# Étude des chaînes de référence en français : liens entre modélisation linguistique et analyse quantitative

Frédéric Landragin

# Étude des chaînes de référence en français : liens entre modélisation linguistique et analyse quantitative

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

DRAFT AUTEUR

## 1. Introduction

Une chaîne de référence – ou chaîne de coréférences selon la terminologie utilisée dans le domaine du traitement automatique des langues (TAL) – regroupe toutes les expressions référentielles qui désignent le même référent (Chastain, 1975), que celui-ci soit un individu, un animal, un objet concret, abstrait, voire un événement ou un élément relevant de la deixis textuelle. Quand on en parle dans le cadre d'une chaîne de référence, le terme « expression référentielle » est parfois remplacé par celui plus court de « mention », voire de « chaînon » ou « maillon », une chaîne – physique – comportant par analogie un certain nombre de maillons.

Que ce soit dans une page *web* telle qu'une fiche de Wikipédia ou dans un roman voire un portrait journalistique (Schnedecker, 2005), la chaîne de référence portant sur le sujet principal s'étend potentiellement de la première jusqu'à la dernière phrase du texte, et peut comprendre un grand nombre de maillons. D'autres référents, mentionnés beaucoup moins fréquemment, peuvent faire l'objet d'une chaîne à un seul maillon (auquel cas on parle de « singleton » plutôt que de chaîne), ou de deux maillons – c'est typiquement le cas d'une anaphore pronominale (Cornish, 1999) désignant un référent éphémère. Ces référents éphémères peuvent s'avérer très nombreux : on en trouve potentiellement dans chaque phrase du texte.

Les chaînes de référence constituent ainsi un objet linguistique un peu particulier : elles varient considérablement en taille, en nombre et dans l'étendue de leur intervention (Schnedecker, 1997). Quand de plus on considère que les expressions référentielles possibles sont multiples (Corblin, 1987 ; Charolles, 2002), on en vient à considérer qu'une chaîne de référence peut suivre un nombre considérable de constitutions, depuis la chaîne essentiellement pronominale (« Monsieur A », « il », « il », « il »), très répétitive, jusqu'à la chaîne incluant des redénominations variées (« Un jeune homme », « Monsieur B », « il », « ce garçon », « celui-ci »), plus riche lexicalement. Par ailleurs, les chaînes ne se succèdent pas l'une après l'autre, mais s'entrecroisent en permanence tout au long du texte. Il arrive que plusieurs pronoms « il » se suivent, alors même qu'ils ne font pas partie d'une chaîne de référence unique. Généralement, le lecteur sait démêler les successions alternatives de références à A et B. Il arrive qu'un pronom « il » soit localement ambigu mais, dans la très grande majorité des cas, le lecteur rattache chaque expression référentielle au bon référent, donc à la bonne chaîne.

Ces premières remarques nous conduisent à identifier trois enjeux de modélisation pour les chaînes de référence : (a) l'identification et la délimitation des expressions référentielles ; (b) la caractérisation des successions des expressions référentielles, par exemple des transitions d'un référent à un autre, ou des continuations sur un même référent ; (c) la mise au jour de typologies aussi bien pour ces transitions que pour les expressions référentielles et les chaînes elles-mêmes.

Le recours à des indicateurs calculables – ou mesures – est souvent utile, si ce n'est indispensable. C'est ainsi qu'un nombre de plus en plus élevé de travaux exploitent les notions de longueur de chaîne, de densité référentielle (Schnedecker, 2019 ; Guillot-Barbance & Quignard, 2019) ou de stabilité référentielle (Rousier-Vercruyssen & Landragin, 2019), avec des quantifications associées.

Le recours à un vocabulaire technique – dont le but est d'accompagner le travail de modélisation linguistique – s'avère également utile.

