Méthodologie pour la préparation d'une campagne d'annotation manuelle d'expressions référentielles

Frédéric Landragin

Méthodologie pour la préparation d'une campagne d'annotation manuelle d'expressions référentielles

Draft auteur

Frédéric Landragin
Laboratoire Lattice
CNRS, ENS Paris, PSL Research University, Université Sorbonne Nouvelle

Résumé : Le contexte de ce travail est l'annotation manuelle des expressions référentielles qui apparaissent dans des textes écrits en français, de différentes périodes et de différents genres textuels. Le but est la constitution du corpus du projet ANR DEMOCRAT, dans lequel sont annotées les expressions référentielles et les chaînes de référence. Nous présentons une série d'expérimentations d'annotation réalisées en 2016 et en 2017, au début du projet, faisant intervenir plusieurs méthodes et plusieurs annotateurs. Les retours d'expériences ainsi collectés ont servi à spécifier la procédure d'annotation du projet, et ont conduit à la mise en œuvre du corpus, disponible en ligne depuis avril 2019. Parmi les aspects que nous mettons en avant se trouvent d'une part la sélection des expressions à annoter, d'autre part l'utilisation – ou non – d'un outil de traitement automatique des langues en tant que pré-annotateur. Nous discutons chacun de ces deux aspects en nous appuyant sur les résultats de sessions d'annotation chronométrées : quantité de texte traitée ; calcul de l'accord inter-annotateurs ; impressions et difficultés rencontrées par les annotateurs. Nous soulignons notamment l'aspect « robotique » de certaines facettes de la tâche d'annotation, que nous discutons. Nous concluons avec la rédaction et le test du manuel d'annotation.

Mots clés : référence, coréférence, expression référentielle, annotation chronométrée.

1. Introduction

La constitution et l'annotation manuelle d'un corpus en français, regroupant des textes écrits de différents genres textuels et de différentes époques, a fait partie des objectifs initiaux du projet ANR DEMOCRAT (LANDRAGIN 2016). L'annotation a concerné les expressions référentielles et les chaînes de référence. Pour chaque expression référentielle rencontrée, les annotateurs ont indiqué le référent concerné, puis les chaînes ont été construites automatiquement *a posteriori*, par regroupement de toutes les expressions désignant le même référent. Le corpus, récemment disponible en ligne – voir le site http://www.lattice.cnrs.fr/democrat/ –, a une taille d'environ 500 000 mots, ce qui a permis

d'obtenir plus de 200 000 expressions référentielles annotées. La tâche d'annotation a donc représenté un important travail, dont l'utilité est triple : il s'agit premièrement de fournir un corpus de référence qui serve à toute la communauté s'intéressant à la référence et à la coréférence, deuxièmement de permettre des analyses statistiques grâce à des données en quantité suffisante, et troisièmement de nourrir des systèmes de traitement automatique des langues (TAL) fonctionnant par apprentissage artificiel, de manière à consolider la voie de la détection automatique des expressions référentielles et des chaînes de référence en français, suite aux expérimentations déjà réalisées (DESOYER *ET AL.* 2014) avec le corpus ANCOR (MUZERELLE *ET AL.* 2014), seul corpus de taille comparable – 115 000 relations anaphoriques annotées – disponible librement pour la langue française.

Compte tenu de la taille du corpus et du caractère chronophage de l'annotation manuelle, la procédure d'annotation a été anticipée de manière scrupuleuse, avec la réalisation de plusieurs pré-annotations sur des échantillons de petite taille. C'est sur la base de ces pré-annotations qu'a été quantifiée la durée d'annotation du corpus complet, et que les détails de la procédure ont été définis. Dans cet article, qui relève pleinement du domaine de la linguistique de corpus et de ses méthodologies, nous décrivons les discussions et les expérimentations qui ont accompagné ces pré-annotations. Nous n'abordons la question des langues de spécialité qu'indirectement, à travers la notion de genre textuel : certains textes du corpus relèvent par exemple de fiches de la Wikipédia et mettent potentiellement en œuvre une langue de spécialité qui influe sur les référents et les expressions référentielles. Nous considérons donc que nos remarques quant aux genres textuels sont valables également pour les langues de spécialité.

Après une première section portant sur les phénomènes linguistiques considérés et leur annotation en corpus, nous présentons plusieurs stratégies d'annotation identifiées *a priori*. Chacune de ces stratégies a mené à des expérimentations dont nous présentons et discutons les résultats. Nous soulevons ensuite la question de l'appui de l'annotateur sur des pré-annotations automatiques, et notamment sur un pré-repérage des chunks nominaux. Là aussi, nous présentons les résultats d'expérimentations d'annotation, effectuées avec ou sans ce pré-repérage. Nous discutons des caractéristiques de l'annotation manuelle des expressions référentielles, et présentons les choix retenus pour le projet DEMOCRAT, dont nous soulignons les exploitations possibles. Nous concluons en décrivant les étapes qui ont suivi dans le projet, ainsi que les perspectives de recherche qui restent ouvertes, notamment celles qui concernent la linguistique de corpus outillée.

2. La référence et les expressions référentielles

2.1. Les phénomènes linguistiques

La référence est un objet linguistique très vaste, qui a fait l'objet de très nombreuses publications (CHAROLLES 2002). Les expressions référentielles classiquement étudiées relèvent de plusieurs catégories morphosyntaxiques : noms propres, groupes nominaux définis, démonstratifs ou indéfinis, pronoms personnels, pronoms possessifs. Selon l'approche

retenue, on peut même retenir comme expressions référentielles les sujets non exprimés de verbes (LANDRAGIN 2011). Dans une simple phrase comme « Jeanne a vu sa fille hier et l'a félicitée », on trouve de nombreuses expressions référentielles : « Jeanne », « sa », « sa fille », « hier », « l' », auxquelles on pourrait ajouter le sujet non exprimé du verbe « féliciter ». Ce dernier, « Jeanne » et « sa » sont tous coréférents et forment ainsi une chaîne de référence (SCHNEDECKER 1997).

2.2. Leur annotation

Les expressions référentielles et les chaînes de référence parsèment les textes, contribuant à leur ancrage contextuel et à leur cohérence textuelle. Les annoter pose des difficultés (LANDRAGIN ET SCHNEDECKER 2014 ; SCHNEDECKER *ET AL.* 2017), qui relèvent d'une part de leur détection – on peut « rater » non seulement un sujet zéro, mais aussi un pronom comme « l' » –, d'autre part de leur renvoi plus ou moins clair vers un référent du monde extralinguistique : l'ambiguïté comme la sous-détermination font partie intégrante de la tâche (LANDRAGIN 2011), et c'est à l'annotateur qu'il revient de rendre compte du phénomène rencontré, en adéquation avec les consignes indiquées dans le manuel d'annotation.

