Étude de la référence et de la coréférence : rôle des petits corpus et observations à partir du corpus MC4

Frédéric Landragin

Étude de la référence et de la coréférence : rôle des petits corpus et observations à partir du corpus MC4

Frédéric LANDRAGIN
C.N.R.S. " Lattice " (Montrouge)[1]

Résumé : Nous présentons un ensemble d'observations et de repères méthodologiques issus de la constitution et de l'annotation du corpus MC4 « Modélisation Contrastive et Computationnelle des Chaînes de Coréférences », corpus de (très) petite taille qui a permis des expérimentations, des études de faisabilité puis l'élaboration d'un corpus de beaucoup plus grande taille, celui du projet DEMOCRAT « Description et modélisation des chaînes de référence : outils pour l'annotation et le traitement automatique ». Nos remarques traitent notamment de la taille d'un corpus annoté et du besoin de documentation complète, même pour un très petit corpus incompatible avec des exploitations ADT et TAL.

Abstract : We present a set of methodological observations and benchmarks derived from the constitution and annotation of the MC4 corpus, "Contrastive and computational modeling of coreference chains". MC4 is a (very) small corpus which allowed experiments, feasibility studies and then the design of a much larger corpus: the DEMOCRAT corpus, "Description and modeling of reference chains: tools for annotation and automatic processing". Our remarks deal in particular with the size of an annotated corpus, and with the need for a complete documentation, even with a very small corpus incompatible with statistical analyses and NLP exploitations.

Mots clés : procédure d'annotation, taille de corpus, documentation de corpus.
Key words: annotation procedure, corpus size, corpus documentation.

1. Laboratoire mixte CNRS, ENS, Université de Paris 3, Université Sorbonne Paris Cité, PSL Research University, UMR 8094. **DRAFT AUTEUR**.

1. Introduction

Nous présentons un ensemble d'observations issues de la constitution et de l'annotation (manuelle) du corpus MC4, « Modélisation Contrastive et Computationnelle des Chaînes de Coréférences » (https://hdl.handle.net/11403/mc4). MC4 est un petit corpus, voire un « tout petit corpus » : les extraits de textes (écrits) qui le composent ont une longueur totale de 20.000 mots ; les marquables concernés par l'annotation manuelle sont des expressions référentielles et sont au nombre de 4.000 ; le schéma d'annotation comprend 78 étiquettes réparties selon 11 traits – comme nous le verrons dans la discussion, n'indiquer qu'un seul de ces nombres (usuellement le nombre de mots) ne peut pas renseigner suffisamment le lecteur sur la « taille » d'un corpus annoté.

Nous avons conscience du fait que ces nombres sont très peu élevés. On peut même considérer qu'un ensemble de 20.000 mots ne constitue pas un « corpus », et c'est l'un des aspects que nous discuterons, en explicitant les objectifs et les difficultés rencontrées dans le cas de MC4. Nous essayerons tout au long de cet article d'articuler nos remarques portant sur ce projet spécifique avec des déductions et des considérations s'appliquant de manière générale à la catégorie dite des petits corpus.

Nos observations seront essentiellement d'ordre méthodologique. Elles concernent les intérêts et les limites des corpus annotés pour trois exploitations courantes dans la communauté de la linguistique de corpus outillée :

1. Études linguistiques : un corpus même « petit » apporte des exemples remarquables (repérés par les marquables qui les délimitent), des interprétations (à travers les annotations linguistiques que portent ces marquables), et par conséquent des arguments qui aident à confirmer ou infirmer des hypothèses linguistiques.
2. Amélioration ou évaluation d'outils de traitement automatique des langues (TAL), qu'il s'agisse de systèmes à base de règles – pour lesquels l'exploration de corpus est source d'inspiration et permet de déterminer efficacement des règles (Godbert & Favre,

2017) – ou de systèmes à base d'apprentissage artificiel, pour lesquels les exemples annotés sont autant de données sur lesquelles le système va se fonder pour identifier ses propres règles (Tellier, 2009 ; Denis, 2007 ; Désoyer *et al.*, 2014 ; Godbert & Favre, 2017).

3. Exploitation de techniques d'analyse statistique de données textuelles (ADT) – cf. entre autres exemples (Lebart & Salem, 1994), et plus récemment (Turenne, 2016) ou (Poudat & Landragin, 2017) – avec notamment l'exploration de techniques tenant compte aussi bien des textes que des annotations linguistiques (Pincemin, 2004 ; Landragin, 2016).

La taille du corpus est une préoccupation commune à ces trois exploitations, mais avec des contraintes différentes. Les possibilités d'exploitation TAL et ADT sont en effet fortement contraintes par la quantité de données fournie par le corpus. C'est donc un aspect qui nécessite des définitions et des repères, et qui contribue même largement à caractériser un « petit corpus » par rapport à un « grand corpus ». La documentation du corpus est également une préoccupation essentielle : « petit corpus » ne veut pas dire « petite documentation ».

Concernant le cas particulier du corpus MC4, notre propos est une remise en perspective d'une part des étapes qui ont amené à la constitution de ce corpus, d'autre part de la procédure de passage à l'échelle actuellement engagée dans le projet DEMOCRAT, « DEscription et MOdélisation des Chaînes de Références : outils pour l'Annotation de corpus (en diachronie et en langues comparées) et le Traitement Automatique » (http://www.lattice.cnrs.fr/democrat/). L'objectif du projet DEMOCRAT est en effet de tenir compte de l'expérience acquise avec le corpus MC4 pour proposer un corpus de plus grande taille, en tout cas d'une taille suffisante pour permettre des applications TAL et ADT des données annotées. L'évolution de MC4 à DEMOCRAT illustre en quelque sorte les préoccupations du passage d'un « petit corpus » à un « grand corpus », et c'est pourquoi nous l'évoquerons. Et ce d'autant plus que le passage par un petit corpus a été nécessaire

à un ensemble de prises de conscience qui ont permis de lancer DEMOCRAT en toute connaissance de cause.

Nous commençons cet article par une présentation du corpus MC4 (section 2), avec notamment la liste des questions scientifiques à l'origine du schéma d'annotation mis en œuvre. En partant de cet exemple concret, nous pouvons alors mettre en avant les intérêts d'un petit corpus : étape méthodologique pour illustrer une théorie ou une approche (section 3) ; étude de faisabilité avant de passer à l'échelle (section 4). Ces deux approches méthodologiques du petit corpus ont des répercussions sur la manière de calculer la taille du corpus (section 5) et d'écrire sa documentation (section 6). L'approche concernée par un passage à l'échelle a également ses propres spécificités, que nous illustrons ici avec le passage du corpus MC4 au corpus DEMOCRAT (section 7). Nous concluons alors sur la notion de petit corpus pour les phénomènes de référence et de coréférence, et nous donnons quelques perspectives qui correspondent à des préoccupations du projet DEMOCRAT.

2. Le corpus MC4 : historique et grandes lignes

Le projet MC4 est un projet PEPS – Projets Exploratoires Premier Soutien – du CNRS, qui s'est déroulé de 2011 à 2012 et a réuni des chercheurs des laboratoires Lattice (Montrouge), LiLPa (Strasbourg) et ICAR (Lyon). Le corpus MC4, avec les publications qui lui sont attachées (Landragin & Schnedecker, 2014), en est son principal résultat.

Ce corpus se compose des textes ou extraits de textes suivants :

- *Gracial d'Adgar*, extrait (en vers) de 2.641 mots (12^{e} siècle) ;
- *Quatre Livres des Rois*, extrait (en prose) de 2.211 mots (12^{e} siècle) ;
- *La vie de Saint Thomas Becket*, extrait (en vers) de 2.067 mots (12^{e} siècle) ;
- *Li Estoires de Chiaus qui conquisent Coustantinoble*, de Robert de Clari, extrait (en prose) de 1.639 mots (12^{e} – 13^{e} siècle) ;

- *Queste del saint Graal*, extrait de 2.224 mots (13e siècle) ;
- *Les quinze Joyes de mariage, Premiere joye*, extrait de 2.457 mots (14e – 15e siècle) ;
- *Les Bijoux*, de Guy de Maupassant, texte de 2.448 mots (19e siècle) ;
- *La Mère sauvage*, de Guy de Maupassant, texte de 2.450 mots (19e siècle) ;
- *L'occupation des sols*, de Jean Echenoz, texte de 1.787 mots (20e siècle) – texte non diffusé dans l'archive MC4 car non libre de droits. Pour ce texte qui a fait l'objet d'études spécifiques, cf. (Charolles & Le Goffic, 2015) et notamment (Landragin *et al.*, 2015).