Dans cet article, nous montrons comment les recherches actuelles sur les chaînes de référence suivent un cycle qui commence par la formulation d'une hypothèse linguistique ou d'une notion caractérisant un phénomène référentiel particulier, et se poursuit par l'analyse qualitative ou quantitative correspondante, en faisant appel aux fonctionnalités des outils disponibles, ce qui ne va pas sans difficulté quand on s'intéresse aux structures discursives que sont les chaînes de référence. Tous les outils existants fonctionnant dans d'étroites limites, l'analyste peut enchaîner par la formulation – à destination des concepteurs d'outils – de nouveaux besoins en termes d'appréhension, de visualisation et de mesures calculables sur les données annotées qu'il manipule. Mieux : il arrive qu'en manipulant un mode de visualisation proposé par un outil, émerge une idée de nouvelle mesure ou de nouvelle méthode de visualisation, qui permettra potentiellement d'opérer un retour de l'analyse quantitative vers la modélisation linguistique. Le concept de concordancier, très utilisé en linguistique de corpus, a ainsi conduit à l'implémentation dans la plateforme TXM (Heiden *et al.*, 2010) d'un concordancier adapté aux chaînes de référence, qui sert désormais d'outil de visualisation et d'exploration efficace (Quignard *et al.*, 2018). En outre, il arrive qu'en exprimant un nouveau besoin, l'analyste s'oblige à préciser le sens d'une notion (compétition ou bifurcation, par exemple), voire à reformuler une problématique pour la tester plus facilement par la méthodologie de la linguistique de corpus outillée.

Ce cycle modélisation-corpus-analyses-modélisation a notamment été suivi par les projets MC4 (Landragin, 2015 ; Landragin, 2018) et Democrat (Landragin, 2019 ; Landragin, 2020), projets qui constituent le contexte général de ce travail et dont nous évoquerons les enjeux et bilans. Avant cela, nous présenterons dans la section 2 les grandes lignes de nos modélisations linguistiques, telles qu'elles se sont construites pas à pas, avant toute matérialisation dans un corpus ou dans un projet. Nous présenterons alors, dans la section 3, les questions et problèmes posés par la matérialisation d'une modélisation linguistique dans un programme de recherche incluant des analyses quantitatives. Les projets MC4 et Democrat illustreront cette étape cruciale du cycle de travail. Nous enchaînerons, dans la section 4, avec l'autre étape cruciale du cycle, consistant à reconsidérer voire à reconstruire les modélisations linguistiques après annotation d'un corpus et analyses quantitatives des données annotées. La conclusion nous permettra de dresser un bilan et de poser des jalons pour de futurs projets de recherche sur la référence et les chaînes de référence.

## 2. Plusieurs objets d'étude

Etudier les chaînes de référence et leurs entrecroisements dans un texte nécessite de distinguer plusieurs objets d'étude : (a) tout d'abord, les phénomènes de référence, qui conduisent à s'intéresser aux expressions référentielles et à leur fonctionnement ; (b) ensuite, la suite des expressions référentielles, dans l'ordre dans lequel elles apparaissent dans le texte ; (c) enfin, les chaînes de référence proprement dites, c'est-à-dire les ensemble d'expressions coréférentes.

Une fois ces trois objets d'étude posés, on peut formuler un ensemble de questions qui constituent autant de programmes de recherche. Nous présentons dans cette section quelques-unes de ces questions, et nous abordons ainsi la modélisation linguistique. La manière de poser une question linguistique et de mettre en avant un aspect plutôt qu'un autre relève en effet d'un travail de modélisation, de même que le simple fait de distinguer (a), (b) et (c) – là où d'autres chercheurs auraient par exemple fondé leur modélisation sur la distinction entre référence en première mention et anaphore. Les choix que nous faisons ici ne sont pas anodins, et auront des répercussions sur l'annotation de corpus et la conception d'outils d'exploration.