Car, comme toute tâche d'annotation manuelle, celle des expressions référentielles et des chaînes de référence nécessite de suivre une méthodologie stricte et bien définie (FORT 2012). Le manuel d'annotation sert un peu de pivot méthodologique : son écriture nécessite d'avoir bien étudié le problème, d'avoir récolté des exemples prototypiques aussi bien que particuliers, et d'avoir spécifié un schéma d'annotation – ou structure des annotations – qui décrit les différentes étiquettes possibles ainsi que les façons de les attribuer à des portions de texte (« marquables ») ou à des structures qui se détachent de celui-ci (relations entre marquables ; chaînes d'annotations).

Pour déterminer les grandes lignes de l'annotation du projet DEMOCRAT, nous nous sommes inspiré de nombreuses réalisations existantes, et notamment des corpus annotés qui sont souvent exploités en TAL, car les possibilités d'exploitation ultérieure du corpus DEMOCRAT influencent fortement le choix de son schéma d'annotation. Nos choix reposent ainsi sur ceux effectués pour d'autres langues que le français : corpus OntoNotes (PRADHAN *ET AL.* 2011), corpus ACE (DODDINGTON *ET AL.* 2004), corpus « Phrase Detectives » (CHAMBERLAIN *ET AL.* 2016), corpus d'articles scientifiques (SCHÄFER *ET AL.* 2012), corpus WikiCoref (GHADDAR AND LANGLAIS 2016), corpus spécifique au polonais (OGRODNICZUK *ET AL.* 2015). Des corpus annotés pour d'autres phénomènes que la référence – ou pour un sous-ensemble des phénomènes qui nous intéressent – existent également, comme le corpus WiNER annoté en entités nommées (GHADDAR AND LANGLAIS 2017). Pour la langue française, il existe une initiative de ce genre : l'annotation référentielle du corpus arboré de Paris 7 (SAGOT *ET AL.* 2012). Mais c'est surtout le corpus ANCOR (MUZERELLE *ET AL.* 2014) qui a été une source d'inspiration décisive, ainsi que le corpus MC4 (LANDRAGIN 2018) en tant que pré-expérimentation à petite échelle.

Le projet MC4 s'était intéressé aux multiples facteurs morphologiques, syntaxiques, sémantiques et pragmatiques qui interviennent lors de la résolution des références. La procédure d'annotation résultante avait impliqué l'annotation manuelle d'une dizaine de facteurs considérés comme déterminants. Au final, le corpus n'a pas dépassé 5 000 expressions référentielles annotées, l'ampleur de la tâche s'avérant incompatible avec des ambitions de grande taille de corpus. Pour le projet DEMOCRAT, il n'était pas question de reproduire une procédure aussi détaillée. Revenant vers les classiques OntoNotes ou ACE, nous avons ainsi choisi d'annoter uniquement le résultat de la résolution de la référence, sans rendre compte des ambiguïtés comme le fait « Phrase Detectives ».

2.3. L'outil d'annotation et le choix du schéma d'annotation

Gérer des expressions référentielles et des chaînes de référence peut se matérialiser de deux façons très différentes : soit on saisit un identifiant de référent pour chaque expression (les chaînes se déduisant alors des annotations), soit on construit les chaînes et on leur affecte l'ensemble des expressions référentielles qui s'y rapportent. Tous les outils ne permettent pas cette dernière construction, qui nécessite une succession d'opérations à la souris. Plus que cela, les outils qui le permettent le font de différentes manières, chacune se caractérisant par une ergonomie qui lui est propre. Nous avons ainsi testé plusieurs outils conçus pour les chaînes d'annotations : MMAX2 (MÜLLER AND STRUBE 2006), GLOZZ (WIDLÖCHER AND MATHET 2012) et ANALEC (LANDRAGIN *ET AL*. 2012). C'est finalement l'ergonomie d'ANALEC qui a été retenue, mais pas pour l'interface de construction de chaînes : au contraire, les tests ont montré que la solution la plus rapide consistait à saisir un identifiant de référent pour chaque expression référentielle, même si ce choix impliquait de saisir plusieurs fois les identifiants des référents fréquents. En effet, manipuler un objet abstrait couvrant potentiellement le texte entier s'est avéré bien plus délicat (et propice à de potentielles erreurs) que de saisir des identifiants localement, au niveau du marquable qu'est l'expression. C'est aussi et surtout l'implémentation de la complétion automatique qui a permis à ANALEC (dans une version améliorée, donc) d'arriver en première place. Grâce à la complétion, la saisie de l'identifiant ne nécessite que ses premiers caractères, ce qui permet une grande efficacité. Nous noterons que, depuis 2016, les outils ont évolué : le projet DEMOCRAT a non seulement permis l'exploration de nouvelles interfaces graphiques (OBERLE 2018), mais aussi et surtout d'implémenter de nouvelles fonctionnalités d'annotation – inspirées d'ANALEC – dans la plateforme TXM (HEIDEN *ET AL*. 2010), bien plus répandue.

Au final, le schéma d'annotation minimal de DEMOCRAT ne comprend qu'un seul champ : celui de l'identifiant du référent. C'est le schéma retenu pour l'ensemble des expérimentations qui font l'objet de cet article.

2.4. L'accord inter-annotateurs

S'il n'est pas possible de valider une procédure d'annotation – à moins de disposer d'un « gold standard » –, il est possible d'évaluer sa reproductibilité, ce qui donne une indication précieuse sur l'intérêt des annotations d'un corpus (MATHET ET WIDLÖCHER 2016). Ceci se

fait en impliquant plusieurs annotateurs – deux ou trois dans nos expérimentations – sur le même texte, puis en calculant l'accord inter-annotateurs.

Plusieurs indicateurs statistiques sont couramment utilisés : α, π, κ (CARLETTA 1996 ; KRIPPENDORFF 2012) et désormais γ (MATHET *ET AL.* 2015 ; MATHET 2017). Avec notre schéma d'annotation, le calcul de ces indicateurs nécessite une adaptation. En effet, deux annotateurs peuvent choisir des identifiants différents pour le même référent. Tous les identifiants doivent donc être homogénéisés pour pouvoir être comparés. Nous avons ainsi développé un script d'adaptation, de manière à obtenir des chiffres signifiants.