Le projet MC4 s'est focalisé sur le choix et l'annotation de textes hétérogènes, en vue de couvrir le maximum de phénomènes référentiels (pour des référents humains uniquement), et ce à travers l'étude de plusieurs états de la langue française. Ce petit corpus réunit ainsi des textes en vers ou en prose, en français médiéval ou en français contemporain. En plus d'une certaine « représentativité » de la référence aux personnages humains, l'objectif était aussi de permettre des comparaisons qualitatives une fois tous les textes annotés. Pour cela, un même schéma d'annotation a été utilisé : compte tenu de la petite taille du corpus, c'était une contrainte absolue.

L'élaboration de ce schéma d'annotation a constitué l'une des principales tâches du projet MC4. Elle s'est appuyée sur un ensemble de discussions qui avaient eu lieu auparavant (2009-2011) dans le cadre d'un groupe de travail sur la coréférence, et dont une partie des résultats est reportée dans (Landragin, 2011). Dans la mesure où la référence, même quand on ne considère que la référence aux personnages humains, est un phénomène linguistique et pragmatique varié et complexe, les discussions ont été longues et nombreuses. Nous reproduisons ci-dessous à titre indicatif la liste des thèmes abordés lors de chacune des réunions de ce groupe de travail – thèmes repris ensuite en partie dans le projet MC4 (sans pour autant reprendre les discussions) :

1. Définition d'une « chaîne de coréférences » : question pas si anodine, dans la mesure où la littérature (Chastain, 1975 ; Corblin, 1995 ; Schnedecker, 1997 ; Cornish, 1999 ; Schnedecker, 2005) évoque aussi les notions de « chaîne de référence », « chaîne anaphorique », « chaîne topicale », et ainsi de suite. Nous retiendrons ici qu'une chaîne de coréférences regroupe toutes les mentions d'un même référent.
2. Définition d'une « expression référentielle » : comment passer de l'objet linguistique au marquable concerné par l'approche de la linguistique de corpus outillée. Les grands types d'expressions référentielles sont les noms propres, les pronoms et les groupes nominaux. Tous les noms propres, tous les pronoms et tous les groupes nominaux ne réfèrent pas pour autant (Charolles, 2002). Globalement, les marquables sont donc des groupes de mots consécutifs, qui comportent le plus souvent un ou deux mots (« Jean Dupont », « il », « le président »), et vont jusqu'à une dizaine voire une vingtaine de mots. Ce sont les « maillons » des chaînes de coréférences.
3. Réflexion sur les référents évolutifs – le poulet vif et bien gras que l'on abat, que l'on coupe en morceaux et que l'on sert à dîner – et leur prise en compte dans des chaînes de coréférences (nécessité de relâcher les contraintes inhérentes à une relation de coréférence stricte).
4. Définition d'une typologie des référents envisageables, notamment des groupes (d'individus, d'objets), en tenant compte de groupes stricts et de groupes flous.
5. Détermination des relations (partie-tout, association) pouvant exister entre deux référents, et pouvant donc faire l'objet d'une annotation reliant deux chaînes de coréférences.
6. Définition d'une typologie des transitions référentielles – « maintien » d'un référent saillant ; « changement » de référent ; « compétition » entre plusieurs référents (cf. aussi 13^{e} item de cette liste, qui reprendra cet aspect lors de la mise en œuvre d'un schéma d'annotation) – et

réflexion sur les possibilités d'annoter ce type de phénomène, ou au contraire d'observer les transitions en étudiant les annotations des expressions et leurs alternances.

7. Réflexion sur les mécanismes d'extractions et de regroupements de référents.
8. Détermination des caractéristiques des expressions référentielles, qui pourraient chacune faire l'objet d'une annotation (ou d'un trait dans une structure de traits).
9. Réflexion sur la prise en compte des relations anaphoriques dans un schéma d'annotation dédié aux chaînes de coréférences, et notamment des anaphores associatives (Kleiber, 2001) ainsi que de certains phénomènes de saillance référentielle (Landragin, 2004).
10. Distinction entre les expressions référentielles telles que les groupes nominaux, et des indices référentiels, par exemple les marques d'accord en genre et en nombre (d'un verbe), qui en rappellent le référent actant et participent à ce titre aux chaînes de coréférences.
11. Détermination des critères syntaxiques (coordination, schéma actanciel, reprise par un pluriel) qui conduisent à considérer un référent de type « groupe » en plus de deux ou plusieurs référents « individuels ».
12. Distinction entre la résolution immédiate d'une référence, c'est-à-dire au cours de la lecture, sans retour en arrière dans le texte pour en réinterpréter une partie, et la résolution différée, pour laquelle une expression référentielle peut être réinterprétée – plus loin dans le texte, mais aussi lors d'une deuxième ou troisième lecture, une fois que le lecteur a acquis des connaissances qui lui manquaient à la première lecture.
13. Réflexion sur une adaptation de la Théorie du Centrage (Grosz *et al.*, 1995) – un peu à la suite de la réunion correspondant au 6^{e} item de cette liste – pour en faire un schéma d'annotation en complément de celui dédié aux chaînes de coréférences.

14. Définition d'une structure de traits qui constituera la base de l'annotation des expressions référentielles, de manière à disposer d'un corpus annoté selon plusieurs dimensions d'analyse de la langue. Parmi ces traits se trouvent ainsi la détermination de l'expression référentielle, sa fonction syntaxique, son rôle actanciel, sa position dans la phrase, etc. L'intérêt est de regrouper les aspects morphologiques, syntaxiques et sémantiques liés à la référence, de manière à faciliter l'exploration des annotations.
15. Détermination d'une typologie des expansions (modifieurs, compléments du nom, relatives).
16. Réflexion sur la prise en compte (par exemple dans une nouvelle couche d'annotation, ou automatiquement) des différents niveaux d'enchâssement syntaxique.
17. Réflexion sur la manière de délimiter et d'annoter les expressions référentielles quand elles s'accompagnent d'une expression prédicative ou attributive.
18. Réflexion sur la manière de délimiter et d'annoter les pronoms relatifs, les sujets zéros, ou encore les pronoms réfléchis : pour chacun de ces phénomènes, plusieurs cas ont été envisagés, menant à plusieurs stratégies d'annotations. L'annotateur doit donc décider à chaque occurrence de la stratégie à suivre en s'aidant d'un test linguistique (par exemple un test permettant de distinguer les relatives déterminatives des relatives explicatives).
19. Discussion sur le repérage (voire l'annotation avec une étiquette spécifique) des cataphores.
20. Réflexion sur l'exploitation (en tant que pré-annotation) des étiquettes obtenues avec un outil de TAL tel qu'un analyseur morphosyntaxique ou syntaxique.
21. Réflexion sur le besoin voire la nécessité de segmenter les chaînes de coréférences en sous-chaînes, que ce soit sur la base d'arguments liés à la structure textuelle (découpage en paragraphes) ou sémantico-discursifs (cadratifs, par exemple).

22. Réflexion sur l'intérêt de tenir compte non seulement des référents humains, mais aussi des référents concrets (objets), des référents abstraits, et notamment des référents temporels et événementiels. Un lien est fait avec le genre textuel, dans la mesure où certains types de référents sont plus fréquents dans certains genres textuels – typiquement, les référents humains dans les textes narratifs (mais pas seulement, bien entendu).
23. Réflexion sur la nécessité de prendre en compte d'une part le plan énonciatif, d'autre part la structure textuelle, soit dans la structure de traits affectée aux expressions référentielles, soit dans une couche d'annotation supplémentaire. L'intérêt, pour une analyse statistique par exemple, est de prendre en compte d'une manière différente les maillons qui apparaissent dans le corps du texte de ceux présents dans un titre, un sous-titre ou une note de bas de page.
24. Réflexion sur la référence vague (« ils ont encore augmenté les impôts »), sur la sous-détermination (par rapport à l'ambiguïté), et sur la référence floue, par exemple une anaphore avec deux antécédents possibles sans qu'il soit nécessaire (pour la compréhension) de décider, ou encore sur les référents évolutifs discutés d'un point de vue linguistique dans le 3^{e} item de cette liste : comment annoter ce type de cas limites ?
25. Réflexion sur d'autres cas d'expressions référentielles rencontrés au fur et à mesure des réflexions et des études préliminaires de textes, par exemple les expressions attributives, ou encore les références de type « l'un… l'autre ».