### 2.1. La référence et les expressions référentielles

Les exemples donnés dans l'introduction avec les référents A et B font intervenir trois types d'expressions référentielles : le nom propre (« Monsieur A », « Monsieur B »), le groupe nominal (« un jeune homme », « ce garçon ») et le pronom (« il », « celui-ci »). Ce sont trois types

d'expressions bien connus, pour lesquels il est relativement aisé de d'affecter ou non le statut d'expression référentielle. Dans la suite de phrases : « Il pleuvait. Monsieur A est entré dans le bar. Il a commandé un café. Il était trop chaud, donc il a attendu avant de le boire », on trouve quatre occurrences du pronom « il » : la première occurrence n'est pas référentielle, alors que les trois suivantes le sont, et réfèrent successivement à Monsieur A, au café, puis à nouveau à Monsieur A. Concernant les groupes nominaux, « le bar » et « un café » sont clairement référentiels. On soulignera au passage que les groupes nominaux définis aussi bien qu'indéfinis sont concernés.

Dans notre petit exemple, seule la forme de surface « il » peut être considérée parfois – mais pas tout le temps – comme un maillon de chaîne de référence. Le même phénomène peut également survenir avec un nom propre et avec un groupe nominal. Dans « Il a pris la rue George Sand », « George Sand » ne constitue pas vraiment une expression référentielle – comme l'est « la rue George Sand ». Même si l'on peut considérer que la mention du nom propre facilite un accès au référent (la romancière), on ne peut que constater que ce référent n'intervient pas dans la structure syntaxique ou thématique de la phrase. « George Sand » n'est pas ici actant du verbe « prendre », et on ne considère pas cette expression comme un maillon, contrairement à la rue qui porte son nom – en tant que référent de type « lieu ». Quant au groupe nominal, comparons « la tête » dans « la tête de Monsieur A est trempée par la pluie » et « Monsieur B m'a pris la tête » : la première occurrence est référentielle, alors que la seconde ne l'est pas, le groupe nominal faisant partie d'une expression plus proche d'une forme figée que de la matérialisation linguistique d'une action de référence.

Il n'existe donc pas de lien direct entre l'expression et sa capacité à référer. C'est le contexte qui permet d'identifier les expressions référentielles, et c'est pourquoi un travail de modélisation portant sur la nature des expressions référentielles relève à la fois de la linguistique et de la pragmatique.

Nous avons évoqué les noms propres, les groupes nominaux et les pronoms. Mais bien d'autres expressions linguistiques peuvent être considérées comme référentielles. C'est le cas des déterminants possessifs. Dans « Monsieur A ouvre son parapluie », on peut considérer que « son » est un maillon de la chaîne relative à Monsieur A. Comme c'était déjà le cas avec « la tête de Monsieur A », ou plus généralement avec tout complément du nom du type « le $N_1$ de le $N_2$ », le groupe nominal « son parapluie » est un exemple d'enchâssement d'une expression référentielle (portant sur Monsieur A, référent de type « humain ») dans une autre expression référentielle (portant sur un parapluie, référent de type « objet concret »). Considérer un déterminant comme un maillon à part entière ne fait pas l'unanimité : nous sommes ici en pleine modélisation des maillons de chaînes de référence, et il va s'agir d'interroger cette modélisation et de la mettre à l'épreuve des corpus.

Encore moins consensuelle est la prise en compte des sujets non exprimés des verbes conjugués, à l'infinitif ou au participe. Dans « Monsieur A entra dans le bar et commanda un café », on peut considérer que le sujet non exprimé du verbe « commander » constitue un maillon de la chaîne de référence relative à Monsieur A. La modélisation emprunte alors sa méthodologie à celle de la syntaxe : selon la prise en compte ou non d'un élément sans forme de surface (mais avec une existence profonde), on retient ou on évacue le sujet « zéro ». L'évacuer présente l'avantage de distinguer l'exemple ci-dessus avec son équivalent où le sujet est exprimé, c'est-à-dire : « Monsieur A entra dans le bar et il commanda un café ». Mais le retenir permet d'aboutir à une modélisation plus complète, surtout si l'on tient compte d'un maillon non exprimé d'une manière différente de celle adoptée pour un maillon exprimé – ce que rend la terminologie « maillon faible » pour le maillon non exprimé *versus* « maillon fort » pour le maillon exprimé (Landragin, 2011).

