Analyse, visualisation et identification automatique des chaînes de coréférences : des questions interdépendantes ?

Frédéric Landragin

Analyse, visualisation et identification automatique des chaînes de coréférences : des questions interdépendantes ?

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

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

Résumé

Une chaîne de coréférences est une structure qui regroupe un ensemble d'expressions référentielles (ou mentions, ou maillons) désignant toutes la même entité extralinguistique. Chaque maillon peut être enrichi par des annotations linguistiques, de même que les relations reliant certains maillons. En conséquence, il est difficile d'appréhender une telle structure et d'en tirer directement des analyses. Nous présentons des repères méthodologiques importants pour favoriser l'exploitation d'un corpus annoté en chaînes, tout en précisant les liens opérant entre linguistique théorique, linguistique de corpus outillée et traitement automatique.

Mots-clés : expression référentielle, chaîne de coréférences, corpus, outil d'annotation, apprentissage artificiel.

Abstract: Analyzing, Visualizing and Automatically Identifying Coreference Chains: Interrelated Issues?

A coreference chain groups a set of referring expressions (or mentions) that all refer to the same extra-linguistic entity. Each mention may be annotated with linguistic interpretations, as well as links between mentions. As a consequence, one can find difficult to apprehend and quickly analyse a coreference chain. We present methodological prerequisites, so that the resulting annotated corpus can be exploited as well for machine learning purpose than for linguistic deep analyses. We therefore give precisions on the links that put together theoretical linguistics, computer-aided corpus linguistics, and natural language processing.

Keywords: referring expression, coreference chain, corpus, annotation tool, machine learning.

1. Introduction

L'étude en corpus de chaînes de coréférences pose des questions nouvelles, du moins par rapport à des études en corpus plus usuelles, comme celles de phénomènes lexicaux et syntaxiques. Les interprétations linguistiques liées aux catégories morphosyntaxiques, à la lemmatisation et à l'analyse syntaxique font depuis des années l'objet de très nombreuses recherches en linguistique et en traitement automatique des langues (TAL), ce qui a entraîné une multiplication des corpus annotés autour de ces phénomènes (Bilger, 2000). Ont été ainsi élaborées des méthodologies de constitution et d'exploration de corpus annotés (Fort, 2012), ainsi que de nombreux outils informatiques (Habert, 2005) qui implémentent ces méthodologies. Mais les corpus annotés focalisés sur des phénomènes sémantiques et pragmatiques sont moins nombreux et souvent moins consensuels. Pour la langue française, il existe quelques tentatives de constitution de corpus annotés avec des relations anaphoriques, et une seule initiative d'ampleur, celle du corpus ANCOR (Lefeuvre *et al.*, 2014). Pour les chaînes de coréférences, il n'existe tout simplement aucun corpus de grande taille pour la langue française. Par conséquent, la méthodologie à mettre en œuvre doit reposer soit sur des travaux développés pour d'autres langues comme l'espagnol (Recasens, 2010), l'anglais (Lassalle, 2015) ou le polonais (Ogrodniczuk *et al.*, 2015), soit sur de nouvelles méthodologies qu'il s'agit encore de préciser.

De fait, la différence principale entre les chaînes de coréférences et les phénomènes morphosyntaxiques ou syntaxiques sont leur étendue dans les textes : une chaîne de coréférences – ou « chaîne de référence » (nous ne ferons pas de distinction ici) – est une structure qui regroupe un ensemble d'expressions référentielles – ou « mentions » ou encore, vu que l'on parle de chaîne, « maillons » – désignant toutes la même entité extralinguistique, par exemple un personnage, un objet concret ou abstrait dont il est question dans le texte. Contrairement aux phénomènes lexicaux, morphosyntaxiques, syntaxiques, et même à un grand nombre de phénomènes sémantiques, cette structure peut contenir des éléments issus de plusieurs phrases. Une chaîne de coréférences n'est donc pas un phénomène local, relié à un mot ou à une phrase. Il s'agit au contraire d'une structure discursive, qui couvre potentiellement toute la longueur du texte et pose ainsi des problèmes d'appréhension : il est difficile de garder en mémoire les caractéristiques des chaînes affectées aux personnages d'un roman, par exemple. Plus peut-être que pour tout autre type d'annotation, la manipulation de chaînes de coréférences nécessite des outils informatiques. De même, il est difficile d'analyser des chaînes de coréférences, que ce soit à l'aide de calculs de fréquences ou d'analyses statistiques plus complexes (Lebart & Salem, 1994 ; Turenne, 2016). Là aussi, le linguiste a besoin d'outils, et les outils couramment utilisés pour la gestion, l'annotation et l'exploration de corpus ne sont malheureusement pas encore bien adaptés aux chaînes (Poudat & Landragin, 2017).

Après une description dans la section 2 des spécificités des chaînes de coréférences, nous détaillons dans la section 3 les approches de la linguistique de corpus outillée, et nous montrons notamment qu'il n'existe que très peu d'outils capables d'aider le linguiste à annoter, visualiser et analyser des chaînes de coréférences. Nous y présentons nos propres propositions en cours d'expérimentation dans le cadre du projet ANR DEMOCRAT mentionné dans l'article de présentation de ce numéro. Avec la section 4, nous abordons alors la facette du traitement automatique, et donc de l'identification automatique de chaînes de coréférences dans des textes tout-venants, ce que les technologies actuelles ne peuvent réaliser qu'en se nourrissant au préalable de données déjà annotées, et donc de corpus annotés manuellement, comme ceux dont il est question dans les sections 2 et 3. Or la communauté du TAL, si elle s'est intéressée aux chaînes de coréférences, l'a fait surtout pour la langue anglaise ou d'autres langues bien représentées dans les compagnes d'évaluation internationales, mais très peu pour la langue française. Nous discutons des techniques récentes et fortes d'enjeux importants. Sur la base des contraintes et apports de chacune des branches méthodologiques décrites – étude linguistique en corpus, visualisation et exploration de données, TAL – nous proposons alors, dans la section 5, un ensemble de précautions méthodologiques pour l'étude en corpus de chaînes de coréférences, en soulignant notamment l'interdépendance des questions soulevées et l'intérêt de travailler avec des préoccupations à la fois théoriques, pratiques et automatiques.