Le but de cette liste est multiple. Elle montre d'une part que constituer un schéma d'annotation pour des phénomènes référentiels n'a rien à voir avec la constitution d'un jeu d'étiquettes morphosyntaxiques, ou celle d'une étude portant sur un marqueur linguistique donné (par exemple un mot auquel on peut attribuer plusieurs interprétations sémantiques – la tâche de l'annotateur revenant à choisir pour chaque occurrence la bonne interprétation). Il a ainsi fallu pas moins de 25

réunions pour aboutir à un schéma d'annotation complet et (presque) consensuel. Cette liste montre d'autre part que le travail de l'annotateur peut s'avérer délicat, pour ne pas dire réservé à des linguistes spécialistes de référence et de coréférence. À titre d'exemple, la délimitation d'une expression référentielle à laquelle est liée une subordonnée relative ne conduit pas à la même procédure selon que la relative est déterminative ou explicative (aspect discuté lors de la réunion correspondant au 18e item de notre liste). Effectuer un test s'avère donc nécessaire à chaque occurrence rencontrée, et c'est un aspect sur lequel un annotateur peut potentiellement se tromper. La même remarque est valable pour les pronoms réfléchis : selon que la valeur est réfléchie, réciproque, passive ou essentiellement pronominale, la procédure d'annotation varie : elle fait donc appel à des connaissances linguistiques, et parfois à des connaissances très spécialisées.

On comprendra ainsi que l'élaboration d'un schéma d'annotation et de la procédure d'annotation associée ait demandé autant de temps. On comprendra également qu'annoter « seulement » 4.000 expressions référentielles n'est pas un « petit » travail, même si le résultat obtenu reste effectivement un « petit » corpus.

Par ailleurs, le groupe de travail sur la coréférence et le projet MC4 ont également nécessité un ensemble de discussions portant sur les aspects méthodologiques et techniques de l'annotation (Habert *et al.*, 1997 ; Habert, 2005 ; Fort, 2012 ; Landragin *et al.*, 2015 ; Poudat & Landragin, 2017) : si l'annotation de marquables avec un ensemble d'étiquettes prédéterminées peut se faire à l'aide d'une multitude d'outils, il n'en est pas de même de l'annotation d'anaphores et de chaînes de coréférences (Van Deemter & Kibble, 2000). À titre indicatif, ces discussions – qui ont fait l'objet de multiples réunions – ont porté sur les aspects classiques de l'annotation manuelle suivants :

26. Choix des textes et des formats de textes ; préparation des fichiers informatiques avec un codage bien choisi (à titre d'exemple, l'un des premiers textes que nous avions choisis comportait des citations en grec ancien et

nous a fait revoir le choix d'encodage pour considérer un format Unicode).

27. Comparaison des outils d'annotation disponibles sur le marché, compte tenu de leur ergonomie et de leurs fonctionnalités quant aux schémas d'annotation possibles : structures de traits typés ; traits éventuellement multivalués ; possibilité d'annoter des relations entre marquables (et pas seulement les marquables eux-mêmes) ; possibilité de construire des ensembles de marquables pour rendre compte des chaînes de coréférences, etc. Ont ainsi été testés et comparés les outils MMAX2 (Müller & Strube, 2006) ; GLOZZ (Widlöcher & Mathet, 2009) et ANALEC (Landragin *et al.*, 2012). Notons au passage que la grande majorité des corpus annotés manuellement en chaînes de coréférences (Schäfer *et al.*, 2012 ; Ogrodniczuk *et al.*, 2015 ; Ghaddar & Langlais, 2016) l'ont été avec MMAX2, et que le seul corpus de grande taille disponible pour la langue française, ANCOR (Muzerelle *et al.*, 2014), l'a été avec GLOZZ. Le choix retenu pour MC4 – ANALEC – est proche de ce dernier dans la mesure où ANALEC partage le même schéma conceptuel que GLOZZ. *Grosso modo*, seule l'ergonomie d'annotation et d'exploration diffère.
28. Détermination d'un compromis entre les possibilités offertes par les outils d'annotation et la complexité des phénomènes référentiels dont on veut rendre compte. À titre d'exemple, l'annotation d'un sujet zéro, une fois qu'elle a été décidée (18^{e} item), pose des problèmes techniques : faut-il faire ressortir par un caractère spécial les sujets zéro dans le texte – auquel cas ce caractère spécial sert de marquable ? Si l'on garde au contraire le texte inchangé, l'outil autorise-t-il l'annotation de l'espace devant le verbe ? Ou faut-il se reporter sur la forme verbale elle-même ?
29. Rédaction – éventuellement incrémentale – du manuel d'annotation, en trouvant un compromis entre concision et exhaustivité, dans la mesure où l'annotateur ne doit

pas être rebuté par un document trop long et difficile à appréhender.

30. Mise en œuvre de l'annotation en faisant appel à plusieurs annotateurs : pré-expérimentations rapides ; évaluation des résultats de ces pré-expérimentations et détermination des modalités du calcul de l'accord inter-annotateurs ; choix d'une procédure définitive ; choix de mettre en place une phase d'« adjuvication » (ou « arbitrage ») pour traiter les cas de désaccord entre annotateurs et finaliser le corpus.
31. Rédaction de la documentation du corpus, en tenant compte des métadonnées des textes sélectionnés et de l'intégralité des facettes de la procédure d'annotation.

Sans plus entrer dans les détails, nous noterons au final que l'élaboration du corpus MC4 a commencé quasiment dès 2009 et a nécessité un temps de travail très important, que la petite taille du corpus ne reflète pas. Plusieurs versions du schéma d'annotation se sont succédé, et le manuel d'annotation a subi de nombreuses mises à jour. Parmi les points peu discutés durant le projet se trouve l'exploitation d'annotations préalables, par exemple en morphosyntaxe et en syntaxe, et ce en choisissant des textes déjà annotés dans le cadre d'autres projets de recherche. En effet, comme MC4 prévoyait dès le départ la constitution d'un « petit corpus », il a été décidé rapidement de tout annoter à la main, et donc de ne pas se lancer dans l'exploitation d'une couche de pré-annotations – la décision inverse sera prise dans le projet DEMOCRAT car on sortira alors du cadre des « petits corpus ». L'annotation de MC4 s'est terminée en 2013, la finalisation (métadonnées et encodage du corpus) en 2014, et la diffusion – sous licence CC BY-NC-SA 3.0 FR – est assurée depuis 2015 par la plateforme Ortolang. Ces deux dernières phases – finalisation et diffusion – ont bénéficié d'un soutien de la part du consortium IR Corpus Écrits (maintenant CORLI, « Corpus, Langues, Interactions », de la Très Grande Infrastructure de Recherche sur les Humanités Numériques). Ce soutien a permis de mettre en œuvre un format XML TEI de représentation de corpus incorporant des annotations complexes comme le sont les

chaînes de coréférences (Mélanie-Becquet & Landragin, 2014), et de développer des modules d'importation et d'exportation spécifiques dans l'outil ANALEC qui a servi tout au long du projet.

Parmi les raisons qui ont contribué à la lourdeur de la procédure d'annotation, la plus importante à nos yeux est le phénomène linguistique choisi : annoter la référence est un défi en soi. Des formes telles que « il », « une belle jambe » ou « Alan Turing » peuvent être aussi bien référentielles – « il est venu », « une belle jambe est apparue dans l'entrebaillement de la porte », « Alan Turing était un mathématicien » – que non référentielles – « il pleut », « ça me fait une belle jambe », « il a reçu le prix Alan Turing » –, avec parfois des discussions sur l'aspect plus ou moins référentiel : « faire une belle jambe » peut être considérée comme une forme plus ou moins figée ; « le prix Alan Turing » est un référent qui rappelle plus ou moins un autre référent : l'homme à l'origine de la dénomination du prix. Il est toujours possible de forcer la présence d'un référent « caché » : « il a reçu le prix Alan Turing. Celui-ci doit se retourner dans sa tombe ». Même chose pour les expressions figées : « Mme Rezeau n'osa pas dire « Ça me fait une belle jambe ! » mais, par suite d'une silencieuse association d'idées, elle se caressa longuement le tibia » (Hervé Bazin, *Vipère au poing*). Dans ce dernier exemple, autant la constitution d'une chaîne de coréférences pour le personnage de Mme Rezeau ne pose pas trop de problèmes – on identifie quatre maillons : « Mme Rezeau », « me », « elle », « se » –, autant celle regroupant « une belle jambe » et « le tibia » peut soulever des questions qui risquent de départager les annotateurs. Chaque phrase peut ainsi faire l'objet de plusieurs questions référentielles et, même en prévoyant un manuel d'annotation suffisamment complet et directif, un annotateur sera forcément confronté à des doutes, voire à des impossibilités à décider. Or l'annotation manuelle d'un grand corpus ne peut pas se permettre de tels questionnements : seule celle d'un petit corpus le peut, et c'est l'approche que nous avons privilégiée pour MC4.