### 2.2. Suite des références d'un texte

Après cette modélisation des maillons des chaînes de référence, passons au deuxième objet d'étude, à savoir la suite des références d'un texte, et exploitons pour cela un exemple de texte attesté – plutôt que des exemples construits comme c'était le cas dans l'introduction et dans la section précédente. Voici ci-dessous le résumé du film *Alien : le huitième passager* (Ridley Scott, 1979), tel qu'on peut le lire sur le site *web* http://cineclap.free.fr/ (consulté en décembre 2020).

« Bien qu'encore éloignés de leur destination, les sept membres de l'équipage du Nostromo, vaisseau commercial intergalactique, sont tirés de leur sommeil artificiel. Le commandant Dallas explique que "Maman", le super-ordinateur de bord, a détecté un signal inconnu sur Zeta-2-Reticuli. Leur contrat leur impose de partir en expédition. Lors de l'atterrissage, la navette est endommagée. Les réparations sont confiées aux deux machinistes, Parker et Brett. Sur son pupitre, l'officier scientifique Ash reste en contact vidéo avec Dallas, Kane et l'angoissée Lambert, sortis chercher l'origine du signal. Lorsque Ripley s'aperçoit qu'il s'agit d'un avertissement, les explorateurs sont injoignables. Le trio repère et pénètre dans un immense vaisseau aux parois osseuses. Ils trouvent un squelette géant d'une race inconnue, au thorax explosé de l'intérieur. Dans un champ d'œufs, une petite créature bondit sur le casque de Kane et s'agrippe à son visage. Ses compagnons le ramènent précipitamment à la navette. Ash les fait entrer malgré l'ordre de mise en quarantaine donné par Ripley. Après analyse, Ash précise que le parasite, constitué de silicone polarisé et d'un acide ultra-puissant, ne peut être ôté sans arracher le visage de Kane. Peu après, l'organisme se détache de son hôte puis est retrouvé mort. Bien que la navette soit partiellement réparée, l'équipage décolle et regagne le vaisseau en orbite. Kane semble se rétablir quand une bestiole – pondue par le parasite – sort de son ventre en lui explosant le thorax. Après les funérailles spatiales pour Kane, l'équipage se répartit en deux groupes munis d'un détecteur de mouvement, dans l'espoir de jeter l'intrus dans le vide. En cherchant à attraper le chat Jones, qui brouille leurs recherches, Brett est embroché par la créature devenue gigantesque. Témoin, Parker donne l'alerte. Faute de trouver une réponse auprès de "Maman" pour détruire l'indésirable bestiole, Dallas se lance à sa poursuite afin de l'acculer dans le sas d'aération. Parker ne retrouvera que son lance-flamme. Devenue la plus gradée, Ripley interroge le super-ordinateur. Elle découvre avec effroi que le Nostromo a été délibérément détourné pour ramener l'Alien aux Services de la Défense, en sacrifiant les humains. Ash tente alors d'étouffer la jeune femme. Surgissant avec Lambert, Parker fracasse la tête de l'officier scientifique... qui se révèle être un androïde. Celui-ci était chargé de protéger la créature indestructible. Bien qu'ils soient trois, Ripley décide de tenter le coup avec la navette, équipée d'un seul habitacle d'hibernation. Elle lance le compte à rebours d'auto-destruction du vaisseau. Tandis qu'ils rassemblent du fréon, Parker et Lambert sont tués à leur tour. Trouvant l'Alien sur son chemin, Ripley tente en vain d'interrompre la procédure d'auto-destruction. Elle parvient enfin à récupérer Jones, à grimper dans la navette et à quitter le vaisseau avant qu'il n'explose. Alors qu'elle se prépare à hiberner, elle découvre la présence de l'Alien. Enfilant une combinaison spatiale, elle l'éjecte dans le sas et le grille d'un coup de réacteur. L'humaine et l'animal peuvent enfin se plonger dans leur long sommeil de retour... »