Comme les résultats ne s'avèrent pas dépendre de la nature des expérimentations réalisées, donnons tout de suite les chiffres obtenus. Nous avons calculé l'ensemble des indicateurs pour chacune des expérimentations, et le premier constat est que les chiffres ne sont jamais très élevés. Ceci est dû à la nature même de l'objet d'étude : la coréférence et l'anaphore sont des phénomènes complexes, pour lesquels les interprétations peuvent varier d'un annotateur à l'autre sans pour autant que les annotations en deviennent inutilisables. La communauté s'est habituée à des taux d'accord modestes (ARTSTEIN AND POESIO 2008), et les corpus constitués dans d'autres langues que le français obtiennent des scores comparables aux nôtres. Le deuxième constat est que les chiffres augmentent expérimentation après expérimentation : comme les annotateurs comparent leurs productions entre chaque série d'expérimentation, cela contribue à rendre peu à peu leur comportement plus homogène. Ainsi, α oscille entre 0,662 et 0,733 ; π entre 0,66 et 0,732 ; κ entre 0,661 et 0,733. Ces résultats sont cohérents avec l'état de l'art : aucune de nos expérimentations ne peut être rejetée sur la seule base d'un mauvais accord inter-annotateurs. Comme c'est γ qui correspond le mieux au type d'annotation qui nous intéresse, notons que sa valeur plutôt médiocre lors de la toute première expérimentation – soit 0,53 – augmente peu à peu pour atteindre la valeur tout à fait honorable de 0,73. Toutes nos expérimentations peuvent donc conduire à un passage à l'échelle.

3. Identifier des stratégies d'annotation

3.1. Annotation systématique des expressions référentielles

Nos premières expérimentations d'annotation se sont déroulées sur des textes narratifs, en l'occurrence des extraits de romans libres et gratuits, disponibles sur la plateforme Wikisource. Il s'agit donc de textes littéraires, et les discussions ont vite porté sur le filtrage ou non des référents : faut-il annoter toutes les expressions référentielles, ou seulement celles qui réfèrent à des personnages humains, à des êtres animés, à des objets concrets ? Nous constatons par exemple que les expressions référentielles temporelles – comme « hier » dans l'exemple de la section précédente – sont très peu reprises et ne forment donc que rarement des chaînes de référence intéressantes à étudier. Dans ce cas, faire l'impasse sur l'annotation de ces expressions, du moins quand elles restent des « singletons », permettrait d'aller plus vite à l'essentiel.

Cependant, plusieurs arguments doivent être pris en compte dans une telle décision. Premièrement, l'annotation systématique de toutes les expressions référentielles permet de nourrir un système d'apprentissage dédié à la détection des expressions référentielles. Certes, il s'agit là d'une tâche différente de celles constituant à détecter les chaînes de référence – la tâche ultime, pourrait-on dire – mais c'est néanmoins une tâche très complexe et intéressante en soi, qu'il ne faudrait pas écarter trop rapidement. Notamment, c'est une tâche différente de celles fréquemment mises en compétition en TAL, à savoir la détection des entités nommées (*grosso modo*, les noms propres seulement) et la résolution des pronoms anaphoriques.

Deuxièmement, il n'est jamais possible de savoir à coup sûr si l'expression en cours d'annotation va être reprise ultérieurement ou non. Autrement dit, la tâche d'annotation peut comporter des retours en arrière dans le texte si l'annotateur s'aperçoit qu'il a initialement considérée une expression comme singleton, à tort. Or un retour en arrière va totalement à l'encontre de la rapidité et de l'efficacité, car il faut alors retrouver l'expression dont on n'a potentiellement qu'un souvenir vague, peut-être pas de contenu, en tout cas de sa localisation dans le texte.

Troisièmement, se demander à chaque expression si elle a des chances d'être un singleton ou de faire partie d'une chaîne de référence va également à l'encontre de l'efficacité. Mieux : tous les annotateurs qui ont été questionnés sont unanimes sur le fait que se poser trop de questions est souvent contre-productif. Cette remarque est importante, car elle hisse l'aspect « robotique » de l'annotation manuelle au statut d'avantage et non d'inconvénient, contrairement à ce que l'on pourrait penser *a priori*. Autrement dit, il vaut mieux annoter l'intégralité des expressions en se posant peu de questions, plutôt qu'annoter moins d'expressions en se posant plus de questions…

Au final, l'annotation systématique a été mise en avant et a fait l'objet de plusieurs expérimentations. Notamment, nous avons vite convenu de choisir des identifiants de référents qui soient simples et faciles à retenir pour les référents fréquents, et de choisir des identifiants complets et discriminants pour des expressions susceptibles d'être peu reprises. Surtout, pour les expressions qui resteront clairement des singletons, nous avons choisi d'employer un code dédié – « SI » comme singleton –, ce qui permet d'augmenter encore l'efficacité. Ce code ne doit être utilisé qu'en cas de certitude, car un retour en arrière s'avère particulièrement préjudiciable, aucun souvenir d'identifiant ne permettant de tenter une recherche dans le texte déjà annoté.

3.2. Annotation de certaines expressions référentielles

Nous avons également réalisé plusieurs expérimentations impliquant une sélection des expressions référentielles à annoter. La sélection a dans un premier temps reposé sur la nature des référents – humains et animaux, par exemple – et dans un deuxième temps sur les thèmes explorés dans le texte : nous faisions notamment l'hypothèse qu'un texte de genre narratif mettait en avant les personnages, alors qu'un texte non narratif comme un extrait de presse écrite (L'Est Républicain) mettait en avant plutôt des événements. Nous avons également

exploré l'annotation de fiches issues de la Wikipédia, en supposant que le titre même de la fiche indiquait le ou les référents les plus pertinents à annoter. De fait, nos expérimentations ont montré que le choix des référents intéressants n'avait pas à être décidé *a priori*, mais revenait surtout à l'annotateur lors de sa lecture : c'est lui qui sait gérer l'orientation thématique de la fiche, ainsi que les spécificités de la langue de spécialité, pour certaines fiches. Ceci nous a conduit à ne conserver qu'un seul critère de sélection : la nature des référents.

Nous avons un moment envisagé d'annoter en deux étapes : une première pour les référents sélectionnés puis, si besoin, une seconde pour l'ensemble des référents. Cette procédure s'est révélée peu convaincante, notamment parce que certains annotateurs ressentaient comme difficile la reprise de leurs annotations en vue de les augmenter : il leur était préférable d'accomplir la totalité de la tâche en un seul passage, plutôt que de devoir se replonger dans un corpus déjà partiellement annoté.