On pourrait objecter qu'il n'est pas raisonnable d'annoter de la littérature. Il est vrai que des exemples comme celui d'Hervé Bazin sont complexes, et que l'on pourrait considérer comme inutile de s'y attarder. Mais les extraits que nous avons étudiés lors des 25 réunions mentionnées ci-dessus se sont avérés encore plus complexes, alors qu'il s'agissait de dépêches de l'AFP, de recettes de cuisine ou de résumés de films trouvés sur le web (Landragin, 2011). Il semble en effet que des textes courts – et parfois mal écrits – regorgent d'ambiguïtés et de flous référentiels qui complexifient la tâche d'annotation. Attribuer un référent précis à une expression référentielle n'est pas une tâche facile, et faire rentrer des phénomènes extrêmement variés dans les cases d'un schéma d'annotation ne l'est pas non plus. En fin de compte, le choix de l'auteur ou du genre textuel ne complexifie pas beaucoup plus une tâche qui l'est déjà de manière inhérente.

Le corpus MC4 a donc été annoté en tenant compte de la complexité des phénomènes référentiels, avec toutefois un filtrage des référents puisque seuls les référents humains ont fait l'objet d'annotations. Une fois les 4.000 expressions référentielles annotées, plusieurs analyses ont été effectuées (Landragin & Schnedecker, 2014). Il est à noter que la petite taille du corpus n'a pas permis d'effectuer des analyses statistiques poussées, et encore moins d'exploitation des données annotées pour des besoins de TAL. Néanmoins, la faisabilité de plusieurs études a été testée, avec notamment l'écriture de scripts pour extraire automatiquement certaines données, par exemple celles qui concernent un aspect linguistique spécifique, tel que la détermination.

Peu après la publication des résultats de ces premières analyses, le corpus a été rendu disponible (à l'exception d'un texte sous droits). Comme pour un grand corpus, la disponibilité d'un petit corpus est essentielle pour les chercheurs qui s'intéressent au phénomène linguistique étudié : elle permet à ceux-ci d'explorer les données annotées, de comparer le schéma d'annotation suivi avec le leur, de confronter leurs idées quant aux types d'annotation souhaités pour d'autres corpus, et notamment pour des corpus de plus grande taille. Plutôt qu'une

fin en soi, le petit corpus peut servir d'étape intermédiaire dans un parcours méthodologique.

3. Petit corpus : étape méthodologique pour illustrer une théorie ou une approche

L'analyse d'un petit corpus reste à taille humaine, c'est-à-dire qu'elle peut s'effectuer par un chercheur unique, par exemple dans le cadre d'une recherche ponctuelle – réalisable en quelques mois voire quelques semaines. Le petit corpus est en quelque sorte la version orientée « linguistique de corpus outillée » de l'ensemble d'exemples récoltés ou fabriqués qui servent à matérialiser des préoccupations de recherche (et à illustrer des articles de linguistique théorique et descriptive). Un petit corpus matérialise une démarche de recherche, fournit des exemples et des preuves de faisabilité. On peut même le considérer comme l'instrument adéquat pour ce faire.

D'une manière générale, la constitution d'un petit corpus peut jouer de nombreux rôles :

1. Le premier recueil de données attestées (textes seulement, ou textes enrichis d'annotations), permettant d'identifier des tendances – non valides statistiquement, mais suffisantes pour encourager ou réorienter un travail de recherche – et de mettre en place des collaborations de recherche autour d'un même objet d'étude. C'est clairement le cas de MC4.
2. L'expérimentation, visant par exemple à tester la faisabilité d'un schéma d'annotation avant le lancement d'une campagne d'annotation de plus grande envergure. Le petit corpus permet de mettre en œuvre une procédure d'annotation et de la valider, notamment *via* le calcul d'accords inter-annotateurs (qui se satisfait de petits corpus). Là aussi, MC4 a eu une utilité certaine, notamment pour le lancement de DEMOCRAT.
3. La validation d'une procédure d'annotation, qui en inclut toutes les facettes – en tout cas le maximum – sur des données en quantité réduite, de manière à tester l'ergonomie de l'annotation, l'intérêt des éventuelles pré-annotations automatiques, la lourdeur de la tâche de

> saisie et/ou de correction manuelle, la possibilité de s'aider d'outils de TAL et de scripts divers, et ainsi de suite. Sur ces aspects, l'expérience de MC4 s'est avérée plutôt négative. Elle a permis en tout cas d'envisager l'exploration de nouvelles directions de travail pour DEMOCRAT.

Bien entendu, la constitution d'un grand corpus visera directement le recueil de données attestées (textes) en quantité, et l'annotation à grande échelle. Elle se reposera sur les trois aspects précédents et ajoutera plusieurs règles quant à la constitution du corpus : règles de représentativité, règles d'échantillonnage des textes, etc. Cela ne veut pas dire pour autant que la constitution d'un petit corpus peut s'affranchir de ces règles. Pour que les procédures d'analyse puissent être validées sur un petit jeu de données, encore faut-il que ce qui fait office d'échantillon soit construit correctement. Concernant le corpus MC4, le but était notamment de comparer les expressions référentielles selon plusieurs états de la langue française (le sujet zéro est un phénomène important en français médiéval et était entre autres exemples un sujet d'étude des participants – diachroniciens ou non – du projet). Logiquement, les quelques textes constituant le corpus sont issus de périodes variées. Chaque siècle n'est cependant pas représenté : ce qui pour un grand corpus nécessiterait une rigueur et un choix systématique pour chaque siècle de textes de tailles comparables prend ici la forme d'une ébauche, jugée satisfaisante compte tenu des moyens mis en œuvre.

La notion de « représentativité » est au cœur de la légitimité des petits corpus. Mais cette notion est complexe et très difficile à valider. Comment peut-on construire un corpus (annoté) représentatif des phénomènes de référence en langue française ? La tâche est quasiment impossible : il faudrait avoir une idée très précise de l'étendue des expressions référentielles, de leurs fonctions syntaxiques, de leurs rôles thématiques, de leurs rôles discursifs – notamment leurs apports dans les chaînes de coréférences – et surtout des mécanismes mis en jeu lors de leur interprétation. Autant il est possible de comptabiliser les aspects purement formels (encore que…),

autant il est impossible de décrire l'intégralité des mécanismes d'interprétation référentielle pour en dresser une liste de cas exhaustive. Ce type d'approche est tout à fait réalisable pour des annotations morphosyntaxiques voire syntaxiques, mais très difficilement dès que l'on aborde des notions comme la référence indirecte, attributive, plus ou moins générique, etc., cf. par exemple les multiples interprétations possibles du pronom « on » (Fløttum *et al*., 2007 ; Cabredo Hofherr, 2008 ; Landragin & Schnedecker, 2014).

À défaut de pouvoir identifier des critères précis de représentativité, nous avons considéré qu'un corpus comprenant un certain nombre d'expressions référentielles annotées constituait un échantillon digne d'intérêt. Pour MC4, ce nombre est de 4.000. C'est le résultat d'un temps de travail plutôt que d'un calcul *a priori*, et le nombre en lui-même n'a aucune valeur : comment fixer un seuil quand la diversité des phénomènes est telle qu'on ne peut même pas en donner un ordre de grandeur ? Le corpus ANCOR (Muzerelle *et al*., 2014) comporte 115.000 mentions annotées, mais peut-on seulement valider sa représentativité ? On ne le peut qu'à partir du jeu d'étiquettes utilisé lors de l'annotation, ou en testant des systèmes d'apprentissage artificiel entraînés sur la totalité ainsi que sur des parties plus ou moins couvrantes du corpus. Mais, même ainsi, comment garantir que le corpus est linguistiquement représentatif ?