2. Les chaînes de coréférences : des objets d'étude particuliers

Pour étudier les chaînes de coréférences d'un texte, il est tout d'abord nécessaire de les repérer. Une chaîne étant un ensemble regroupant plusieurs expressions référentielles, une première étape consiste donc à repérer ces expressions référentielles. Comme une chaîne regroupe les expressions qui réfèrent au même objet, il est nécessaire de résoudre les références : une deuxième étape consiste donc à résoudre toutes les références, problème bien connu en linguistique et en philosophie du langage (Charolles, 2002), qui a entraîné la publication de plusieurs milliers d'articles de recherche, fondés sur des exemples extraits de corpus aussi bien que sur l'intuition des chercheurs. Beaucoup parmi ces travaux se sont intéressés aux ambiguïtés référentielles, à des distinctions entre plusieurs catégories de références (spécifique *versus* générique, directe *versus* indirecte, réelle *versus* virtuelle, et ainsi de suite). Tous ces centres d'intérêt, toutes ces distinctions sont à prendre en considération lors de cette deuxième étape, car les interprétations permettront d'affecter une expression référentielle à une chaîne plutôt qu'à une autre, voire à aucune chaîne – faute d'un consensus interprétatif – ou à deux chaînes à la fois. Nous ne reviendrons pas ici sur les détails de cette étape, car notre objet d'étude est la chaîne de coréférences : quel que soit le chemin parcouru pour la construire, c'est le résultat qui compte plutôt que les détails qui ont mené à tel ou tel choix. Enfin, une troisième étape consiste à construire les chaînes de coréférences proprement dites, c'est-à-dire à regrouper – conceptuellement, mais aussi par le biais de manipulations informatiques – les expressions référentielles dont les résolutions ont abouti au même résultat, à savoir le même référent. Si celui-ci est clairement identifié et peut être caractérisé, une quatrième et dernière étape revient à décrire ces caractéristiques dans des annotations, qui sont spécifiques aux chaînes et permettent de catégoriser et de comparer celles-ci.

Un référent pouvant apparaître du début jusqu'à la fin d'un texte, la chaîne de coréférences regroupant les mentions de ce référent peut couvrir toute la longueur du texte. Cette couverture peut selon les cas être qualifiée de dense (quand le nombre de mentions par paragraphe est élevé), de régulière (quand le référent est mentionné dans chaque paragraphe par exemple), d'éparse (quand des passages où le référent est mentionné alternent avec d'autres passages où il est totalement absent), etc. Étudier les chaînes de coréférences d'un texte peut ainsi amener à comparer, en plus de leur cardinal (nombre de mentions), des notions comme la densité référentielle, ratio entre le nombre de mentions référentielles et le nombre de mots d'un paragraphe. Ce qui fait autant d'observables envisageables, à analyser en parallèle avec les formes de référence constituant les maillons de la chaîne. L'objet d'étude est donc complexe : il s'agit d'un ensemble de *formes linguistiques*, prises *çà et là dans le texte*, et dont le regroupement s'est opéré via une *interprétation contextuelle*, connue pour sa complexité (Charolles, 2002). Or chacun de ces aspects pose ses propres problèmes.

Tout d'abord quand on parle de *formes linguistiques*. Nous avons jusque-là utilisé le terme *expression référentielle*, qui regroupe « classiquement » les noms propres, les groupes nominaux et les pronoms, du moins pour des référents concrets. Pour des référents abstraits tels que des propositions (reprises par exemple avec « ça arrive et c'est dommage »), on étend la notion à des propositions, voire des paragraphes complets, qui sont considérés comme le premier maillon de la chaîne de coréférences incluant aussi « ça » et « ce ». Nous avons réalisé deux expériences d'annotation : une qui se restreint aux seules expressions référentielles « classiques », et l'autre qui étend la délimitation des maillons à toute forme de référence. Dans le premier cas, les chaînes de coréférences liées à des référents abstraits sont souvent incomplètes : elles ne sont complètes que lorsque le premier maillon est lui-même un groupe nominal. Dans le second cas, la notion de *maillon* devient polymorphe, ce qui peut perturber non seulement l'annotateur, mais aussi les résultats d'analyses statistiques : on aimerait en effet pouvoir parcourir et comparer les expressions référentielles, sans être gêné par toutes les autres formes de référence qui ne sont pas aussi fréquentes. Pour rendre possibles de telles comparaisons, il est nécessaire de catégoriser chaque maillon : soit « expression référentielle », soit « autre forme de référence ». Même si elle restreint le problème, la référence aux objets concrets – comme les référents humains – pose elle aussi des questions spécifiques. Dans (Landragin & Schnedecker, 2014), l'article de F. Landragin et N.

Tanguy montre par exemple que le pronom « on » peut être considéré selon les cas comme plus ou moins référentiel, voire impersonnel. Une occurrence de « on » peut donc n'intervenir dans aucune chaîne de coréférences, ou – vaguement – dans une voire plusieurs chaînes. Enfin, le cas des sujets non exprimés, par exemple dans « il entra et prit un café », pose aussi un problème de définition : considère-t-on que le vide devant le verbe est une forme de référence ? Cette solution n'est pas viable informatiquement, mais les solutions envisageables sont peu nombreuses : soit on fait apparaître artificiellement le sujet zéro dans le texte, par exemple avec le symbole « Ø » (que l'on délimite alors comme maillon), soit on délimite la forme verbale. Une indication spécifique est alors indispensable : elle permettra aux analyses de données textuelles (ADT, cf. la série de conférences JADT, avec par exemple (Landragin, 2016)) et de TAL d'ignorer les formes verbales dans leurs processus.

Tournons-nous ensuite vers l'expression *çà et là dans le texte*. La chaîne de coréférences n'a d'existence qu'à travers cette dispersion dans le corpus. Or la linguistique de corpus s'est surtout focalisée sur des phénomènes locaux. Visualiser et explorer des données annotées n'est possible que parce que ces données sont circonscrites : les annotations sont affectées à des *markables* qui, dans la très grande majorité des cas, sont des mots ou des suites de mots. En ADT, choisir des lois statistiques et des méthodes de calcul n'est possible que parce que les éléments de base du texte le permettent : ces éléments de base sont souvent les mots présents dans le texte, auxquels sont parfois ajoutés les lemmes, et éventuellement – mais c'est déjà plus rare – les étiquettes morphosyntaxiques. Tenir compte à la fois des mots et de structures annotées complexes comme le sont les chaînes de coréférences nécessite de reprendre les méthodologies d'ADT et, avant toute chose, d'identifier des mesures adaptées aux chaînes.