Le texte est un peu long, mais riche d'un point de vue référentiel (et coréférentiel). Une première lecture montre une alternance constante entre des noms propres de personnes individuelles, des groupes nominaux tels que « l'équipage », des références à des objets concrets comme le vaisseau spatial et le super-ordinateur ou la navette qui en sont des équipements, ainsi que des références au fameux Alien, l'extraterrestre caractérisé par trois phases de développement : (a) petite créature parasite qui s'agrippe au visage (*facehugger* en version originale) ; (b) bestiole pondue par ce parasite et qui sort du corps de son hôte en explosant le thorax (*chestburster*) ; (c) monstre incarnant le prédateur ultime, appelé parfois xénomorphe (*Alien*). Nous avons là des références à des individus, des groupes d'individus, des objets de natures différentes et dont les relations d'appartenance de l'un à l'autre sont complexes, ainsi qu'un exemple tout à fait saisissant de référent évolutif (Achard-Bayle, 2001 ; Charolles, 2001).

Dès la première phrase, le texte montre toute la complexité d'une modélisation des expressions référentielles : il est question d'un équipage comportant sept membres, autrement dit d'un groupe de personnes dont le cardinal est sept. Sans aucun doute, ce groupe fait l'objet de plusieurs expressions référentielles, mais encore faut-il arriver à délimiter celles-ci. Une modélisation favorisant les expressions complètes (plutôt que les seules têtes nominales) conduit à identifier « les sept membres de l'équipage du Nostromo » et « leur », à la fois dans « leur destination » et « leur sommeil artificiel ». Mais la première expression contient également « l'équipage du Nostromo », expression coréférente bien qu'enchâssée... Une modélisation incluant de plus les sujets zéro de verbes et les phénomènes similaires pourrait même ajouter un maillon « zéro » à « éloignés de leur destination »...

Voici trois questions de recherche qu'il s'agit de trancher dès la première phrase. Les choix ont une importance car, si l'on note A les références à l'équipage, B celle à la destination, C celle au vaisseau nommé Nostromo et D celle au sommeil artificiel, on observe la suite de références « A A B A A C A D » avec le choix le plus inclusif, et « A B A C A D » avec les restrictions les plus fortes. Si les implications relatives à la prégnance du référent A diffèrent, les deux cas permettent de décrire un premier phénomène : la suite des références consiste en l'insertion très nette de référents éphémères (B, C, D) à l'intérieur d'une suite de références portant sur un référent important (A).

Typiquement, une autre suite de références aurait pu ressembler à « A A A B B B C C C », voire à « A B A B A B ». La première se caractérise par des abandons successifs d'un référent au profit d'un autre, et correspond de fait à des chaînes qui se suivent sans s'entrecroiser. Le second cas illustre une compétition durable entre deux référents. Nous aboutissons ici à une première modélisation de la notion de suite de références du texte.

Bien entendu, l'observation de l'exemple montre que les choses ne s'avèrent pas si simples. De nouveaux référents sont introduits dans la deuxième phrase, notamment une référence à l'un des membres de l'équipage et une autre à un mystérieux signal (appelons-le « E »), puis dans la troisième, par exemple la navette dont on comprend qu'il s'agit d'un élément du vaisseau Nostromo. On retrouve ensuite une reprise du signal, d'autres références individuelles à des membres de l'équipage, etc. Il devient très difficile de retrouver l'une des trois suites prototypiques du paragraphe précédent : d'une part E est répété, ce qui l'exclut de la catégorie des « référents éphémères », d'autre part la multiplicité des relations partie-tout complique la modélisation : soit on considère les membres de l'équipage comme $A_1$, $A_2$,... $A_7$ (avec des renvois explicites au référent A « équipage »), soit comme des référents à part entière, sans aucun lien avec le référent A. Dans ce dernier cas, la suite des références du texte devient une succession totalement imprévisible de référents divers et variés. Et les choses se compliquent encore quand le texte décrit les trois phases successives de l'Alien, avec dans un cas la mention explicite d'une métamorphose : « la créature devenue gigantesque ». Doit-on considérer les trois phases comme trois référents distincts, ou comme un référent évolutif ne faisant l'objet que d'une seule chaîne de référence ?

