fait l'objet de développements en parallèle et ont abouti à deux systèmes.

Le premier, nommé COFR, « *COreference for FRench* » (WILKENS *et al*. 2020), est un système bout-en-bout – c'est-à-dire capable de traiter du texte brut tout-venant – qui n'utilise aucune autre ressource que des plongements de mots ; il s'agit d'une adaptation du système à base de réseaux de neurones de (KANTOR & GLOBERSON 2019). Puisque le corpus Democrat dispose des annotations des singletons (expressions référentielles non reprises), contrairement au corpus de référence en langue anglaise (CoNLL-2012), COFR a été adapté pour détecter l'ensemble des mentions, qu'elles soient coréférentes ou qu'elles restent des singletons. Cela nous a conduits à diviser le système original en deux modules spécialisés, chacun avec un modèle entraîné séparément : un détecteur de mentions et un résolveur de coréférences. Le score CoNLL – métrique standard d'évaluation des systèmes de résolution de la coréférence – obtenu par COFR est de 75,00 %.

Comme un tel score n'est pas forcément parlant pour un linguiste, nous reproduisons en figure 7 un exemple de résultat obtenu par le système COFR, après son exécution sur un extrait de texte brut qui provient du roman *La Chartreuse de Parme* de Stendhal. Il va de soi que ce texte était inconnu de COFR, c'est-à-dire qu'aucun extrait de *La Chartreuse de Parme* ne faisait partie du corpus Democrat. Le système fait des erreurs, mais on peut observer la qualité des résultats sur les référents principaux, par exemple, pour cet extrait, l'abbé Blanès, qui apparaît ici en tant que référent n°4 :

> **[Le marquis]$_1$** professait [une haine vigoureuse] pour [les lumières] ; ce sont **[les idées]$_2$** , disait **[-il]$_1$** , **[qui]$_2$** ont perdu [l' Italie] ; **[il]$_1$** ne savait trop comment concilier [cette sainte horreur de [l' instruction]] , avec le désir de voir **[[son]$_1$ fils Fabrice]$_3$** perfectionner [l' éducation si brillamment commencée chez [les jésuites]] . Pour courir [le moins de risques possible] , **[il]$_1$** chargea **[le bon abbé Blanès]$_4$** ,