Enfin, reste l'*interprétation contextuelle*. Nous avons mentionné rapidement *supra* quelques distinctions classiques dans les recherches sur la référence, dont la connaissance permet d'attribuer à une forme de référence un référent plus ou moins bien identifié. D'autres phénomènes ont fait également l'objet de nombreux travaux et se rencontrent dès que l'on annote les chaînes de coréférences. Mentionnons les référents évolutifs, pour lesquels, même quand les expressions référentielles semblent coréférentes deux à deux, le résultat final – c'est-à-dire la chaîne de coréférences – peut s'avérer pour le moins hétérogène. On pensera au poulet que l'on déplume et que l'on cuit (référent évolutif quant à sa nature), mais aussi à un référent de type groupe, tel qu'un parti politique (référent évolutif quant à ses constituants). Dans un même ordre d'idée, la référence reste parfois floue, c'est-à-dire qu'il est impossible de déterminer avec exactitude ou avec exhaustivité quels sont les membres d'un groupe, ne serait-ce que lorsqu'un pronom pluriel comme « ils » apparaît après la mention de plusieurs personnes individuelles. Dans (Landragin, 2011), nous avons montré comment une étude de corpus s'était intéressée à des références qui sont réinterprétées plus loin dans le texte (avec un référent mis à jour), et comment cette réinterprétation pose des problèmes d'annotation : annote-t-on avec les connaissances qu'a le lecteur en fin de lecture, ou avec ses connaissances au fur et à mesure de la lecture, même si cela se fait au prix de quelques chaînes de coréférences – qui se trouvent séparées en chaînes distinctes plutôt qu'en une seule chaîne globale plus cohérente ?

Non seulement les choix d'annotation peuvent varier d'un projet à l'autre, d'une étude à l'autre, mais, de plus, il peut être difficile de mettre en œuvre une méthodologie d'annotation de corpus avec un accord inter-annotateurs élevé (Fort, 2012). L'interprétation contextuelle fait en effet appel aux connaissances et donc à la subjectivité de l'annotateur. Ce problème – pour lequel il existe des solutions, ne serait-ce que la spécification d'un manuel d'annotation précis et contraignant – n'est pas spécifique aux chaînes de coréférences. Mais il est particulièrement prégnant, et ses conséquences sur la justesse des annotations des chaînes sont importantes.

3. Approches de la linguistique de corpus outillée

Nous présentons dans cette section les bilans de plusieurs expériences d'annotation et d'étude de chaînes de coréférences. Un état de l'art des outils de gestion de corpus a été réalisé au début de chaque expérience, et a montré que les outils permettant d'annoter des chaînes de manière ergonomique et efficace étaient très peu nombreux. Il s'agit soit d'outils dédiés aux relations

anaphoriques puis adaptés aux chaînes – c'est le cas notamment de MMAX2 (Müller & Strube, 2006) – soit d'outils dédiés à des structures discursives complexes comme GLOZZ (Widlöcher & Mathet, 2009) – qui a servi pour l'annotation du corpus ANCOR – et ANALEC, ANALyse de l'ECrit, outil d'annotation manuelle et d'exploration de corpus (Landragin *et al*., 2012), dont nous suivons le développement depuis plusieurs années et qui s'impose donc comme le choix le plus logique dans notre cas.

3.1. Constitution et analyse d'un corpus écrit : bilan du PEPS MC4

Le projet PEPS – Projet Exploratoire Premier Soutien du CNRS – MC4 « Modélisation Contrastive et Computationnelle des Chaînes de Coréférences » a été décrit en détail dans (Landragin & Schnedecker, 2014). Nous voulons ici, trois ans après la fin de ce projet, faire un bilan de la procédure d'annotation et de l'exploitation des données annotées.

MC4 était un projet de petite taille, réunissant une dizaine de chercheurs pendant deux ans pour réfléchir à la référence et à la coréférence, et fournir un corpus qui puisse servir de point de départ et de test de faisabilité pour un projet de plus grande envergure, qui est le projet DEMOCRAT. L'objectif initial était d'annoter les formes de référence non seulement avec les référents attribués, mais aussi et surtout avec de multiples dimensions d'analyse, de manière à croiser *a posteriori* les données, pour observer par exemple la fonction grammaticale privilégiée des maillons de telle ou telle chaîne. Des annotations syntaxiques, sémantiques et pragmatiques s'ajoutent donc à l'identifiant du référent, au point de construire pour chaque maillon une structure de traits comportant pas moins de 11 propriétés, pour un total de 78 valeurs possibles. La taille de la structure de traits a des conséquences sur le temps pris par l'annotation manuelle. Comme de plus le corpus n'était pas la seule tâche du projet, seulement 4.066 formes de référence ont été annotées, permettant la construction de 285 chaînes de coréférences.

Si cette quantité d'annotations permet de réaliser des analyses linguistiques qualitatives, elle en interdit toute analyse quantitative sérieuse. Le nombre de maillons annotés est clairement insuffisant pour entraîner un système d'apprentissage artificiel : à titre de comparaison, le corpus ANCOR comporte un peu plus de 100.000 expressions référentielles annotées (mais avec une structure de traits bien plus réduite), ce qui fait entrer ANCOR dans le club des corpus utilisables pour des applications TAL. De même, le nombre de maillons annotés dans MC4 est insuffisant pour lancer des analyses de type ADT fondées à la fois sur les mots et sur les annotations. Toute analyse quantitative ne peut donc être qu'indicative.

Or l'un des objectifs de MC4 était de faire émerger des tendances à partir des données annotées, sans avoir à explorer manuellement, valeur après valeur, les exemples rencontrés dans le corpus. Mais un constat fait *a posteriori* réside dans la difficulté de faire émerger des observations pertinentes et statistiquement intéressantes, malgré la richesse des dimensions d'analyse. Dans la majorité des études reportées dans (Landragin & Schnedecker, 2014), c'est la conjonction des données annotées et d'une intuition linguistique qui permet de faire émerger des observations, et non les données elles-mêmes. Le travail d'analyse repose ainsi sur une lecture attentive du texte et sur des aspects linguistiques qui ne sont pas encodés dans le corpus. Autrement dit, l'analyse en corpus de chaînes de coréférences reste encore en grande partie manuelle. Logiquement, on s'attend à ce que les outils d'exploration de corpus viennent aider cette analyse, en apportant des facilités de repérage de phénomènes intéressants et des mesures adaptées. Améliorer les outils devient donc un enjeu pour faire progresser les possibilités d'analyse linguistique, et pas seulement un support ou une application.

3.2. Panel d'outils pour analyser : bilan de l'étude d'une nouvelle d'Échenoz

Une étude de la nouvelle *L'occupation des sols* de Jean Echenoz vient compléter le bilan précédent : ce texte faisait partie au départ du corpus MC4, mais des études spécifiques lui ont été appliquées. Dans cette étude décrite dans (Landragin *et al.*, 2015), plusieurs types d'analyse sont menés, avec à chaque fois des calculs particuliers et une spécification des besoins attendus en termes d'outillage informatique.