Au final, nous soulignons deux avantages et trois inconvénients : filtrer les référents selon leur nature peut être un avantage quand le filtrage va dans le même sens que les préoccupations de recherche de l'annotateur, dans le cas où l'annotateur est aussi un chercheur membre du projet (mais est-ce vraiment souhaitable pour la qualité du corpus final ?). Le deuxième avantage réside dans la rapidité de la première passe : quand on ne s'intéresse qu'aux humains et aux animaux, l'annotation s'avère bien plus rapide, et parfois aussi plus agréable. Du côté des inconvénients, notons que si les types de référents retenus sont nombreux, on a tendance à se mélanger un peu les pinceaux et, à l'extrême, à passer plus de temps à se demander si tel ou tel référent est pertinent, plutôt qu'à annoter. Notons également que les types de référents varient d'un genre textuel à l'autre, et que l'homogénéité des annotations peut être mise en péril. Retenons surtout que tout corpus annoté uniquement pour certains référents restera incomplet pour une étude exhaustive de la référence et pour des applications de TAL comme la détection des expressions référentielles. Rapproché de l'inconfort observé dans le cas d'une annotation multi-passes, cet inconvénient suffit à écarter, pour le corpus DEMOCRAT, toute stratégie partielle.

3.3. Annotations réalisées

À titre d'indication sur la quantité de travail réalisé, le tableau (1) présente la liste des textes annotés lors des expérimentations décrites dans cette section. Chaque texte a été choisi pour comprendre environ 10 000 mots. Quand il s'agit d'une nouvelle, nous avons cherché une nouvelle de cette taille, avec une marge de plus ou moins 20%. Quand il s'agit d'un roman, nous avons extrait le ou les premiers chapitres. Concernant le texte de presse, il s'agit d'extraits de L'Est Républicain rassemblés dans un seul texte, *grosso modo* d'articles écrits par le même journaliste, mais sur des sujets très différents. Les neuf textes présentés dans le tableau ont tous été annotés intégralement, parfois en une seule passe, souvent après une voire deux expérimentations (peu importe la nature de l'expérimentation réalisée) puis intégralement, de manière à obtenir un corpus cohérent. Au final, tous les textes n'ont pas été

retenus, pour diverses raisons dont la représentativité de tel genre textuel et de telle période plutôt que les autres.

Texte (extrait)	Annotateur	nombre de mots de l'extrait	nombre d'expressions référentielles	Nombre de chaînes de référence
Bouvard et Pécuchet	A	10 086	4 280	284
Nemoville	A	12 992	3 252	286
De la Ville au moulin	A	10 101	4 187	292
Le Portrait de Dorian Gray	B	10 723	3 708	304
Le Ventre de Paris	A	10 037	3 147	329
Le Capitaine Fracasse	A	8 289	3 075	347
La Morte amoureuse	B	12 177	4 226	354
Sarrasine	B	12 787	4 406	399
articles de l'Est Républicain	C	10 360	2 626	393

Tableau 1 : textes annotés complètement à l'issue des expérimentations.

Toujours à titre d'indication quantifiée, il est à noter que pour les toutes dernières expérimentations d'annotation exhaustive, le rythme atteint par les annotateurs a été d'environ 500 expressions référentielles annotées par jour. Annoter un texte de 10 000 mots représente donc à peu près deux semaines de travail pour un annotateur. Les expérimentations décrites dans cette section ont nécessité plus de deux mois de travail.

4. Pré-annotation automatique : aide ou gêne ?

4.1. Quelles pré-annotations ?

Le schéma d'annotation retenu, avec son champ unique, met en avant une facette essentielle de l'annotation : la délimitation des expressions référentielles. Les annotateurs l'ont confirmé : ils passent plus de temps à délimiter qu'à catégoriser. L'ergonomie de l'outil d'annotation est alors déterminante. Rien n'est plus pénible que de repérer une expression référentielle et de se tromper d'un caractère en la délimitant. On doit dans ce cas corriger les frontières et, si ANALEC propose bien cette fonctionnalité, aucune aide n'est proposée pour ce faire, qu'il s'agisse d'un zoom automatique ou d'une proposition d'extension de la frontière du marquable à celle du mot courant.

Comme de plus l'aspect robotique de l'annotation avait été mis en avant lors des expérimentations déjà réalisées, nous avons envisagé de nouvelles expérimentations en utilisant un système de TAL avec le rôle de pré-annotateur. Il nous semblait en effet intéressant de quantifier l'apport d'une automatisation partielle du travail, de discuter de l'aspect robotique de la procédure, et de voir si partir d'un texte pré-annoté aidait un nouvel annotateur à se mettre au travail – autrement dit à observer (ou non) un équivalent du syndrome de la page blanche.

La pré-annotation a été envisagée au départ uniquement pour les problèmes de frontières puis, le débat évoluant, pour tous les aspects de l'annotation. Mais encore fallait-il trouver un système de TAL qui soit adapté au français et dont le taux d'erreur ne soit pas une entrave à l'annotation : quand les erreurs sont nombreuses, l'annotateur a vite l'impression de passer son temps à les corriger plutôt qu'à exploiter directement les pré-annotations. Les performances constituent donc un critère de choix important.

4.2. Repérages automatiques de haut niveau

Quitte à exploiter un système de TAL qui aide l'annotateur, pourquoi ne pas utiliser un analyseur automatique correspondant à la tâche visée ? Le mieux serait ainsi d'exploiter un système de détection automatique des expressions référentielles, ce qui est généralement la première étape de la détection des chaînes de référence, voire directement un système de détection des chaînes de référence dans du texte tout venant.

Comme il est impossible d'utiliser un système de TAL conçu pour une autre langue que le français, il ne reste que peu de choix. Depuis la mise en œuvre des expérimentations ici décrites, de nouveaux systèmes sont apparus, notamment (GODBERT ET FAVRE 2017). En 2016, très peu de systèmes existaient – RefGen notamment (LONGO AND TODIRASCU 2010) – et se caractérisaient soit par des performances moyennes, soit par leur non accessibilité. Un test a été fait avec des sorties de RefGen, et a montré que le nombre d'erreurs était suffisamment important pour que les annotateurs aient très vite l'impression de passer leur temps à corriger ces pré-annotations, plutôt qu'à les exploiter en tant que base de travail.