Nous considérons que les phénomènes référentiels sortent du cadre méthodologique nécessaire à de telles considérations. Par conséquent, nous considérons qu'un petit corpus annoté en référence ne peut pas prétendre à une représentativité quelconque de ce phénomène (en suivant une interprétation stricte du terme « représentativité »), et donc ne peut *a priori* pas être exploité pour des mesures statistiques significatives. Le corpus MC4 n'a pas de telles prétentions, de même qu'il ne peut pas servir à entraîner des systèmes d'apprentissage artificiel ou à élaborer des analyses telles que celles faites dans le domaine de l'ADT. De son côté, le corpus ANCOR permet ce type d'exploitation (Désoyer *et al*., 2014 ; Godbert & Favre, 2017).

En revanche, nous considérons qu'un petit corpus peut prétendre à un certain nombre de tâches préalables :

1. L'exploration voire l'approfondissement d'une théorie linguistique de la référence, avec par exemple l'analyse approfondie d'exemples pour identifier des cas intéressants voire des tendances.
2. L'illustration d'une démarche scientifique précise, depuis la délimitation des phénomènes linguistiques jusqu'à la détermination des méthodes, calculs et procédures d'analyse.
3. La réalisation de systèmes de TAL ou de prototypes de systèmes (voire de *baseline*) tels que des systèmes de résolution des anaphores ou de détection des chaînes de coréférences sur la base de règles. Le concepteur d'un tel système peut en effet parcourir l'intégralité des exemples regroupés dans un petit corpus, et s'en inspirer pour déterminer les règles de son système. Tous les cas ne seront pas pris en compte, mais le système pourra malgré tout s'avérer performant sur les cas les plus typiques ou les plus fréquents. Ce qui n'est déjà pas si mal, même si les techniques d'apprentissage artificiel s'avèrent bien plus performantes et bien plus prometteuses au fur et à mesure que des grands corpus se constituent.
4. La validation d'une démarche d'annotation, impliquant par exemple plusieurs annotateurs, ce qui passe par des calculs d'accords inter-annotateurs et, après interprétation de ceux-ci (ainsi éventuellement qu'une phase d'ajustement du schéma d'annotation), ce qui permet d'envisager un passage à l'échelle, c'est-à-dire l'application de la même procédure pour l'obtention maîtrisée – en qualité comme en temps de travail – d'un corpus de grande taille.

Concernant MC4, seuls les deux premiers points ont vraiment fait l'objet d'efforts : la réalisation d'un système de TAL n'était pas (encore) à l'ordre du jour, et la validation de la procédure d'annotation n'a pas été faite dans les règles de l'art (Fort, 2012), la raison étant un nombre trop élevé d'étiquettes – qui

plus est sur des phénomènes très variés – ce qui a conduit à un taux d'accord inter-annotateurs (estimé sur un extrait du corpus) trop faible. Dans ce sens, il a été décidé de ne pas conserver le schéma d'annotation de MC4 pour la constitution d'un corpus de plus grande taille. Autrement dit, MC4 a conduit à déterminer une nouvelle procédure d'annotation pour le passage à l'échelle. Cela ne revient pas forcément à considérer MC4 comme un échec, mais plutôt comme une initiative close, qui se doit de rester dans le domaine des petits corpus.

4. Petit corpus : étude de faisabilité avant passage à l'échelle

L'exploitation d'un petit corpus ne peut pas donner des résultats à la hauteur de l'exploitation d'un grand corpus. Elle peut en revanche permettre de tester toutes les facettes d'une procédure outillée d'exploitation, depuis les modes de requêtes envisagés jusqu'aux formats de fichiers traités et exportés.

L'exploration et l'analyse du corpus MC4 ont ainsi permis de mettre à jour un grand nombre de difficultés quant à l'exploration de données annotées portant sur la référence et la coréférence. Comme le montre (Landragin *et al.*, 2015) sur l'un des textes du corpus, les linguistes participant à MC4 ont vite exprimé un ensemble de besoins :

- visualisation (graphique) des chaînes de coréférences dans le texte, avec par exemple un surlignage coloré des expressions référentielles (une couleur par référent) ;
- filtrage des référents lors de cette visualisation, pour ne faire apparaître par exemple que les personnages principaux (en utilisant pour cela un seuil sur le nombre de maillons des chaînes) ;
- extraction des chaînes de coréférences pour en visualiser directement – par exemple sous la forme d'un tableau – la liste des maillons (avec éventuellement un contexte gauche et un contexte droit pour chacun d'entre eux) ;
- affichage aligné des maillons de deux ou de plusieurs chaînes de coréférences, de manière à permettre des comparaisons visuelles rapides de leur constitution ;

- calculs de fréquences et de ratios – par exemple de la proportion de pronoms dans l'ensemble des maillons d'une chaîne – pour permettre de caractériser les chaînes : chaînes majoritairement pronominales *versus* chaînes comportant de nombreuses redénominations du référent, notamment ;
- calculs de nouvelles mesures adaptées à une analyse des références d'un texte, notamment une mesure de la densité référentielle des paragraphes, c'est-à-dire le nombre d'expressions référentielles par rapport au nombre de mots que comporte le paragraphe – ce qui permet de distinguer les paragraphes « très » référentiels des paragraphes « peu » référentiels ;
- croisement des données annotées, de manière à n'explorer que les maillons affectés de telle fonction grammaticale ou de tel rôle thématique, etc.

La plupart de ces fonctionnalités ne sont pas implémentées dans les outils tels que MMAX2, GLOZZ et ANALEC. S'il est effectivement possible de visualiser graphiquement les chaînes d'annotation (ce sont d'ailleurs trois outils conçus dans ce but), l'éventail des possibilités offertes ne va pas jusqu'à l'affichage aligné des maillons de même catégorie d'un ensemble de chaînes, tout simplement parce que ces besoins sont trop spécifiques des chaînes de coréférences. Soit l'on tente d'utiliser un outil existant et de contourner ses fonctionnalités pour un usage adapté aux chaînes, soit l'on développe un nouvel outil – ou un nouveau module d'un outil existant – dédié aux chaînes. Suite à l'expérience MC4 et face aux besoins exprimés, c'est cette dernière solution qui a été retenue pour ANALEC et qui a été présentée dans (Landragin, 2016). Dans sa dernière version, ANALEC permet de visualiser les chaînes de coréférences avec quelques facilités ergonomiques inspirées des besoins exprimés ci-dessus.

Le petit corpus MC4 a ainsi permis de préciser une démarche outillée d'exploration de corpus et d'enrichir un outil existant pour en faire un outil adapté aux chaînes de coréférences. Les prochains corpus annotés en références et en chaînes pourront bénéficier de ces nouvelles fonctionnalités.

Par ailleurs, nous noterons que c'est sur la base des analyses effectuées avec le corpus MC4 qu'un jeu de données annotées a pu servir de support de collaboration pour le montage d'un nouveau projet de recherche, le projet DEMOCRAT. Même si le schéma d'annotation de MC4 n'a pas été reconduit, même si de nouvelles réflexions et de nouvelles expérimentations (chronométrées, impliquant plusieurs annotateurs) ont été nécessaires lors de la mise en place du projet DEMOCRAT, c'est bien l'expérience du projet MC4 qui a permis de tester les possibilités d'exploration et de prévoir les exploitations TAL et ADT explicitées dans les objectifs de DEMOCRAT.

5. Calcul de la taille d'un petit corpus

MC4 comporte 20.000 mots, DEMOCRAT en prévoit 1.000.000. On pourrait ainsi croire que le passage à l'échelle revient à une multiplication par 50. Sauf que la refonte du schéma d'annotation a conduit à une réduction drastique de la complexité de celui-ci : annoter 1.000 mots dans DEMOCRAT prendrait *grosso modo* autant de temps qu'annoter 200 mots dans MC4. Et encore, il s'agit d'une estimation très imprécise dans la mesure où la variabilité des phénomènes de référence fait qu'annoter un paragraphe d'un texte A peut prendre 10 fois plus de temps qu'annoter un paragraphe de taille comparable dans un texte B. En fin de compte, le nombre de mots n'a aucune importance quand on parle d'un corpus annoté sur des phénomènes comme l'anaphore, la référence ou la coréférence.