Un premier effort a été fait pour la visualisation des chaînes de coréférences. En complément des modes de visualisation déjà proposés par l'outil GLOZZ (Widlöcher & Mathet, 2009), nous avons utilisé une procédure et une représentation graphique simple, implémentées dans l'outil ANALEC (Landragin *et al.*, 2012). On choisit l'une des 11 propriétés, par exemple la catégorie de la forme de référence (nom propre, groupe nominal, pronom) ou sa fonction grammaticale (sujet, c.o.d., c.o.i.), et l'outil affecte un code couleur à chaque valeur possible (trois couleurs pour les exemples donnés dans les parenthèses précédentes). Les formes de référence sont alors présentées dans l'ordre linéaire du texte, sous la forme de nœuds colorés. L'analyste peut alors appréhender les annotations avec cette représentation graphique et, par exemple, repérer des zones textuelles remarquables par leur absence de telle ou telle couleur, ou au contraire leur plus grande fréquence. Comme nous le disions à propos du bilan du projet MC4, c'est à l'analyste que revient la tâche de rapprocher sa perception de la représentation graphique, de son intuition par rapport au texte : l'outil ne fait pas émerger de nouvelles observations ; son rôle se limite à indiquer des endroits intéressants qu'il reste à analyser ponctuellement.

Un deuxième effort a été fait pour que l'outil fournisse quelques indicateurs numériques adaptés aux chaînes de coréférences. Il s'agit notamment de décomptes des formes de référence, chaîne par chaîne, paragraphe par paragraphe, et de calculs de la densité référentielle pour chaque paragraphe du texte. Ces indicateurs numériques restent à enrichir, et à rapprocher des mesures envisageables en ADT. Pour l'étude du texte de Jean Echenoz, ils ont néanmoins facilité et quantifié des observations sur les rôles relatifs des personnages principaux.

Un troisième effort concerne l'exploitation de calculs de bi-grammes et de tri-grammes pour proposer une étude semi-automatique des chaînes de coréférences, ainsi que de la suite complète de toutes les formes de référence du texte. Cette exploitation a permis de quantifier des hypothèses linguistiques quant aux transitions référentielles et aux rapports entre changements de paragraphe et constitution de chaînes. Elle a été mise en œuvre à l'aide d'outils disponibles sur le web, et surtout à l'aide de procédures manuelles, ne serait-ce que pour la transformation de la suite des formes de référence en un message écrit dans un alphabet exploitable pour des calculs de bi-grammes et de tri-grammes. Les résultats obtenus, quoique modestes, ont validé la démarche et amené à spécifier un travail de développement informatique pour enrichir ANALEC avec des fonctionnalités ciblées sur les chaînes.

3.3. Outiller l'étude des chaînes de coréférences : les évolutions d'Analec

Le projet MC4 et l'étude du texte de Jean Echenoz ont joué le rôle de révélateurs des lacunes des outils de gestion de corpus et plus spécifiquement d'ANALEC. Nous avons donc augmenté ANALEC de trois nouvelles interfaces : une interface « statistiques » dédiée à l'affichage de statistiques générales sur le texte en cours d'étude dans le logiciel ; une interface « analyse d'une chaîne » pour visualiser et explorer les formes de référence et les données annotées propres à une chaîne de coréférence ; une interface « analyse de la suite des références » pour visualiser et explorer l'ensemble des formes de références annotées, toutes chaînes confondues. Les détails de ces nouvelles fonctionnalités et de leur implémentation ont été présentés lors de la conférence « Journées internationales d'Analyse statistique des Données Textuelles (JADT) » (Landragin, 2016). Ce qui nous importe ici, c'est l'approche suivie : c'est à partir des besoins exprimés par des linguistes – directement (demande de fonctionnalité supplémentaire) ou indirectement (expression

d'une hypothèse linguistique, difficilement vérifiable sans outil informatique) – que nous avons envisagé l'évolution d'un outil d'annotation. Et ces évolutions autorisent de nouvelles possibilités d'exploration, que les linguistes s'approprient petit à petit (Poudat & Landragin, 2017), et qui vont potentiellement leur permettre d'exprimer des hypothèses linguistiques plus précises, ainsi que de nouveaux besoins d'outillage. ANALEC sert actuellement de plateforme d'expérimentation de cette approche caractérisée par des allers-retours incessants entre théorie et pratique. Par rapport à des logiciels plus complexes comme Le Trameur (Fleury, 2013) ou TXM (Heiden *et al.*, 2010), l'intérêt est que l'ajout et le test d'une nouvelle fonctionnalité est peu coûteux en temps et en efforts humains : les allers-retours peuvent se dérouler selon un rythme satisfaisant, en tout cas plus rapide qu'avec un logiciel plus institutionnalisé.

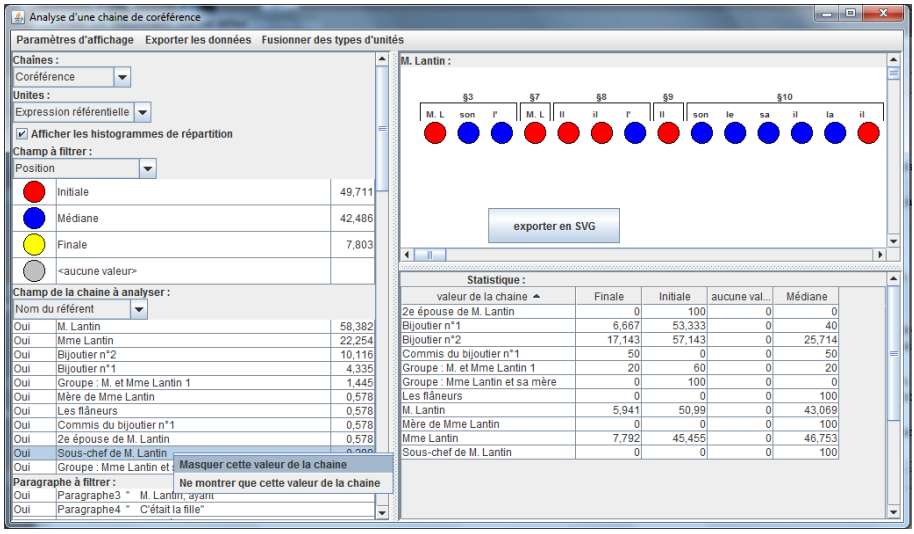
L'interface « statistique » présente un ensemble de décomptes, de répartitions des annotations tout au long du texte (en tenant compte du découpage en paragraphes) et de densités référentielles. Son intérêt est surtout informatif : il ne s'agit pas ici d'explorer les données annotées, mais seulement d'en appréhender la nature et l'étendue à travers quelques quantifications.