Nous avons donc envisagé d'utiliser un système de détection automatique des entités nommées (NOUVEL *ET AL.* 2015). Sauf que les expressions référentielles ainsi identifiées ne représentent qu'un sous-ensemble très restreint des expressions à annoter. Le résultat est donc un « silence » important, préjudiciable pour l'annotateur : soit celui-ci ne voit plus que les expressions détectées par l'outil, et dans ce cas beaucoup de pronoms sont oubliés, soit le travail revient à tout relire et à se comporter, en fin de compte, comme avec du texte brut sans pré-annotation. Avec un résolveur d'anaphores pronominales, le problème s'avère similaire. Une solution possible – que nous n'avons pas testée pour des raisons de complexité technique – aurait été de combiner les résultats d'un détecteur d'entités nommées avec celui d'un résolveur d'anaphores pronominales. La plus-value aurait peut-être été plus probante, mais il aurait toujours manqué les possessifs ainsi que la majorité des groupes nominaux.

4.3. Repérage automatique des groupes nominaux

Face à ces obstacles, nous avons envisagé d'exploiter d'autres systèmes opérationnels pour le français, notamment des analyseurs syntaxiques et morphosyntaxiques. Plus que ceux-ci, ce sont les chunkers qui ont attiré notre attention. Après un comparatif des différents outils disponibles, notre choix a porté sur le chunker nominal de SEM (TELLIER *ET AL.* 2012 ; DUPONT ET PLANCQ 2017). Même si un analyseur pour le français reste un enjeu de recherche, et que des erreurs sont inévitablement commises, c'est un outil tout à fait utilisable, qui a

l’avantage de repérer tous les groupes nominaux, y compris les pronoms, y compris les noms propres. Il permet ainsi à l’annotateur, et c’est son principal avantage, de n’oublier aucune expression référentielle.

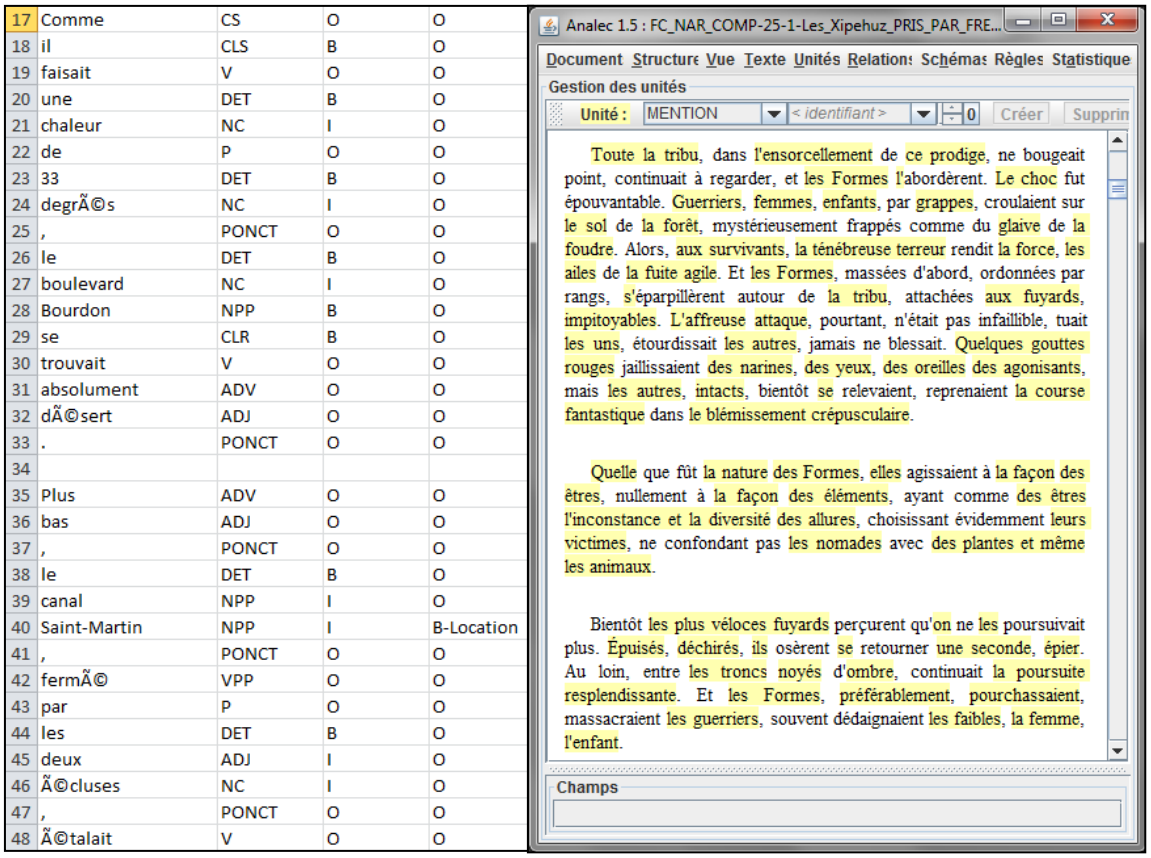
Par rapport à un détecteur d’entités nommées ou d’anaphores pronominales, le problème principal n’est cette fois plus le « silence » mais le « bruit » : l’outil pré-délimite tous les groupes nominaux, même ceux qui ne sont pas référentiels (mention non référentielle d’une partie du corps comme dans « avoir la grosse tête », par exemple), ainsi que tous les pronoms, même ceux qui sont impersonnels (« il » dans « il y a » ou « il pleut », par exemple).

Un deuxième problème, que l’on retrouve dans beaucoup d’autres outils de TAL, est qu’un chunker repère par définition des portions de texte non enchâssées. Or des expressions référentielles peuvent tout à fait s’enchâsser, comme c’est le cas avec les compléments du nom. Pour obtenir des pré-annotations vraiment exploitables, des adaptations des résultats du chunker et/ou de l’outil d’annotation sont donc nécessaires. Nous avons rapidement écarté la solution consistant à ré-enchâsser deux chunks consécutifs. En effet, pour déterminer si le deuxième concerne une expression référentielle enchâssée et non pas une juxtaposition, il est nécessaire de faire appel à un analyseur syntaxique. La mise en œuvre technique devient alors assez lourde et, surtout, la justesse des résultats est loin d’être garantie. Un chunker nominal ayant déjà un taux d’erreur non négligeable, il devient dangereux d’ajouter un post-traitement susceptible d’introduire de nouvelles erreurs.