Cette remarque nous semble juste quel que soit le corpus, et les termes de « petit corpus » et de « grand corpus » ne sont d'aucune aide : un corpus de 10 millions de mots contenant 10.000 anaphores annotées – par exemple un certain type d'anaphore associative, en suivant par exemple des caractérisations linguistiques approfondies (Kleiber, 2001) – reste finalement un petit corpus, alors qu'un corpus de 50.000 mots annoté selon toutes les dimensions d'analyse de la langue qui ont un lien avec la (co)référence (c'est-à-dire beaucoup !) peut avoir demandé des mois de travail. Le terme de « petit corpus » sera alors pour le moins vexant pour les annotateurs.

À ce titre, la durée requise par l'annotation manuelle peut être considérée comme un indicateur qui complète efficacement ceux correspondant au nombre de mots (20.000 pour MC4), au nombre de marquables (4.000 pour MC4), voire au nombre d'étiquettes incluses dans le schéma d'annotation (78 pour MC4). Beaucoup de concepteurs de corpus s'en tiennent au nombre de mots, mais la multiplicité des indicateurs est bien plus informative et, surtout, justifie beaucoup plus la notion de petit corpus. À titre d'exemple parmi d'autres, (Grouin *et al.*, 2011) indique clairement qu'un sous-corpus de 400 phrases (11.400 tokens) a été extrait d'un corpus pour être ré-annoté manuellement par quatre annotateurs, avec une durée de travail estimée à 90 heures (durée d'annotation cumulée avec la durée d'adjuvication). Ce sous-corpus peut être considéré comme un « petit corpus », non seulement par son but de validation d'une procédure (et d'obtention d'une référence), mais aussi par le faisceau d'indicateurs donné.

D'une manière générale, pour renseigner la taille d'un corpus, il nous semble utile de présenter plusieurs indicateurs complémentaires, par exemple sous la forme d'un tableau de nombres qui comporterait :

1. Le nombre de mots : c'est bien entendu utile et nécessaire, mais le chiffre ne tient pas compte des annotations – même en comptant le nombre de mots que comportent ces annotations elles-mêmes, notamment quand il s'agit d'interprétations verbalisées plutôt que codées *via* des étiquettes. Le nombre de mots caractérise la quantité de textes regroupés dans le corpus. C'est un indicateur précieux, surtout pour les applications ADT et TAL. Ce n'est cependant pas totalement suffisant non plus : dans certains cas particuliers de corpus, par exemple quand celui-ci sert à comparer plusieurs versions successives d'un même texte, il serait utile de le compléter par la taille en nombre de mots du texte initial (ou final). On pourrait aller jusqu'à caractériser numériquement la redondance, mais cela concerne certaines analyses ADT et non la description du corpus lui-même.

2. Le nombre de marquables : il s'agit cette fois d'une indication précieuse sur la quantité des annotations ajoutées aux textes. En complément du nombre de mots, on peut ainsi calculer le ratio : nombre de marquables par nombre de mots, c'est-à-dire la densité d'annotations. Le principal problème est de déterminer quels sont les marquables concernés. L'outil GLOZZ annote automatiquement tous les paragraphes dès l'importation d'un texte brut ; l'outil TXM repère tous les mots comme des marquables potentiels et les délimite : ce sont autant de marquables qu'il ne faudrait pas comptabiliser. Un corpus dont les textes sont passés par un analyseur morphosyntaxique et/ou syntaxique pose une question légèrement différente : faut-il comptabiliser les annotations produites par des outils de TAL ? La réponse varie selon l'attention portée à ces annotations : des annotations issues d'un outil mais vérifiées et corrigées manuellement sont bien entendu bien plus précieuses que les sorties brutes d'un logiciel, avec le bruit et le silence qui les caractérisent. Pour bien faire, il faudrait indiquer deux chiffres – le nombre de marquables obtenus automatiquement et le nombre de marquables obtenus manuellement – voire trois chiffres : ceux obtenus automatiquement, ceux obtenus par correction manuelle d'annotations automatiques, et ceux créés manuellement. N'indiquer que le nombre total de marquable revient à donner plus d'importance à la procédure d'annotation automatique qu'à la procédure d'annotation manuelle, celle-ci concernant généralement moins d'occurrences.
3. Le nombre de structures de traits qui constituent les annotations (manuelles comme automatiques), avec éventuellement le nombre de traits et le nombre total d'étiquettes (ou le nombre moyen d'étiquettes par trait). Les annotations du corpus MC4 prennent la forme d'une seule structure de traits, répartie selon 11 traits avec un nombre moyen de 7 étiquettes par traits. Pour d'autres corpus et notamment pour le corpus

DEMOCRAT en cours de constitution, le cas de figure est plus complexe : certains traits sont remplis automatiquement à partir des étiquettes saisies manuellement dans d'autres traits (et à partir des formes de surface des marquables). Comme pour MC4, les annotations ne prennent la forme que d'une seule structure de traits mais, pour la clarté de l'indication, mieux vaut distinguer une structure de traits avec les seuls traits saisis manuellement, et une autre avec les traits remplis automatiquement.

4. Le nombre d'heures de travail qu'ont demandé la constitution et surtout l'annotation d'un corpus – en incluant éventuellement l'adjuvication, ou au contraire en la comptabilisant séparément. C'est une information intéressante pour le chercheur qui se pose la question d'exploiter tel ou tel corpus disponible. En effet, on peut au premier regard sous-estimer le travail que requiert l'annotation manuelle de plusieurs milliers de marquables ou la correction manuelle d'annotations automatiques. Avec l'indication explicite du nombre d'heures que ce travail a demandé, il est plus facile de se faire une idée des efforts effectués, même si un chiffre ne résume jamais un travail, quel qu'il soit. L'intérêt de ce nombre est par ailleurs de mettre en avant les procédures manuelles : toutes les procédures automatiques – annotation morphosyntaxique, par exemple – ne l'augmentent que très peu. C'est peut-être l'indicateur le plus valorisant pour un corpus qui a demandé beaucoup de travail, y compris pour un petit corpus. En revanche, il peut s'avérer difficile pour une équipe comme celle d'un projet collaboratif d'évaluer le nombre d'heures dépensées, surtout quand la tâche s'est répartie entre plusieurs laboratoires, entre plusieurs années de travail, voire entre plusieurs projets institutionnels. C'est le cas du projet MC4, pour lequel nous serions à l'heure actuelle incapable d'évaluer un tel chiffre. Pire : les concepteurs d'un corpus peuvent tout à fait vouloir cacher ce nombre – qui d'ailleurs

> n'est absolument pas vérifiable, même quand les annotations sont étiquetées (comme avec l'outil GLOZZ) par une date GMT. Les raisons sont multiples : trop peu de temps passé par rapport aux ambitions initiales ; trop de temps passé compte tenu du résultat (modeste) obtenu (c'est un peu le cas du corpus MC4) ; non prise en compte du temps important passé à déterminer le schéma d'annotation ; trop grande variabilité entre le temps d'annotation d'un texte en fin de projet par rapport au temps requis en début de projet ; impossibilité de mesurer l'efficacité d'un annotateur, et encore moins son degré de concentration variable d'un texte à l'autre, etc. Comme tout indicateur, ce nombre est informatif mais doit être considéré avec précaution – et avec indulgence.

Le calcul et l'interprétation de ces nombres – potentiellement une dizaine – peut s'avérer difficile, notamment quand le corpus se focalise sur des annotations ponctuelles ou très spécifiques. Par ailleurs, les petits corpus qui sont réalisés manuellement et avec des règles de constitution et d'annotation peu précises (relâchement de contraintes de représentativité, par exemple) sont plus susceptibles que les grands corpus de présenter des manques d'homogénéité dans la réalisation des annotations. Comme ils n'ont pas volonté à nourrir des systèmes de TAL ou des analyses ADT, on leur pardonne volontiers – pas aux grands corpus ! –, mais c'est aussi une source de biais dans le calcul et l'interprétation de tous ces indicateurs de taille.

6. Petit corpus ne doit pas entraîner petite documentation

La conception d'un petit corpus requérant généralement moins d'efforts que celle d'un grand corpus, il peut en être de même de la rédaction de la documentation. Quand un corpus regroupe plusieurs centaines de textes, le temps passé à documenter chacun des textes par des métadonnées (parfois en grand nombre) s'avère en effet beaucoup plus coûteux que le temps passé à documenter les 8 ou 9 textes d'un petit corpus comme MC4.