L'interface « analyse d'une chaîne » rend opérationnelle l'étude des chaînes de coréférences : l'utilisateur peut se concentrer sur un référent et procéder à diverses visualisations graphiques. Avec le corpus du projet MC4, chacune des 11 propriétés peut ainsi faire l'objet d'un code couleur, ce qui permet d'appréhender en un coup d'œil les écarts de valeurs (figure 1). Visualisation et exploration peuvent se faire globalement sur le texte, ou en sélectionnant quelques paragraphes seulement, ou encore paragraphe par paragraphe.

L'interface « analyse de la suite des références » reprend les mêmes principes, mais sans distinction de chaîne de coréférence : la visualisation graphique comme l'exploration portent sur l'ensemble des formes de référence du texte, ce qui permet d'exploiter des codes graphiques (couleur, taille, texture) pour appréhender les transitions référentielles, les alternances d'une valeur à une autre (pour une propriété donnée) et ainsi de suite.

D'un point de vue informatique, ces trois interfaces sont à l'heure actuelle dans un état plus expérimental que véritablement stable. C'est l'un des objets du projet DEMOCRAT que d'opérationnaliser avec de nouveaux efforts, linguistiques comme informatiques, la démarche ainsi amorcée.

Figure 1 : interface d'ANALEC pour l'analyse d'une chaîne de coréférences. On y voit une représentation graphique à base de ronds colorés : un rond par maillon, une couleur par type d'annotation. On y voit également un ensemble de calculs de fréquences : fréquence de telle annotation (à gauche), fréquence de tel référent (en bas à gauche), fréquence de telle annotation par référent (tableau à droite).



4. Approches du traitement automatique des langues (TAL)

Jusqu'à présent, nous n'avons abordé l'étude des chaînes de coréférences que sous l'angle de la linguistique de corpus outillée : pour mieux étudier ces structures, le linguiste utilise un outil qui lui permet de les construire (tâche d'annotation coûteuse en temps), puis de les explorer. Mais la construction des données constituant le corpus reste manuelle. Dans cette section, nous envisageons l'automatisation de l'annotation, et donc l'appel à des systèmes de TAL dont la tâche est d'identifier automatiquement, en partant d'un texte brut, les chaînes de coréférences liées aux référents mentionnés dans le texte.

4.1. Systèmes à base de règles et systèmes à base d'apprentissage

Qu'on l'envisage comme un système totalement automatique visant à enrichir de gros corpus, ou comme un outil d'aide à l'annotation manuelle, la détection automatique de chaînes de coréférences est un enjeu crucial, qui permettra à terme d'améliorer les outils d'indexation, les moteurs de recherche et d'une manière générale le web sémantique et les humanités numériques.

Cependant, réaliser un système suffisamment souple et caractérisé par un taux d'erreur faible est une tâche extrêmement complexe. En effet, il s'agit d'implémenter la résolution de la référence, ni plus ni moins, donc l'attribution d'un référent à toutes les formes de référence présentes dans le texte de départ, puis le regroupement des formes coréférentielles pour en faire les maillons d'une même chaîne. La tâche inclut donc, sans s'y réduire, celle de la résolution automatique des anaphores, qui a fait l'objet d'efforts importants en TAL et de nombreux systèmes (Mitkov, 2002).

Avant les années 2000, cette tâche était réalisée en spécifiant un système de règles qui – en s'appliquant dans un ordre précis – permettaient de détecter les formes de référence puis les chaînes (en faisant à chaque fois des oublis et des erreurs). Parmi les règles déclarées dans le système, certaines s'inspirent des théories linguistiques, par exemple l'accord en genre et en nombre du pronom anaphorique avec son antécédent, règle qui permet au système de ne pas envisager comme coréférentielles les mentions « la jeune femme » et « il ». On peut paramétrer un système avec quelques dizaines de règles, mais on tombe très vite dans la gestion d'exceptions. En reprenant l'exemple de l'accord en genre et en nombre, les exemples classiques de « la sentinelle » repris par « il », et de « le maire du village » repris par « elle » peuvent être à l'origine de nouvelles règles, ou du moins de conditions dans lesquelles une règle ne doit pas s'appliquer. Dresser à la main la liste des noms de métier va dans ce sens. C'est là que l'on constate les principaux inconvénients des systèmes à base de règles : beaucoup d'efforts sont faits pour les cas particuliers, sans jamais aucune garantie de couvrir un maximum de phénomènes linguistiques. En outre, à chaque fois qu'un nouveau terme, par exemple un nouveau nom de métier, apparaît dans la langue, il est nécessaire de reprendre le système de règles pour le mettre à jour. Toutes ces opérations sont manuelles, donc susceptibles d'erreurs, d'imprécisions et ainsi de suite.

La détection automatique de chaînes de coréférences, comme beaucoup d'autres tâches relevant du TAL, se réalise désormais à l'aide de techniques d'apprentissage artificiel. Ces techniques s'avèrent plus souples, car c'est le système qui définit de lui-même ses propres règles, et amènent globalement à de meilleurs résultats. Mais, pour être mises en œuvre, ces techniques nécessitent d'une part des corpus déjà annotés – en tant qu'exemples d'apprentissage – et d'autre part des ressources linguistiques telles que des dictionnaires de noms propres ou de synonymes – en tant que facteurs potentiels de décision, et donc en tant qu'aides ponctuelles pour réaliser la tâche efficacement. Ces dernières ressources étant utilisées par ailleurs pour l'analyse syntaxique ou d'autres tâches du TAL, leur recours ne pose pas de problème particulier : soit on en dispose et on en fait profiter la détection automatique des chaînes de coréférences, soit on fait sans elles. La principale contrainte est surtout de disposer d'un corpus de qualité, donc avec des annotations vérifiées par des linguistes, et qui plus est de grande taille. En effet, plus le stock d'exemples d'apprentissage sera grand, mieux le système arrivera à spécifier ses propres règles. Même s'il

existe des techniques permettant de se satisfaire d'un corpus de taille moyenne, on considère – pour le problème des chaînes de coréférences – que la limite inférieure « raisonnable » se situe autour de 100.000 formes de référence annotées.