Autrement dit, nous nous sommes reposé sur la possibilité qu’offre l’outil d’annotation de modifier les frontières d’un marquable. Le retour d’expérience des annotateurs a été assez clair : modifier les bornes d’un chunk pour qu’il englobe le chunk suivant est une opération qui, bien que simple, nécessite de la précision, et donc de la concentration. Or elle s’avère suffisamment fréquente pour diminuer l’intérêt de la pré-annotation.

La figure (1) présente un exemple de texte pré-annoté par le chunker nominal de SEM. Comme nous nous y attendions, les chunks – de même que les expressions référentielles – sont nombreuses et rendent les pré-annotations très denses. En conséquence, on hésite à les considérer immédiatement comme des aides à l’annotation.

Le bruit est présent, mais aussi le silence, c’est-à-dire des chunks non repérés par SEM alors qu’ils correspondent à des expressions référentielles. Ce cas est particulièrement problématique, comme nous l’avons vu précédemment avec les entités nommées, et sans doute plus encore ici : s’il est déjà difficile pour un annotateur de détecter les cas de silence quand ceux-ci sont nombreux, il l’est plus encore quand ceux-ci sont très peu nombreux. En effet, seuls quelques rares groupes nominaux sont oubliés, ce qui les rend difficiles à trouver. Dans l’extrait de la figure (1), aucun chunk n’a été oublié par SEM, et l’on se surprend à essayer d’en trouver un ! On passe du temps à le faire, cela demande de la réflexion et va à l’encontre de l’aspect robotique de l’annotation.



17	Comme	CS	O	O
18	il	CLS	B	O
19	faisait	V	O	O
20	une	DET	B	O
21	chaleur	NC	I	O
22	de	P	O	O
23	33	DET	B	O
24	degrÃ©s	NC	I	O
25	,	PONCT	O	O
26	le	DET	B	O
27	boulevard	NC	I	O
28	Bourdon	NPP	B	O
29	se	CLR	B	O
30	trouvait	V	O	O
31	absolument	ADV	O	O
32	dÃ©sert	ADJ	O	O
33	.	PONCT	O	O
34				
35	Plus	ADV	O	O
36	bas	ADJ	O	O
37	,	PONCT	O	O
38	le	DET	B	O
39	canal	NPP	I	O
40	Saint-Martin	NPP	I	B-Location
41	,	PONCT	O	O
42	fermÃ©	VPP	O	O
43	par	P	O	O
44	les	DET	B	O
45	deux	ADJ	I	O
46	Ã©cluses	NC	I	O
47	,	PONCT	O	O
48	Ã©talait	V	O	O

Figure 1 : exemple de fichier tabulaire obtenu en sortie du chunker nominal à gauche, et exemple d'interface d'annotation avec pré-repérage des chunks nominaux à droite.

4.4. Expérimentations chronométrées

Au final, plusieurs expérimentations ont été réalisées. Le tableau (2) présente une série d'annotations réalisées dans le temps limité de 30 minutes – temps retenu car il encourage une concentration en continu. Pour préserver la santé mentale des annotateurs, nous avons pris soin de ne pas réaliser plus de deux expérimentations chronométrées par jour. On remarque que les performances des annotateurs varient d'un texte à l'autre, et pas toujours dans le même sens : la variabilité est grande, et ne permet pas de mettre clairement en avant une stratégie plutôt qu'une autre, du moins sur le seul critère de la rapidité.

Texte (extrait)	Stratégie d'annotation	Annotateur A	Annotateur B
Le Collier des jours	systématique sans chunks	95 ER, 386 mots	93 ER, 541 mots
Boule de Suif	systématique sans chunks	100 ER, 392 mots	78 ER, 348 mots
L'Enfant	systématique sans chunks	114 ER, 338 mots	111 ER, 345 mots
La Recherche de l'absolu	systématique sans chunks	85 ER, 275 mots	80 ER, 301 mots
Manon Lescaut	objets, animaux et humains	120 ER, 397 mots	100 ER, 390 mots
Douce lumière	systématique avec chunks	105 ER, 543 mots	81 ER, 311 mots
Le Capitaine Fracasse	systématique sans chunks	130 ER, 410 mots	141 ER, 510 mots

Tableau 2 : liste chronologique des expérimentations chronométrées réalisées.

Pour choisir une stratégie, il est nécessaire d'interroger de manière approfondie les annotateurs. C'est surtout leur ressenti qui permet d'affirmer qu'annoter avec une pré-annotation est parfois un poids plutôt qu'une aide. En effet, l'annotateur doit non seulement repérer le silence, corriger les erreurs de frontières – ce qui arrive très fréquemment, on le voit dans la figure (1) avec d'une part les enchâssements qui sont « aplatis », d'autre part avec les coordinations et juxtapositions qui conduisent systématiquement à plusieurs chunks, et jamais à la délimitation de l'expression référentielle complète –, mais doit de plus traiter les chunks non référentiels. Au départ, notre première expérimentation consistait à supprimer ceux-ci. Or c'est une opération supplémentaire, coûteuse en temps, et pour laquelle un traitement rapide devrait être proposé. Dans un second temps, c'est donc un script qui s'est occupé de supprimer tous les chunks pour lesquels aucun identifiant de référent n'avait été saisi. Sauf qu'il arrive qu'un annotateur oublie de saisir cet identifiant. En temps normal, comme l'expression référentielle est délimitée, une simple revue des valeurs permet de se rendre compte de l'oubli et de le corriger. En utilisant le script de suppression, la délimitation est perdue et aucune correction n'est plus possible. L'annotateur doit donc redoubler d'attention à chaque fois qu'il traite une expression référentielle…

Comme le ressenti des annotateurs variait, nous avons décidé de laisser le choix à chaque annotateur DEMOCRAT d'exploiter ou non une pré-annotation en chunks. Tous les textes ont donc été fournis en deux versions, une avec et une sans chunks. Comme beaucoup d'aspects de la tâche d'annotation, l'appropriation de la procédure se fait généralement mieux quand une certaine souplesse est autorisée, sous condition bien entendu que les annotations finales ne soient pas impactées. Cette fois, c'est justement le risque d'hétérogénéité des annotations qui a conduit à revenir en arrière, et à imposer à tous les annotateurs de travailler sans chunks. Ce choix était d'autant plus pertinent que le corpus du projet comporte des textes en ancien français et en moyen français, pour lesquels aucun chunker similaire à SEM n'était disponible. Les expérimentations de cette section se sont donc soldées par un échec, mais elles ont eu l'avantage énorme de mettre noir sur blanc les habitudes, les préférences et les ressentis des annotateurs, contribuant ainsi à mieux connaître et caractériser la tâche d'annotation manuelle.