curé de [Grianta] , de faire continuer **[Fabrice]$_3$** **[ses]$_3$** études en **[latin]$_5$** . Il eût fallu que **[le curé lui-même]$_4$** sût **[cette langue]$_6$** ; or **[elle]$_6$** était [l' objet de [**[ses]$_3$** mépris]] ; [**[ses]$_3$** connaissances en [ce genre]] se bornaient à réciter , par cœur , [les prières de **[[son]$_3$ missel]$_7$**] , **[dont]$_7$** **[il]$_3$** pouvait rendre à peu près [le sens] à [**[ses]$_3$** ouailles] . Mais **[ce curé]$_4$** n' en était pas moins fort respecté et même redouté dans [le canton] ; **[il]$_4$** avait toujours dit que ce n' était point en [treize semaines] ni même en [treize mois] , que l' on verrait s' accomplir [la célèbre prophétie de [saint Giovita]] , le patron de [Brescia] . **[Il]$_4$** ajoutait , quand **[il]$_4$** parlait à [des amis sûrs] , que [ce nombre treize] devait être interprété d' **[une façon]$_8$** **[qui]$_8$** étonnerait bien de **[le monde]$_9$** , s' il était permis de tout dire ([1813]) . [Le fait] est que **[l' abbé Blanès]$_4$** , personnage d' [[une honnêteté] et d' [une vertu primitives]] , et de plus homme d' esprit , passait [toutes les nuits] à [le haut de **[[son]$_4$ clocher]$_{10}$**] ; **[il]$_4$** était fou d' [astrologie] . Après avoir usé [**[ses]$_4$** journées] à calculer **[[des conjonctions] et [des positions d' étoiles]]$_{11}$** , **[il]$_4$** employait [la meilleure part de [**[ses]$_4$** nuits]] à **[les]$_{11}$** suivre dans [le ciel] . Par suite de [**[sa]$_4$** pauvreté] , **[il]$_4$** n' avait d' [autre instrument] qu' [une longue lunette à [tuyau de carton]] . **[On]$_{12}$** peut juger de **[le mépris]$_{13}$** **[qu']$_{13}$** avait pour [l' étude de [les langues]] **[un homme]$_{14}$** **[qui]$_{14}$** passait [**[sa]$_{14}$** vie] à découvrir [l' époque précise de [**la chute de [les empires] et de [les révolutions]]$_{15}$**] **[qui]$_{15}$** changent [la face de **[le monde]$_9$**] . Que sais **[-je]$_{16}$** de plus sur **[un cheval]$_{17}$** , disait **[-il]$_{16}$** à **[Fabrice]$_3$** , depuis qu' **[on]$_{12}$** **[m']$_{16}$** a appris qu' en **[latin]$_5$** **[il]$_{17}$** s' appelle [equus] ? **[Les paysans]$_{18}$** redoutaient **[l' abbé Blanès]$_4$** comme [un grand magicien] : pour **[lui]$_4$** , à l' aide de **[la peur]$_{19}$** **[qu']$_{19}$** inspiraient [**[ses]$_4$** stations] dans **[le clocher]$_{10}$** , **[il]$_4$** **[les]$_{18}$** empêchait de voler . [**[Ses]$_4$** confrères les curés de [les environs]] , fort jaloux de [**[son]$_4$** influence] , **[le]$_4$** détestaient ; **[le marquis del Dongo]$_{20}$** **[le]$_4$** méprisait tout simplement , parce qu' **[il]$_{20}$** raisonnait trop pour [un homme de si bas étage] . **[Fabrice]$_3$** **[l']$_4$** adorait ; pour **[lui]$_4$** plaire **[il]$_3$** passait quelquefois [des soirées entières] à faire [des additions] ou [des multiplications énormes] . Puis **[il]$_3$** montait à **[le clocher]$_{10}$** : c' était [une grande faveur] et que **[l' abbé Blanès]$_4$** n' avait jamais accordée à personne ; mais **[il]$_4$** aimait **[cet enfant]$_4$** pour [**[sa]$_4$** naïveté] . Si **[tu]$_4$** ne deviens pas hypocrite , **[lui]$_4$** disait **[-il]$_4$** , peut-être **[tu]$_4$** seras [un homme] .

Figure 7 : exemple de résultat obtenu par le système COFR (WILKENS et al. 2020)

4.2. DeCOFR

Autre système développé dans le cadre du projet Democrat, DeCOFR, « *Detecting Coreference for Oral FRench* » (GROBOL 2019), est une adaptation du système de LEE *et al*. (2018). La première raison de l'adaptation réalisée est que le système de LEE *et al*. (2018) – qui dérive de travaux et donc de systèmes précédemment développés (LEE *et al*. 2017) – opère systématiquement au niveau d'un document entier, ce qui paraît raisonnable au vu de la nature discursive des chaînes de coréférences, mais pose un problème de taille de mémoire : le document entier doit être gardé en mémoire, ce qui entraîne des besoins de calculs potentiellement très élevés. LEE *et al*. (2018) proposent de compenser ce problème en effectuant à chaque étape une série d'élagages un peu brutaux, mais le revers de la médaille est que cela complique la mise en œuvre et rend le processus d'apprentissage moins efficace. Au final, l'apprentissage est toujours très gourmand en mémoire et en calculs.

La deuxième raison de l'adaptation réalisée pour DeCOFR est le fait que le système de LEE *et al*. (2018) ne fait pas de distinction entre des expressions référentielles et des expressions ayant la même forme de surface mais non référentielles. Il ne détecte que les mentions susceptibles d'appartenir à des chaînes de coréférences, et pas les mentions qui restent isolées – les fameux singletons. Ce n'est pas un problème pour son entraînement avec le corpus CoNLL-2012 (cadre dans lequel le système de LEE *et al*. (2018) a été développé), mais c'en est un quand on considère un corpus comprenant des singletons. Ce qui est le cas du corpus ANCOR (MUZERELLE *et al*. 2013) exploité lors des premières expérimentations du système DeCOFR – et qui est à l'origine du nom de ce système. Et c'est le cas également du corpus Democrat. Pour pallier ces problèmes, DeCOFR cible en tenant compte du contexte immédiat plutôt que du document entier, et opère une détection des mentions, en tant que telles, avec prise en compte des singletons, en préalable à la détection des coréférences.