Pour la langue française, il n'existe à l'heure actuelle qu'un seul corpus permettant un apprentissage artificiel de qualité : le corpus ANCOR que nous avons déjà mentionné (Lefeuvre *et al.*, 2014). C'est ce corpus que nous avons exploité pour réaliser une première expérimentation d'apprentissage artificiel, le système CROC – *Coreference Resolution using Oral Corpus* (Désoyer *et al.*, 2014). Il ne s'agit cependant que d'une partie d'un système complet (ou « *end-to-end* »), c'est-à-dire d'un système capable de construire automatiquement les chaînes de coréférences rien qu'à partir du texte brut. En effet, pour fonctionner correctement, le système CROC doit partir d'un texte déjà annoté en formes de référence. Il ne correspond en quelque sorte qu'au dernier processus de la chaîne de traitement automatique, c'est-à-dire de la succession des opérations réalisées.

4.2. Chaîne de traitement automatique

Séparer une tâche en sous-tâches est parfois un moyen commode de délimiter les problèmes et de trouver des solutions adaptées à chaque sous-tâche. De même que l'on peut envisager plusieurs systèmes de règles successifs pour identifier les chaînes de coréférences dans un texte brut, on peut envisager plusieurs traitements en cascade lors de la conception d'un système *end-to-end* basé sur des techniques d'apprentissage artificiel. Comme pour tout système de traitement automatique, plusieurs étapes s'enchaînent, chacune partant des résultats obtenus par la précédente. Ainsi, les deux étapes principales classiquement distinguées (Recasens, 2010 ; Lassalle, 2015) sont premièrement la détection des formes de référence, et deuxièmement l'appariement de plusieurs formes de référence pour en faire une chaîne de coréférence.

Un système commence donc par « lire » le texte pour en détecter les expressions référentielles et, si besoin, les formes de référence plus implicites comme les sujets non exprimés de verbes conjugués ou à l'infinitif. Celles-ci sont alors annotées en tant que telles dans le texte. Autrement dit, elles sont étiquetées. La première étape correspond à une tâche d'étiquetage, pour laquelle certaines techniques d'apprentissage s'avèrent plus performantes que d'autres. C'est aussi l'un des intérêts de dissocier le traitement en deux étapes.

Sur la base du texte étiqueté obtenu, l'étape suivante revient à construire les chaînes de coréférences. Une méthode consiste à sélectionner deux formes de référence, notamment deux formes relativement proches dans le texte (dans la même phrase, dans deux phrases consécutives, ou séparées par un maximum de – par exemple – cinq autre formes de référence), puis à déterminer si ces deux formes sont coréférentielles ou non. Il s'agit cette fois d'une tâche de classification, donc d'une nature différente de l'étape précédente. Les chaînes de coréférences se construisent ainsi par appariements de formes, deux par deux, exactement comme une chaîne réelle s'assemble en joignant des maillons deux par deux. Le résultat, c'est-à-dire les ensembles de formes de référence, est sauvegardé en parallèle du texte, éventuellement sous la forme d'une annotation supplémentaire affectée à chaque forme. Il est exploitable aussi bien informatiquement que par un outil d'annotation comme ceux que nous avons mentionnés pour l'annotation manuelle.

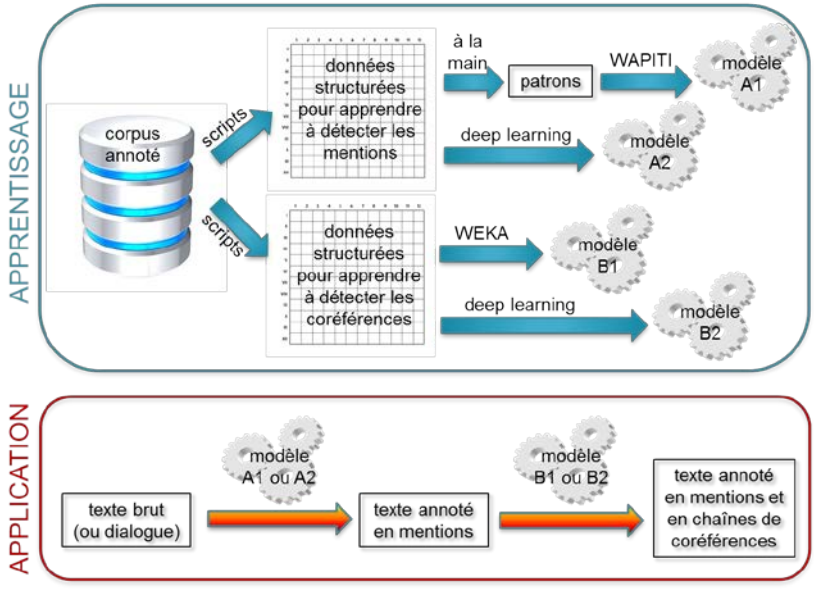
Tout cela concerne le système exécutable, c'est-à-dire l'application finale. La réalisation de ce système est un processus différent, qui consiste en un ensemble de traitements et de séances d'apprentissage. Comme le point de départ est l'exploitation d'un corpus de référence annoté en chaînes de coréférences, il y a tout d'abord une phase – inévitable – de transformation des fichiers composant le corpus. Dans ANCOR par exemple, l'annotation originelle est en relations anaphoriques. Il faut donc reconstituer les chaînes de coréférences avant de les fournir au système d'apprentissage artificiel. C'est l'objectif d'un script de transformation, première étape du processus d'apprentissage. En effet, un système d'apprentissage ne fonctionne pas directement sur un corpus annoté, mais travaille avec des données structurées. Notamment, ces données incluent un certain nombre de facteurs, que l'on considère comme des paramètres potentiels de décision, et qui doivent être explicités en tant que tels. Le ou les scripts ont donc comme tâche d'expliciter tous les facteurs envisageables, et de structurer les données en fonction de ces facteurs. Un immense tableau peut être construit, avec un exemple par ligne, et un facteur par colonne. Pour un corpus comportant

100.000 exemples annotés, c'est ainsi un tableau de 100.000 lignes qui va être construit et dont le système d'apprentissage va se nourrir.