5. Exploitation des annotations

Les objectifs d'exploitations statistiques et TAL de notre corpus incitent à privilégier certaines stratégies d'annotation par rapport à d'autres. La rapidité d'annotation prime, car l'apprentissage artificiel nécessite des corpus de grande taille. La résolution des anaphores et des coréférences a fait l'objet de nombreux travaux depuis des dizaines d'années : les approches à base de règles ont montré leurs avantages, surtout quand les règles sont faciles à implémenter (MITKOV 2002), mais ont aussi montré leurs inconvénients face à l'apprentissage (URYUPINA 2007 ; RECASENS 2010 ; LASSALLE 2015), qui est désormais considéré comme l'approche la plus prometteuse, bien qu'elle n'ait pas encore vraiment fait ses preuves pour la

détection automatique des coréférences en langue française. Nous espérons que le corpus DEMOCRAT, de même qu'ANCOR, contribuera aux avancées à venir du TAL.

Comme la rapidité prime, nous avons mis en avant le caractère robotique de l'annotation et fait donc en sorte que l'annotateur ait à se poser le moins de questions possibles. Même en tenant compte des avantages des autres stratégies, cette démarche privilégie la stratégie consistant à annoter systématiquement tous les référents, sans aucune pré-annotation automatique.

Mieux : les expérimentations décrites dans cet article nous ont permis de clarifier la procédure complète d'annotation du corpus DEMOCRAT. Nous voulions initialement que les annotateurs saisissent non seulement l'identifiant du référent, mais aussi la catégorie morphosyntaxique de l'expression référentielle : nom propre, groupe nominal, pronom, sans oublier sa détermination (définie, démonstrative, indéfinie), voire son genre et son nombre. Or, une fois délimitées, les expressions référentielles peuvent être analysées automatiquement pour faire ressortir ces propriétés. Des scripts ont été développés dans ce but, et la procédure d'annotation définitive alterne des phases manuelles avec des lancements de scripts. Cette manière de procéder permet de consacrer le temps des annotateurs à un travail qui ne peut être fait qu'à la main, et à se reposer sur des outils automatiques pour le reste, en terminant par la construction des chaînes de référence.

6. Conclusion et perspectives

Nous avons présenté plusieurs expérimentations visant à définir la procédure d'annotation manuelle d'un corpus de grande taille. Ces expérimentations, parfois chronométrées, ont couvert plusieurs paramétrages possibles. Les annotations obtenues, de même que le calcul de l'accord inter-annotateurs et – surtout ! – les impressions et difficultés rencontrées par les annotateurs, ont pesé dans la balance lorsqu'on a dû procéder à un choix. Malgré l'exploration de nombreuses voies techniques parfois complexes, ce choix a porté sur la procédure la plus simple. Le manuel d'annotation a été rédigé dans la foulée, ce qui a permis d'y mentionner non seulement les exemples prototypiques et particuliers les plus intéressants, mais aussi les arguments ayant conduit à tel ou tel choix d'annotation – notamment pour la délimitation des expressions possédant de multiples modifieurs.

Les discussions avec les annotateurs ont permis de caractériser le travail d'annotation : une ou plusieurs passes ; passe unilatérale ou avec d'éventuels retours en arrière ; aléas et dérives lors de l'exploitation de pré-annotations imparfaites. Tous ces aspects nous semblent susceptibles d'être creusés, avec par exemple des expérimentations psycholinguistiques. On pourrait ainsi envisager l'utilisation d'un oculomètre pour observer le parcours de lecture d'un annotateur : va-t-il de mot en mot, ou de marquable pré-repéré en marquable pré-repéré ? Passe-t-il moins, ou plus de temps sur les expressions référentielles quand celles-ci sont pré-repérées ?

D’autres perspectives pouvant prolonger ce travail relèvent de la gestion des retours des annotateurs : des questionnaires pourraient par exemple être rédigés, de manière à anticiper les points d’intérêt et les différentes catégories de difficultés rencontrées, puis à mettre en perspective de manière plus méthodique les réponses fournies.

7. Remerciements

Ce travail a été réalisé avec le soutien de l’ANR dans le cadre du projet DEMOCRAT (ANR-15-CE38-0008). Un grand merci aux annotateurs impliqués dans les expérimentations, notamment à Meryl Bothua et Juliette Potier (LANDRAGIN *ET AL*. 2017).

Références bibliographiques

Artstein Ron and Poesio Massimo, 2008, « Inter-Coder Agreement for Computational Linguistics », *Computational Linguistics, 34,* p. 555-596.

Carletta Jean, 1996, « Assessing Agreement on Classification Tasks: the Kappa Statistic », *Computational Linguistics, 22*, p. 249-254.

Chamberlain Jon, Poesio Massimo and Kruschwitz Udo, 2016, « Phrase Detectives Corpus 1.0 Crowdsourced Anaphoric Coreference », in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia, p. 2039-2046.

Charolles Michel, 2002, *La référence et les expressions référentielles en français*. Paris, Ophrys, 260 p.

Désoyer Adèle, Landragin Frédéric, Tellier Isabelle, Lefeuvre Anaïs et Antoine Jean-Yves, 2014, « Les coréférences à l’oral : une expérience d’apprentissage automatique sur le corpus Ancor », *Traitement Automatique des Langues, 55(2)*, p. 97-121.

Doddington George, Mitchell Alexis, Przybocki Mark, Ramshaw Lance, Strassel Stephanie and Weischedel Ralph, 2004, « The Automatic Content Extraction (ACE) Program – Tasks, Data, and Evaluation », in *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, Lisbon, Portugal, p. 837-840.

Dupont Yoann et Plancq Clément, 2017, « Un étiqueteur en ligne du français », dans *Actes de la 24e Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2017), session des démonstrations*, Orléans, p. 15-16.

Fort Karën, 2012, Les ressources annotées, un enjeu pour l’analyse de contenu : vers une méthodologie de l’annotation manuelle de corpus, thèse de doctorat, Université Paris-Nord – Paris 13, 254 p.

Ghaddar Abbas et Langlais Philippe, 2016, « WikiCoref: An English Coreference-annotated Corpus of Wikipedia Articles », in *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia, p. 136-142.