Soulignons que les recherches entreprises dans le projet Democrat sont initiatrices et ouvrent la voie à la détection automatique des chaînes de références pour la langue française. Le projet Democrat livre et rend publics deux systèmes bout-en-bout fondés sur les techniques les plus récentes d'apprentissage profond. Il permet ainsi à la communauté internationale de disposer d'équivalents des systèmes récemment développés pour la langue anglaise, espagnole ou polonaise. Avec la livraison effectuée en juin 2019 du corpus Democrat, chaque chercheur peut de plus tester la détection automatique de coréférences sur le français et développer son propre système.

Notons qu'une comparaison des systèmes COFR et DeCOFR, ainsi que des analyses des spécificités de chacun des systèmes en fonction de leurs points forts et de leurs points faibles, dépassent largement les objectifs initiaux du projet Democrat, et en constituent des perspectives que nous n'aborderons pas plus ici.

5. Linguistique textuelle, linguistique de corpus et projet Democrat

5.1. Un bilan des apports du projet Democrat

Lors de la soumission du projet Democrat, en 2015, nous ne disposions : (i) ni de description intégrée permettant de modéliser les chaînes de référence, et de prédire leur typologie ou leurs comportements textuels, en français ; (ii) ni de corpus permettant d'apprécier l'évolution historique de leur composition ; (iii) ni d'outil de visualisation et d'exploration des chaînes de référence ; (iv) ni de logiciel de traitement automatique des langues capable d'extraire de textes rédigés en français les expressions référentielles et les chaînes de référence. Dans la perspective de remédier à ces manques, et ainsi de rattraper des initiatives du même genre explorées pour d'autres langues que le français (RECASENS 2010 ; OGRODNICZUK *et al*. 2015), le projet Democrat s'est donné pour ambition d'apporter de nouveaux résultats sur ces quatre aspects, qui ont constitué les quatre volets du projet – présentés dans les quatre premières sections de cet article.

Aujourd'hui, un an après la fin du projet, nous disposons : (i) d'une description intégrée des chaînes de référence, qui se décline en termes discursifs, diachroniques et inter-genres (SCHNEDECKER *et al*. 2017) ; (ii) d'un corpus de textes écrits en français dont les chaînes de référence ont été annotées (LANDRAGIN 2019) ; (iii) de plusieurs outils qui permettent de visualiser et d'explorer ces chaînes (QUIGNARD *et al*. 2018 ; OBERLÉ 2018) ; (iv) de deux systèmes de traitement automatique des langues capables de traiter du texte tout-venant pour en extraire les expressions référentielles et les chaînes de référence (GROBOL 2019 ; WILKENS *et al*. 2020). Nous pouvons affirmer que le programme de travail et les résultats du projet ont permis des avancées significatives pour l'étude des chaînes de référence.

5.2. Mise en perspective des apports par rapport à la thématique du numéro

Surtout, nous pouvons affirmer que le projet Democrat a permis de franchir plusieurs pas décisifs dans la méthodologie d'appréhension, de visualisation, d'exploration et d'analyse des chaînes de référence, à la croisée de la linguistique textuelle et de la linguistique de corpus. Nous avons vu que, par définition, une chaîne de référence constitue un objet d'étude relevant de la linguistique textuelle. Nous avons vu également à quel point le projet Democrat a mis en avant une méthodologie quasi systématiquement centrée sur le recours à un corpus – corpus imaginé, élaboré, conçu, constitué, annoté, analysé puis exploité par les membres du projet. Chacune des avancées explore ainsi de nouveaux rapports entre linguistique textuelle et linguistique de corpus. Le corpus Democrat est un résultat concret dont la richesse n'a pas d'équivalent en langue française, surtout si l'on considère l'aspect diachronique. Les extensions de la plateforme TXM en permettent une exploitation ergonomique, autorisant une multiplicité d'approches et de méthodes.

Si nous ne devions retenir que deux exemples ouvrant des perspectives intéressantes à la croisée de la linguistique textuelle et de la linguistique de corpus, citons le concordancier de la figure 3 et les diagrammes de progression de la figure 4. Avec son adaptation à un objet linguistique qui dépasse largement le cadre du marquable ainsi que celui de la phrase, le concordancier est la preuve qu'il est

désormais possible de ré-exploiter des outils classiques de linguistique de corpus pour des besoins qui relèvent de la linguistique textuelle. Avec sa capacité à transformer des phénomènes linguistiques en courbes colorées, les diagrammes de progression montrent qu'il est possible de procéder à des visualisations sur l'ensemble d'un texte. Si l'on examine la figure 4, on s'aperçoit que l'axe horizontal – correspondant à la linéarité du texte – est scindé en quatre parties. Il s'agit de fait d'une structuration textuelle : chaînes de référence et structures textuelles peuvent ainsi être analysées simultanément, sur un même graphique.