Mais documenter les textes et les données annotées, si c'est nécessaire, n'est pas forcément le type de documentation prioritaire pour un petit corpus. Comme le petit corpus a une fonction d'étape méthodologique pour illustrer une théorie ou une approche, ou encore un rôle d'étude de faisabilité avant un passage à l'échelle, ce sont ces aspects qui devraient être documentés en priorité. Ainsi, la procédure prime plus que les données obtenues – qui d'ailleurs ne sont statistiquement pas suffisantes –, et il nous semble important de documenter pleinement la démarche qui a conduit à la constitution du petit corpus : théories sous-jacentes, hypothèses linguistiques, méthodologie adoptée, schéma d'annotation déterminé. Ces éléments sont documentés pour un grand corpus, mais devront faire l'objet d'un soin particulier pour un petit corpus, et surtout un petit corpus susceptible de faire l'objet d'un passage à l'échelle. Comme le font dans une certaine mesure (Landragin, 2011), (Landragin & Schnedecker, 2014) ainsi que cet article pour le corpus MC4, il s'agit notamment de décrire :

1. les réflexions ayant conduit au choix d'un schéma d'annotation plutôt qu'un autre ;
2. les tests d'annotations, avec par exemple la réalisation d'expérimentations chronométrées par plusieurs annotateurs (et les accords inter-annotateurs obtenus) ;
3. les tests d'exploitation des données annotées.

Le deuxième point est particulièrement sensible pour des phénomènes linguistiques comme la référence et la coréférence. Nous avons vu à quel point ces phénomènes étaient multi-facettes et pouvaient conduire à diverses manières d'annoter. Même à l'aide d'une documentation complète et d'un manuel d'annotation précis, tout annotateur humain peut interpréter un phénomène de manière subjective et aboutir à une annotation différente de celle que ferait un autre annotateur. L'accord inter-annotateurs reste ainsi peu élevé pour une telle tâche (Artstein & Poesio, 2008). Au-delà de ce constat, il nous semble important que la documentation explicite plusieurs calculs, c'est-à-dire fasse appel à plusieurs métriques classiques – par exemple α, π, κ et désormais γ (Mathet & Widlöcher, 2016) – et, si elle utilise une métrique spécifique, décrive comment

s'opère le calcul et avec quels repères le nombre résultant peut être interprêté.

7. D'un petit corpus à un grand corpus : MC4 à DEMOCRAT

Nous l'avons dit, le « petit » corpus MC4 a servi de préalable pour la préparation du « grand » corpus DEMOCRAT, en cours de constitution et d'annotation. Celui-ci n'étant pas finalisé, nous ne pouvons que donner des indications sur sa nature : il s'agira d'un corpus d'environ 1.000.000 mots, d'environ 200.000 marquables (expressions référentielles) annotés selon une structure de traits comportant une dizaine de traits – un seul annoté manuellement, les autres automatiquement. Le trait annoté manuellement est l'identifiant du référent et permet ensuite à l'outil utilisé – indifféremment ANALEC ou TXM (Heiden *et al.*, 2010) – de construire les chaînes de coréférences. Les chaînes elles-mêmes, qui ne sont plus des marquables mais des éléments décorrélés du texte, seront annotées selon une structure de traits comportant elle aussi une dizaine de traits – la moitié annotée manuellement, l'autre automatiquement.

Sans entrer plus loin dans les détails de ce qui reste à l'état de spéculations, nous noterons que la taille du corpus DEMOCRAT n'aura rien à voir avec celle du corpus MC4, et que le nombre d'heures passées à la constitution, à l'annotation et à l'adjuvication n'auront elles non plus rien à voir avec le temps passé pour MC4 – la différence d'heures étant cependant plus raisonnable que la différence de tailles.

Plusieurs points peuvent être soulignés concernant la démarche d'élaboration du corpus DEMOCRAT :

- Le schéma d'annotation a été complètement revu de manière à simplifier la tâche des annotateurs. Alors que des tests linguistiques étaient nécessaires dans MC4 pour annoter différemment les relatives déterminatives des relatives explicatives (par exemple), une certaine unification de la procédure a été privilégiée dans DEMOCRAT.
- Le schéma d'annotation a été complètement revu de manière à maximiser la partie automatisable. Tout ce

qui a trait à la morphosyntaxe et à la syntaxe a ainsi été spécifié plutôt en fonction des outils de TAL disponibles (et envisageables pour les textes choisis) qu'en fonction des besoins linguistiques. Pour les annotations de ce type, les étiquettes sont désormais issues du monde du TAL plutôt que des modalités d'analyse attendues – les premières pouvant heureusement couvrir les secondes.

- Le schéma d'annotation a été complètement revu de manière à optimiser l'ergonomie de la tâche d'annotation. La construction des chaînes de coréférences se fait ainsi non pas manuellement, en incluant des expressions référentielles dans un ensemble de type « schéma GLOZZ » (Widlöcher & Mathet, 2009), mais automatiquement à partir d'un champ « identifiant du référent » saisi manuellement au niveau de chaque expression référentielle. Ce mode de saisie avait été écarté dans MC4, tout simplement parce que l'outil utilisé – ANALEC – ne permettait pas la complétion automatique des noms de référents, ce qui rendait la saisie répétitive, et donc couteûse et laborieuse. Pour DEMOCRAT, ce sont une version mise à jour d'ANALEC ainsi qu'une version adaptée de TXM qui sont utilisées. Ces deux outils ne présentent plus cette limitation, et des tests de rapidité ont permis de valider la saisie « répétitive », dans la mesure où la complétion la rend « partiellement répétitive », et au final peu coûteuse en temps.
- La procédure d'annotation a été testée et évaluée par le biais d'expérimentations chronométrées (Landragin *et al.*, 2017), ce qui a permis de privilégier des choix – linguistiques, ergonomiques et techniques – plutôt que d'autres.
- Le manuel d'annotation a été écrit en collaboration par les trois annotateurs qui ont mis en œuvre les expérimentations chronométrées, et a été revu ensuite par l'ensemble des participants au projet.

- La rédaction de la documentation a fait l'objet d'un site web participatif, de manière à partager les expériences de chacun, les difficultés rencontrées, les exemples remarquables (dont les spécificités ne sont pas forcément décrites dans le manuel d'annotation), ainsi que les cas particuliers qui résistent aux catégories du schéma d'annotation, et qui peuvent faire l'objet – pour les annotateurs qui le souhaitent – d'une procédure particulière.
- L'exploitation des données annotées, avant même l'obtention de celles-ci dans des proportions intéressantes (ne serait-ce que 5% du corpus), a fait l'objet de plusieurs expérimentations, qui ont conduit notamment à déterminer des mesures adaptées aux chaînes de coréférences, à implémenter des macros TXM capables de calculer ces mesures et de faire ressortir les résultats de manière conviviale.

8. Conclusions et perspectives

La référence et la coréférence sont des phénomènes linguistiques et pragmatiques difficiles à formaliser sous la forme d'une procédure d'annotation de corpus. Nécessairement, les données annotées relèvent d'un compromis entre la finesse des interprétations linguistiques et la rigueur d'un schéma d'annotation ressenti parfois comme un carcan – mais un carcan nécessaire à la faisabilité de l'annotation manuelle et aux possibilités d'exploitations TAL et ADT ultérieures.

Le corpus MC4 présente tous les caractères d'un corpus de très petite taille mais d'utilité pour les chercheurs intéressés par la référence et la coréférence. Il montre les intérêts en nombre non négligeable que présente la réalisation d'un petit corpus avant d'envisager un passage à l'échelle : réflexion méthodologique, étude de faisabilité, expérimentations diverses.

Si sa (très) petite taille peut être considérée comme un inconvénient majeur, il reste néanmoins un avantage certain : celui d'avoir permis d'envisager en toute connaissance de cause et de préparer l'élaboration d'un corpus de plus grande taille, le

corpus DEMOCRAT, y compris en délaissant certaines des voies qui avaient commencé à être explorées.

Au final, il nous semble important que les petits corpus, surtout ceux qui constituent une première étape ou une étude de faisabilité, soient rendus disponibles pour la communauté. Même si les données annotées ne sont pas homogènes, ou pas validées par les incontournables calculs d'accords et métriques diverses. Comme nous l'avons dit en début d'article, le corpus MC4 est disponible librement sur la plateforme Ortolang (https://hdl.handle.net/11403/mc4). Avec en temps voulu le corpus DEMOCRAT, nous espérons qu'il permettra à des chercheurs de s'inspirer de cette expérience – ou au contraire de ne pas la reproduire (mais avec de bons arguments !).