Plusieurs techniques d'apprentissage existent, et chacune d'entre elles apporte ses propres contraintes. Si nous avons distingué une étape de détection des formes de référence et une étape d'appariement de paires de formes, c'est parce que les techniques exploitées diffèrent : champs aléatoires conditionnels (CRF, *Conditional Random Field*) pour la première étape ; machines à vecteurs de support (SVM, *Support Vector Machines*) pour la deuxième étape. Si les SVM se nourrissent directement des données structurées, les CRF nécessitent une sorte de guidage, de manière à mieux exploiter la nature textuelle des données initiales. S'ajoute ainsi une phase de spécification de « patrons », dont le but est de focaliser l'attention de l'apprentissage sur, par exemple, deux ou trois mots consécutifs. Même si elle consiste en quelques déclarations simples, cette phase est manuelle et donc à la discrétion du spécialiste de TAL. D'une manière générale, celui-ci peut, en fonction de ses compétences en apprentissage artificiel, exploiter des plateformes d'apprentissage existantes, notamment les plateformes WAPITI (pour les CRF, cf. https://wapiti.limsi.fr/) et WEKA (pour les SVM, cf. http://weka.wikispaces.com/), ou implémenter son propre algorithme d'apprentissage. Il peut également exploiter la puissance d'abstraction de l'apprentissage profond sur la base de réseaux neuronaux, comme cela se fait de plus en plus actuellement, non seulement pour des tâches de vision artificielle (dont on a pu constater les performances dans l'actualité scientifique récente) mais aussi pour des tâches de TAL.

Pour le système CROC (Désoyer *et al.*, 2014), c'est la plateforme WEKA qui a été utilisée sur une version transformée du corpus ANCOR. La figure 2 donne une idée du processus complet envisagé dans le projet DEMOCRAT. Les résultats obtenus pour le moment ne concernent que la deuxième étape de traitement, et sont donc partiels. En gros, et donc pour la tâche d'appariement seulement, ces résultats sont meilleurs que ceux obtenus par les anciens systèmes à base de règles, et sont à peu près au niveau de l'état de l'art pour les systèmes réalisés sur d'autres langues que le français. On peut donc espérer améliorer les performances, d'une part en continuant le développement informatique, d'autre part en exploitant le futur corpus DEMOCRAT à la place ou en plus du corpus ANCOR.

Figure 2 : les étapes de réalisation et d'exécution d'un système de détection automatique des chaînes de coréférences.



5. De la linguistique au TAL et du TAL à la linguistique

Nous l'avons vu dans la section précédente, détecter automatiquement les chaînes de coréférences est un enjeu en soi, qui peut s'éloigner des études linguistiques et de leurs préoccupations : il s'agit de concevoir des modèles d'apprentissage de plus en plus performants, et pour cela, de constituer des corpus et des ressources linguistiques de taille de plus en plus grande. On peut donc se demander quels sont les liens entre ces efforts du TAL qui concernent des aspects très techniques de l'apprentissage artificiel, et les efforts réalisés par ailleurs en linguistique.

Nous avons identifié plusieurs disciplines, et notamment les trois suivantes : la linguistique théorique et descriptive ; le développement d'outils pour la linguistique de corpus ; le développement de systèmes de TAL. Ce que nous voulons montrer avec l'exemple des chaînes de coréférences, c'est que les préoccupations de ces trois disciplines sont totalement imbriquées. Un linguiste qui soulève une hypothèse sur le fonctionnement des chaînes de coréférences, par exemple que le démonstratif ne sera utilisé à l'intérieur d'une chaîne que sous certaines conditions, ou que le changement de paragraphe n'est pas forcément lié à un changement de référent, mais correspond au contraire souvent à une continuation sur le même référent avec un changement de point de vue (dans un genre textuel donné), doit pouvoir se reposer sur des outils de linguistique de corpus pour vérifier cette hypothèse. Si l'hypothèse n'est pas vérifiable, il y a deux causes possibles : soit les outils s'avèrent insuffisants ; soit l'hypothèse est mal formulée.

Dans les deux cas, le linguiste a un travail à faire : aider les concepteurs d'outils à formuler de nouvelles spécifications (cas du projet MC4) ; réfléchir à une autre formulation de son hypothèse (cas de l'étude du texte de Jean Echenoz). À chaque fois, le spécialiste d'outils – et des statistiques sous-jacentes – peut aider le linguiste à s'y retrouver, parmi les hypothèses pouvant amener à des calculs concrets et les hypothèses qui ne pourront que rester sans autre réponse que celle de l'intuition. Dans les deux cas également, les données qui permettent la vérification sont celles d'un corpus annoté. Or, là aussi, linguistes, spécialistes d'outils et spécialistes de TAL ont tout intérêt à collaborer : c'est au moment de spécifier le schéma d'annotation d'un corpus et la procédure d'annotation que tout se joue. Pour qu'un corpus soit exploitable par des techniques d'ADT et de TAL, nous avons vu qu'une première condition résidait dans sa taille. Annoter quelques dizaines ou centaines d'exemples ne suffira pas : l'annotation est un travail de longue haleine, et c'est dans les possibilités d'exploitation ultérieure que réside son intérêt. L'annotation manuelle de corpus peut servir à des systèmes d'apprentissage artificiel et pas seulement à de l'exploration manuelle. Tout projet d'annotation de corpus devrait envisager cette opportunité d'apprentissage, et ce, dès la spécification du schéma d'annotation. Ce qui nécessite d'avoir des connaissances précises sur les besoins d'un système d'apprentissage artificiel comme CROC. En effet, ce n'est que lorsqu'on a une idée des facteurs de décision potentiels que l'on peut décider de la présence de tel ou tel trait dans la structure affectée à une expression référentielle. Les spécialistes de TAL savent quels traits sont automatisables et quels traits ne le sont pas. Face à une proposition de schéma d'annotation, ils peuvent donc évaluer quelle partie est automatisable, quelle partie ne l'est pas, et quels seront les apports du corpus envisagé.

Qui plus est, l'annotation automatique de corpus, même sur des phénomènes plus locaux que des chaînes de coréférences, n'est jamais parfaite, et on peut envisager de concevoir un système qui serve d'aide à l'annotation manuelle. Dans notre cas, ce serait un système qui suggère des coréférences, l'annotateur pouvant accepter ou refuser les suggestions. Une telle approche permet d'envisager des collaborations entre spécialistes d'outils et spécialistes de TAL, afin de proposer les aides à l'annotation les plus efficaces possibles, c'est-à-dire celles qui exploitent au mieux les dernières avancées du TAL et qui augmentent significativement l'ergonomie de l'interface de l'outil, à savoir la nature et l'ordre des manipulations effectuées par l'utilisateur linguiste.