C'est l'un des apports du projet Democrat, et c'est surtout un principe qui ne dépend pas uniquement du contexte de notre projet. Typiquement, n'importe quel objet relevant de la linguistique textuelle – du moment qu'il est annoté dans un corpus – peut faire l'objet de diagrammes de progression, ainsi que d'une comparaison croisée avec les structures textuelles (à condition qu'elles aussi soient annotées dans le corpus). Les perspectives de recherche qui sont rendues possibles par les avancées de Democrat auront probablement des répercussions sur la manière d'articuler linguistique textuelle et linguistique de corpus, au-delà des seuls phénomènes de référence et de coréférence.

Conclusion

Un an après la fin du projet Democrat, alors que les derniers résultats sont en cours de publication, nous avons dressé un bilan du projet en insistant particulièrement sur trois aspects. Premièrement, sur la nécessité de disposer d'un corpus annoté et, qui plus est, d'un corpus de grande taille. Atteindre le million de mots n'est pas forcément nécessaire. En revanche, dépasser le seuil symbolique de 100 000 annotations l'est, d'une part parce que cela permet des analyses qualitatives et quantitatives pouvant faire appel à des indicateurs statistiques de significativité, d'autre part parce que cela permet de nourrir des systèmes de TAL fondés sur l'apprentissage profond – de tels systèmes étant par nature gourmands en données.

Deuxièmement, nous avons insisté sur l'outillage systématique de notre méthodologie. Aussi bien la constitution du corpus, que son annotation et son exploration sont outillées. Cet outillage ne sert pas seulement à produire des tableaux et des graphiques informatifs. Il sert également à dynamiser les recherches : plus on visualise de diagrammes et de représentations diverses de l'objet d'étude exploré, plus on peut envisager de nouvelles hypothèses de travail, qui conduisent elles-mêmes à de nouvelles explorations des données annotées.

Enfin, nous avons souligné l'interconnexion permanente entre linguistique théorique et descriptive, linguistique textuelle, linguistique de corpus, interfaces humain-machine, informatique et apprentissage profond, c'est-à-dire intelligence artificielle. Le projet Democrat a exploré et produit des résultats dans chacune de ces disciplines, et a montré à quel point celles-ci étaient complémentaires. Avec les chaînes de référence, nous avons montré que l'on pouvait garder une démarche de linguistique textuelle tout en constituant un corpus, et que l'on pouvait explorer les données incluses dans un corpus en gardant en tête des préoccupations relevant de la linguistique du discours.

BIBLIOGRAPHIE

CHAROLLES Michel (2002), *La référence et les expressions référentielles en français*, Paris-Gap, Ophrys.

CHASTAIN Charles (1975), Reference and Context, in : GUNDERSON Keith (éd.), *Language Mind and Knowledge*, Minneapolis, University of Minnesota Press, p. 194-269.

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

DELABORDE Marine, LANDRAGIN Frédéric (2019), En quoi le pronom *on* a-t-il une valeur anaphorique ? Le cas des successions d'occurrences de *on*, *Cahiers de praxématique 72*, p. 1-19.

DÉSOYER Adèle, LANDRAGIN Frédéric, TELLIER Isabelle, LEFEUVRE Anaïs, ANTOINE Jean-Yves, DINARELLI Marco (2018), Coreference resolution for French oral data: Machine learning experiments with ANCOR, in : *Computational Linguistics and Intelligent Text Processing, Seventeenth International Conference (CICLing 2016, Konya, Turquie)*, Berlin, Springer Verlag, p. 507-519.

GROBOL Loïc (2019), Neural Coreference Resolution with Limited Lexical Context and Explicit Mention Detection for Oral French, in : *Second Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC19-NAACL)*, Minneapolis, United States.

HEIDEN Serge (2010), The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme, in : *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation*, Waseda University, Sendai, Japan, p. 389-398.