Ghaddar Abbas et Langlais Philippe, 2017, « WiNER: A Wikipedia Annotated Corpus for Named Entity Recognition », in *Proceedings of the The 8th International Joint Conference on Natural Language Processing*, Taipei, Taiwan, p. 413-422.

Godbert Elisabeth et Favre Benoît, 2017, « Détection de coréférences de bout en bout en français », dans *Actes de la 24e Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2017)*, Orléans.

Heiden Serge, Magué Jean-Philippe et Pincemin Bénédicte, 2010, « TXM : Une plateforme logicielle open-source pour la textométrie – conception et développement », dans *Proceedings of Tenth International Conference on the Statistical Analysis of Textual Data*, Vol. 2, p. 1021-1032.

Krippendorff Klaus, 2012, *Content Analysis: An Introduction to its Methodology (third edition)*, Thousand Oaks, Sage Publishing, 456 p.

Landragin Frédéric, 2011, « Une procédure d'analyse et d'annotation des chaînes de coréférence dans des textes écrits », *Corpus, 10*, p. 61-80.

Landragin Frédéric, 2016, « Description, modélisation et détection automatique des chaînes de référence (Democrat) », *Bulletin de l'Association Française pour l'Intelligence Artificielle (AFIA), 92*, p. 11-15.

Landragin Frédéric, 2018, « Étude de la référence et de la coréférence : rôles des petits corpus et observations à partir du corpus MC4 », *Corpus, 18*, p. 1-20.

Landragin Frédéric, Poibeau Thierry et Victorri Bernard, 2012, « Analec: a New Tool for the Dynamic Annotation of Textual Data », In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, Istanbul, Turquey, p. 357-362.

Landragin Frédéric, Potier Juliette et Bothua Meryl, 2017, « Annotation manuelle d'expressions référentielles : expérimentations pour simplifier les prises de décisions et optimiser le processus », in *9èmes Journées Internationales de la Linguistique de corpus (JLC 2017)*, Grenoble, p. 43-46.

Landragin Frédéric et Schnedecker Catherine, 2014 (dir.), *Les chaînes de référence. Langages, 195*, Paris, Larousse, 142 p.

Lassalle Emmanuel, 2015, Structured Learning with Latent Trees: A Joint Approach to Coreference Resolution, thèse de doctorat, Université Paris Diderot – Paris 7, 157 p.

Longo Laurence et Todiraşcu Amalia, 2010, « RefGen: A Tool for Reference Chains Identification », in *Proceedings of the International Multiconference on Computer Science and Information Technology*, p. 447-454.

Mathet Yann, 2017, « The Agreement Measure Gamma-Cat, a Complement to Gamma Focused on Categorization of a Continuum », *Computational Linguistics, 43(3)*, p. 661-681.

Mathet Yann et Widlöcher Antoine, 2016, « Évaluation des annotations : ses principes et ses pièges », *Traitement Automatique des Langues, 57(2)*, p. 73-98.

Mathet Yann, Widlöcher Antoine et Métivier Jean-Philippe, 2015, « The Unified and Holistic Method Gamma (γ) for Inter-Annotator Agreement Measure and Alignment », *Computational Linguistics, 41(3)*, 437-479.

Mitkov Ruslan, 2002, *Anaphora Resolution*, Upper Saddle River, NJ, Pearson Education, 234 p.

Müller Christoph and Strube Michael, 2006, « Multi-level Annotation of Linguistic Data with Mmax2 », in S. Braun, K. Kohn and J. Mukherjee (eds.), *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*, Frankfurt, Peter Lang, p. 197-214.

Muzerelle Judith, Lefeuvre Anaïs, Schang Emmanuel, Antoine Jean-Yves, Pelletier Aurore, Maurel Denis, Eshkol Iris et Villaneau Jeanne, 2014, « Ancor_centre, a Large Free Spoken French Coreference Corpus: Description of the Resource and Reliability Measures », in *Proceedings of*

the Ninth International Conference on Language Resources and Evaluation (LREC 2014), Reykjavik, Iceland, p. 843-847.

Nouvel Damien, Ehrmann Maud et Rosset Sophie, 2015, *Les entités nommées pour le traitement automatique des langues.* Londres, Éditions ISTE, 168 p.

Oberlé Bruno, 2018, « SACR: A Drag-and-Drop Based Tool for Coreference Annotation », in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyasaki, Japan, p. 389-394.

Ogrodniczuk Maciej, Głowińska Katarzina, Kopeć Mateusz, Savary Agata and Zawisławska Magdalena, 2015, *Coreference in Polish: Annotation, Resolution and Evaluation*, Berlin, Walter De Gruyter, 297 p.

Pradhan Sameer, Ramshaw Lance, Marcus Mitchell, Palmer Martha, Weischedel Ralph and Xue Nianwen, 2011, « CoNLL-2011 Shared Task: Modeling Unrestricted Coreference in OntoNotes », in *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, Portland, Oregon, USA, p. 1-27.

Recasens Marta, 2010, Coreference: Theory, Annotation, Resolution and Evaluation, Ph.D. dissertation, Universitat de Barcelona, 239 p.

Sagot Benoît, Richard Marion et Stern Rosa, 2012, « Annotation référentielle du corpus arboré de Paris 7 en entités nommées », dans *Actes de la 19e Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2012)*, Grenoble, p. 535-542.

Schäfer Ulrich, Spurk Christian and Steffen Jörg, 2012, « A Fully Coreference-annotated Corpus of Scholarly Papers from the ACL Anthology », in *Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012)*, Mumbai, p. 1059-1070.

Schnedecker Catherine, 1997, *Nom propre et chaîne de référence*, Paris, Klincksieck, 231 p.

Schnedecker Catherine, Glikman Julie et Landragin Frédéric, 2017 (dir.), *Les chaînes de référence en corpus. Langue Française, n° 195*, Paris, Armand Colin, 134 p.

Tellier Isabelle, Duchier Denys, Eshkol Iris, Courmet Arnaud et Martinet Mathieu, 2012, « Apprentissage automatique d'un chunker pour le français », dans *Actes de la 19e Conférence sur le Traitement Automatique des Langues Naturelles, Volume 2*, Grenoble, p. 431-438.

Uryupina Olga, 2007, Knowledge Acquisition for Coreference Resolution. Ph.D. dissertation, Saarbrücken, Universität des Saarlandes, 268 p.

Widlöcher Antoine et Mathet Yann, 2012, « The Glozz Platform: A Corpus Annotation and Mining Tool », in *Proceedings of the 2012 ACM Symposium on Document Engineering*, ACM, p. 171-180.