Restent les retours du TAL vers la linguistique, qui sont peut-être plus difficiles à cerner, notamment depuis l'avènement des techniques d'apprentissage. En effet, les représentations internes et les arbres de décision construits automatiquement sont beaucoup moins lisibles que les classiques systèmes de règles, ce qui a une conséquence importante : ils n'ont que peu d'utilité pour les études linguistiques. Bien sûr, le TAL reste une application privilégiée des recherches

linguistiques, mais les termes de cette application ont quelque peu évolué. Au niveau de la référence, on peut remarquer que le TAL n'a pas suivi la voie ouverte par la linguistique : la résolution automatique des anaphores a été laissée un peu de côté, en tout cas en comparaison avec la détection des entités nommées ou d'autres tâches fréquemment mises en avant dans des campagnes d'évaluation internationales. La détection des entités nommées a ainsi drainé toute une communauté autour des typologies d'entités (donc de référents) et des entités structurées (Nouvel *et al.*, 2015). Or ce sont des aspects qui n'ont inspiré que peu de linguistes. Par ailleurs, les modèles d'apprentissage sont de plus en plus globaux, c'est-à-dire qu'ils se nourrissent de l'intégralité des données annotées, et repèrent d'eux-mêmes – avec leurs propres représentations internes (pour ce qui concerne les réseaux neuronaux) – des « classes » qui pourraient correspondre à des pronoms impersonnels, des groupes nominaux, des entités nommées et des formes de référence plus complexes. Les retours que le TAL peut fournir résident donc surtout dans l'aide à la conception d'outils de linguistique de corpus et dans des recommandations pour la conception de schémas et de procédures d'annotation manuelle.

6. Conclusion et perspectives

Plusieurs expériences nous ont amené à faire un tour d'horizon des difficultés rencontrées et des allers-retours envisagés entre linguistes, concepteurs d'outils et spécialistes de TAL. Ce sont ces expériences qui ont été à l'origine du projet DEMOCRAT, qui est en train de proposer un nouveau schéma d'annotation et une nouvelle procédure pour obtenir un corpus annoté en chaînes de coréférences, de taille comparable au corpus ANCOR. Signe de la fréquence des allers-retours pluridisciplinaires, la procédure d'annotation mise en œuvre dans des expérimentations préalables regroupe pas moins de cinq étapes, alternativement manuelles et automatiques : une première étape d'annotation manuelle permet d'exécuter une deuxième étape automatique, qui fournit les données utiles à une troisième étape manuelle et ainsi de suite. Le projet DEMOCRAT matérialise donc les propositions de cet article.

Qu'en est-il par rapport aux autres projets de la communauté ? Nous avons cité dans l'introduction des travaux sur l'espagnol (Recasens, 2010), l'anglais (Lassalle, 2015) et le polonais (Ogrodniczuk *et al.*, 2015). Ces travaux et leurs équivalents ont suivi des approches variées, souvent orientées TAL, et qui laissent de côté de nombreux aspects relevant de la linguistique outillée, notamment la visualisation et l'exploration ergonomique de corpus annotés en chaînes de coréférences. Quand ils s'intéressent à ces aspects, c'est un peu comme dans MC4, à l'aide d'outils d'annotation qui sont adaptés à l'occasion, parfois au prix de difficultés techniques et d'aménagements spécifiques des formats de fichiers (Ogrodniczuk *et al.*, 2015). Mais il n'existe pas d'outil performant pour gérer un corpus annoté en chaînes.

Notre objectif dans DEMOCRAT est d'opérationnaliser notre utilisation d'ANALEC : pour regrouper en une seule plateforme intégrée et cohérente les fonctions d'annotation, de visualisation, d'exploration, de gestion de corpus et d'ADT, nous envisageons l'intégration des fonctionnalités d'ANALEC décrites dans cet article vers la plateforme TXM (Heiden *et al.*, 2010). Comme les articles de ce numéro le montrent, nous envisageons également l'intégration de nouvelles mesures, qui soient adaptées aux chaînes. TXM devrait ainsi s'ajuster pleinement à l'étude des chaînes de coréférences.

Remerciements

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

Références

BILGER M. (éd., 2000), *Corpus. Méthodologie et applications linguistiques*, Paris : Honoré Champion.

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

DESOYER 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* 55(2), 97-121.

FLEURY S. (2013), « Le Trameur. Propositions de description et d'implémentation des objets textométriques », Publication sur le site de l'Université Paris 3. [http://www.tal.univ-paris3.fr/trameur/trameur-propositions-definitions-objets-textometriques.pdf]

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.

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

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

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

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

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

LANDRAGIN F. & SCHNEDECKER C. (éds) (2014), *Langages* n° 195 : *Les chaînes de référence*, Paris : Larousse/Armand Colin.

LANDRAGIN F., TANGUY N. & CHAROLLES M. (2015), « Références aux personnages dans *L'occupation des sols* : apport de la linguistique outillée », *Revue Sciences/Lettres* 3, ENS. [https://rsl.revues.org/].

LASSALLE E. (2015), Structured Learning with Latent Trees: a joint approach to coreference resolution, Thèse de l'Université Paris Diderot.

LEBART L. & SALEM A. (1994), *Statistique textuelle*, Paris : Dunod.

LEFEUVRE A., ANTOINE J.-Y. & SCHANG E. (2014), « Le corpus ANCOR_Centre et son outil de requêtage : application à l'étude de l'accord en genre et en nombre dans les coréférences et anaphores en français parlé », *Actes du quatrième Congrès Mondial de Linguistique Française*, Berlin, SHS Web of Conferences, 2691-2706.

MITKOV R. (2002), *Anaphora Resolution*, London/New York: Longman.

MÜLLER C. & STRUBE M. (2006), "Multi-level annotation of linguistic data with MMAX2", *in* S. Braun, K. Kohn & J. Mukherjee (eds.), *Corpus technology and language pedagogy: New resources, new tools, new methods*, Frankfurt: Peter Lang, 197-214.

NOUVEL D., EHRMANN M. & ROSSET S. (2015), *Les entités nommées pour le traitement automatique des langues*, Londres : Editions ISTE.

OGRODNICZUK M., GŁOWIŃSKA K., KOPEĆ M., SAVARY A. & ZAWISŁAWSKA M. (2015), *Coreference in Polish: Annotation, Resolution and Evaluation*. Berlin: Walter De Gruyter.

POUDAT C. & LANDRAGIN F. (2017), *Explorer un corpus textuel. Méthodes, pratiques, outils*, Louvain-la-Neuve : De Boeck Supérieur.

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

TURENNE N. (2016), *Analyse de données textuelles sous R*, Londres : Editions ISTE.

WIDLÖCHER A. & MATHET Y. (2009), « La plate-forme Glozz : environnement d'annotation et d'exploration de corpus », *Actes de TALN*, Senlis. [http://www.glozz.org]