Le corpus MC4 correspond à une étape dans nos réflexions sur la référence et n'est plus voué à évoluer. Les seules perspectives qui le concernent relèvent du projet DEMOCRAT. En plus de la constitution du corpus de ce projet (finalisation du schéma d'annotation et de la procédure complète d'annotation, phases manuelles et automatiques incluses), il s'agit notamment de mettre au jour des modalités d'analyse des données annotées : l'analyse linguistique en corpus d'objets linguistiques aussi complexes que des chaînes de coréférences reste encore à enrichir. Des outils comme MMAX2, GLOZZ, ANALEC et TXM proposent chacun des fonctionnalités, mais il manque probablement une méthodologie d'analyse qui serait applicable à n'importe quel texte, quels que soient sa longueur, ses types de référents les plus fréquents, ou encore son genre textuel.

Remerciements

Ce travail a été réalisé avec le soutien de l'ANR dans le cadre du projet DEMOCRAT – ANR-15-CE38-0008 – qui s'est fondé sur le projet PEPS (Projet Exploratoire Premier Soutien – CNRS) MC4. Il a bénéficié de réflexions et de discussions avec des chercheurs des laboratoires Lattice, LiLPa, ICAR et IHRIM : merci à eux.

Références bibliographiques

Artstein R., Poesio M. (2008). « Inter-Coder agreement for Computational Linguistics », *Computational Linguistics*, vol. 34 : 555-596.

Cabredo Hofherr P. (2008). « Les pronoms impersonnels humains – syntaxe et interprétation », *Modèles linguistiques*, tome XXIX-1, vol. 57 : 35-56.

Charolles M. (2002). *La référence et les expressions référentielles en français*. Paris : Ophrys.

Charolles M., Le Goffic P. (2015). *Beaucoup de sens en si peu de mots.* L'Occupation des sols *de Jean Echenoz. Analyse linguistique d'un texte littéraire. Revue Sciences/Lettres*, vol. 3, ENS : https://rsl.revues.org/.

Chastain C. (1975). « Reference and context ». In: Gunderson K. (Ed.), *Language, mind, and knowledge*. Minneapolis : University of Minnesota Press.

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

Cornish, F. (1999). *Anaphora, Discourse and Understanding*. Oxford: Oxford University Press.

Denis P. (2007). *New Learning Models for Robust Reference Resolution*. Ph.D. dissertation, Austin, University of Texas.

Désoyer A., Landragin F., Tellier I., Lefeuvre A., Antoine J.-Y. (2014). « Les coréférences à l'oral : une expérience d'apprentissage automatique sur le corpus ANCOR ». *Traitement Automatique des Langues*, vol. 55, n° 2 : 97-121.

Fløttum K., Jonasson K., Norén C. (2007). *On : pronom à facettes*. Bruxelles : De Boeck/Duculot.

Fort K. (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 13.

Ghaddar A., Langlais P. (2016). « Wikicoref: An english coreference-annotated corpus of wikipedia articles ». In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia.

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

Grosz B. J., Joshi A. K., Weinstein S. (1995). « Centering: a framework for modeling the local coherence of discourse », *Computational Linguistics*, vol. 21, n° 2 : 203-225.

Grouin C., Rosset S., Zweigenbaum P., Fort K., Galibert O., Quintard L. (2011). « Proposal for an extension of traditional named entities: From guidelines to evaluation, an overview ». In *Proceedings of the 5th Linguistic Annotation Workshop*, Portland, Oregon, pp. 92-100.

Habert B. (2005). *Instruments et ressources électroniques pour le français*, Paris : Ophrys.

Habert B., Nazarenko A., Salem A. (1997). *Les linguistiques de corpus*. Paris : Armand Colin.

Heiden S., Magué J.-P., Pincemin B. (2010). « TXM : une plateforme logicielle open-source pour la textométrie – conception et développement ». Actes de *10th International Conference on the Statistical Analysis of Textual Data (JADT 2010)*, Vol. 2 : 1021-1032.

Kleiber G. (2001). *L'anaphore associative*. Paris : PUF.

Landragin F. (2004). « Saillance physique et saillance cognitive », *Cognition, Représentation, Langage (CORELA)* vol. 2, no 2 : https://corela.revues.org/603.

Landragin F. (2011). « Une procédure d'analyse et d'annotation des chaînes de coréférence dans des textes écrits », *Corpus*, vol. 10 : 61-80.

Landragin F. (2016). « Conception d'un outil de visualisation et d'exploration de chaînes de coréférences », *Thirteen*

International Conference on Statistical Analysis of Textual Data (JADT), Nice, 109-120.

Landragin F., Poibeau T., Victorri B. (2012). « ANALEC: A new tool for the dynamic annotation of textual data », *8th International Conference on Language Resources and Evaluation*, Istanbul, 357-362.

Landragin F., Potier J., Bothua M. (2017). « Annotation manuelle d'expressions référentielles : expérimentations pour simplifier les prises de décisions et optimiser le processus », *9e Journées Internationales de la Linguistique de Corpus (JLC 2017)*, Grenoble.

Landragin F., Schnedecker C. (2014). *Les chaînes de référence. Langages*, vol. 195, Paris : Larousse.

Landragin F., Tanguy N., Charolles M. (2015). « Références aux personnages dans *L'occupation des sols* : apport de la linguistique outillée », *Revue Sciences/Lettres*, vol. 3, ENS : https://rsl.revues.org/.

Lebart L., Salem A. (1994). *Statistique textuelle*, Paris : Dunod.

Mathet Y., Widlöcher A. (2016). « Évaluation des annotations : ses principes et ses pièges », *Traitement Automatique des Langues*, vol. 57, n° 2 : 73-98.

Mélanie-Becquet F., Landragin F. (2014). « Linguistique outillée pour l'étude des chaînes de référence : questions méthodologiques et solutions techniques », *Langages*, vol. 195 : 117-137.

Müller C, Strube M. (2006). « Multi-Level Annotation of Linguistic Data with MMAX2 ». In: Braun S., Kohn K., Mukherjee J. (Eds.). *Corpus Technology and Language Pedagogy. New Resources, New Tools, New Methods*. Frankfurt : Peter Lang, pp. 197-214.

Muzerelle J., Lefeuvre A., Schang E., Antoine J.-Y., Pelletier A., Maurel D., Eshkol I., Villaneau J. (2014). « ANCOR CENTRE, a large free spoken French coreference corpus: description of the resource and reliability measures ». In *Proceedings of the 9th International Conference on*

Language Resources and Evaluation (LREC 2014), Reykjavik, Iceland.

Ogrodniczuk M., Głowińska K., Kopeć M., Savary A., Zawisławska M. (2015). *Coreference in Polish: Annotation, Resolution and Evaluation*. Berlin : Walter De Gruyter.

Pincemin B. (2004). « Lexicométrie sur corpus étiquetés », *Le poids des mots. Actes des 7e Journées internationales d'Analyse statistique des Données Textuelles (JADT 2004)*, Louvain-la-Neuve : Presses universitaires de Louvain, pp. 865-873.

Poudat C., Landragin F. (2017). *Explorer un corpus textuel. Méthodes – pratiques – outils.* Louvain-la-Neuve : De Boeck Supérieur.

Schäfer U., Spurk C., Steffen J. (2012). « A fully coreference-annotated corpus of scholarly papers from the ACL anthology ». In *Proceedings of 24th International Conference on Computational Linguistics (COLING 2012)*, Mumbai, pp. 1059-1070.

Schnedecker C. (1997). *Nom propre et chaîne de référence*. Paris : Klincksieck.

Schnedecker C. (2005). « Les chaînes de référence dans les portraits journalistiques : éléments de description », *Travaux de Linguistique*, vol. 51 : 85-133.

Tellier I. (2009). « Apprentissage automatique pour le TAL : Préface ». *Traitement Automatique des Langues*, vol. 50, n° 3 : 7-21.

Turenne N. (2016). *Analyse de données textuelles sous R*. Londres : Éditions ISTE.

Van Deemter K., Kibble R. (2000). « On Coreferring: Coreference in MUC and related annotation schemes », *Computational Linguistics*, vol. 26, n° 4 : 629-637.

Widlöcher A., Mathet Y. (2009). « La plate-forme Glozz : environnement d'annotation et d'exploration de corpus »,

16e Conférence sur le Traitement Automatique des Langues Naturelles, Senlis.