KANTOR Ben, GLOBERSON Amir (2019), Coreference Resolution with Entity Equalization, in : *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*, Florence, Italy, p. 673-677.

KARTTUNEN Lauri (1976), Discourse Referents, in : MCCAWLEY James D. (éd.), *Syntax and Semantics 7*, New York, Academic Press, p. 363-385.

LANDRAGIN Frédéric (2016), Conception d'un outil de visualisation et d'exploration de chaînes de coréférences, in : *Proceedings of the Thirteen International Conference on Statistical Analysis of Textual Data (JADT 2016)*, Nice, France, p. 109-120.

LANDRAGIN Frédéric (éd.) (2019), Democrat Corpus, https://hdl.handle.net/11403/democrat.

LANDRAGIN Frédéric (2021), Méthodologie pour la préparation d'une campagne d'annotation manuelle d'expressions référentielles, in : FREROT Cécile, PECMAN Mojca (éds), *Des corpus numériques à l'analyse linguistique en langues de spécialité*, Grenoble, UGA Éditions, p. 37-60.

LANDRAGIN Frédéric, POIBEAU Thierry, VICTORRI Bernard (2012), ANALEC: a New Tool for the Dynamic Annotation of Textual Data, in : *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC)*, Istanbul, Turkey, p. 357-362.

LANDRAGIN Frédéric, SCHNEDECKER Catherine (éd.) (2014), *Les chaînes de référence, Langages 195*, Paris, Larousse.

LEE Kenton, HE Luheng, LEWIS Mike, ZETTLEMOYER Luke (2017), End-to-end Neural Coreference Resolution, in : *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)*, Copenhagen, Denmark, p. 188-197.

LEE Kenton, HE Luheng, ZETTLEMOYER Luke (2018), Higher-Order Coreference Resolution with Coarse-to-Fine Inference, in : *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics, ACL*, New Orleans, Louisiana, Vol. 2, p. 687-692.

MUZERELLE Judith, LEFEUVRE Anaïs, ANTOINE Jean-Yves, SCHANG Emmanuel, MAUREL Denis, VILLANEAU Jeanne & ESHKOL Iris (2013), ANCOR, premier corpus de français parlé d'envergure annoté en coréférence et distribué librement, in : *Actes de la vingtième Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2013)*, Les Sables-d'Olonne, p. 555-563.

OBERLÉ Bruno (2018), SACR: A Drag-and-Drop Based Tool for Coreference Annotation, in : *Proceedings of the 11th Language Resources and Evaluation Conference (LREC 2018)*, Miyazaki, Japan, p. 389-394.

OGRODNICZUK Maciej, GŁOWIŃSKA Katarzina, KOPEĆ Mateusz, SAVARY Agata, ZAWISŁAWSKA Magdalena (2015), *Coreference in Polish: Annotation, Resolution and Evaluation*, Berlin, Walter De Gruyter.

QUIGNARD Matthieu, HEIDEN Serge, LANDRAGIN Frédéric, DECORDE Matthieu (2018), Textometric Exploitation of Coreference-annotated Corpora with TXM: Methodological Choices and First Outcomes, in : *Fourteenth International Conference on the Statistical Analysis of Textual Data (JADT 2018)*, Roma, Italy, p. 610-615.

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

ROUSIER-VERCRUYSSEN Lucie, LANDRAGIN Frédéric (2019), Interdistance et instabilité au sein des chaînes de référence : indices textuels ?, *Discours 25*, p. 3-32.

SCHNEDECKER Catherine (1997), *Nom propre et chaînes de référence*, Paris, Klincksieck.

SCHNEDECKER Catherine, GLIKMAN Julie, LANDRAGIN Frédéric (éd.) (2017), Les chaînes de référence en corpus, *Langue française 195*, Paris, Armand Colin.

VAN DEEMTER Kees, KIBBLE Roger (2000), On Coreferring: Coreference Annotation in MUC and Related Schemes, *Computational Linguistics 26(4)*, p. 615-623.

WIDLÖCHER Antoine (2008), Analyse macro-sémantique des structures rhétoriques du discours : cadre théorique et modèle opératoire, thèse de doctorat, Caen, Université de Caen.

WILKENS Rodrigo, OBERLÉ Bruno, LANDRAGIN Frédéric, TODIRASCU Amalia (2020), French Coreference for Spoken and Written Language, in : *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*, Marseille, France, p. 80-89.