Le cas du référent évolutif est particulièrement intéressant car une modélisation qui sépare toutes les phases d'évolution aboutit à la construction de chaînes de référence distinctes. Par conséquent, l'Alien ne revenant pas de manière réversible à une phase antérieure, on observera alors – si l'on exclut les références aux autres personnages – le patron « A A A B B B C C C ». Au contraire, une modélisation intégrant toutes les phases dans une seule chaîne (« A A A A A A ») permettra de renforcer la primauté du référent Alien dans le texte, ce qui correspond bien au résumé du film éponyme. Là encore, une observation vient nuancer cette remarque : les maillons relatifs aux membres de l'équipage s'avèrent plus nombreux que ceux relatifs à l'Alien. Mais encore faut-il considérer chaque personnage humain comme partie d'un groupe appelé « équipage », mot qui par ailleurs s'avère ambigu dans le texte : comme nous l'avons vu, sa première occurrence réfère à l'équipage du Nostromo. Or, à peu près au milieu du texte, le mot réfère à l'équipage de la navette, c'est-à-dire une partie du précédent. Trois lignes plus loin, « l'équipage » apparaît encore, avec cette fois une référence correspondant à l'équipage du Nostromo – mais dans une nouvelle acception, l'un des membres étant mort entretemps. Le cardinal du groupe est réduit de sept à six, ce n'est donc plus le même groupe. Et les événements décrits dans la suite du texte ne font que répéter ce processus.

Par ailleurs, que faire de l'expression indéfinie « un détecteur de mouvement » ? Deux groupes de personnes en possèdent chacun un. Le problème de distribution conduit à considérer une seule expression référentielle comme faisant partie de deux chaînes de référence distinctes, ce qui pose une question de modélisation : doit-on restreindre l'appartenance d'un maillon à une seule chaîne ? Si l'on autorise qu'une expression relève à la fois de la chaîne A et de la chaîne B, comment modéliser son intervention dans la suite des références du texte ? Par un code spécifique tel que « A/B » ? Nous choisissons ici de conserver la contrainte d'unicité, mais c'est un aspect tout à fait discutable. Autre exemple, faut-il mettre « le compte à rebours d'auto-destruction » et « la procédure d'auto-destruction » dans la même chaîne de référence ? Après tout, le décompte n'est qu'une des facettes de la procédure, donc il n'y a pas coréférence stricte. Or considérer deux chaînes distinctes (soit « A B ») donne l'impression de perdre une information référentielle intéressante. Il faut pourtant choisir, et nous retenons ici l'importance de la coréférence stricte, avec donc deux chaînes distinctes – de fait des singletons.

Avec ce texte d'une quarantaine de lignes dont nous sommes pourtant loin d'avoir étudié l'intégralité des aspects référentiels, nous voulons souligner ici plusieurs problèmes, ou questions, de modélisation. Premièrement, étudier un objet d'étude qui pouvait au premier abord sembler

relativement simple, s'avère en fait tributaire de multiples interprétations linguistiques, parfois très fines, souvent sujettes à débat dans la littérature (Charolles, 2001). Deuxièmement, tenter d'identifier des patrons – ou prototypes – dans les successions de références d'un texte fonctionne mieux si l'on filtre les référents, plutôt que d'en considérer la totalité. Troisièmement, la nature de la suite des références dépend des choix opérés pour l'identification et la délimitation des expressions référentielles, de même qu'elle dépend des choix retenus pour la construction des chaînes de référence.

### 2.3. Chaînes de référence

Avec une modélisation se restreignant aux formes de surface clairement référentielles, la chaîne de référence relative à l'équipage se réduit à quelques maillons, principalement des occurrences de « l'équipage » et du possessif « leur ». De fait, et c'est ce que l'on observe dans la figure 1, il existe deux chaînes de référence : une, que l'on appelle $CR_1$, pour l'équipage du Nostromo, l'autre, appelée $CR_2$, pour celui de sa navette.

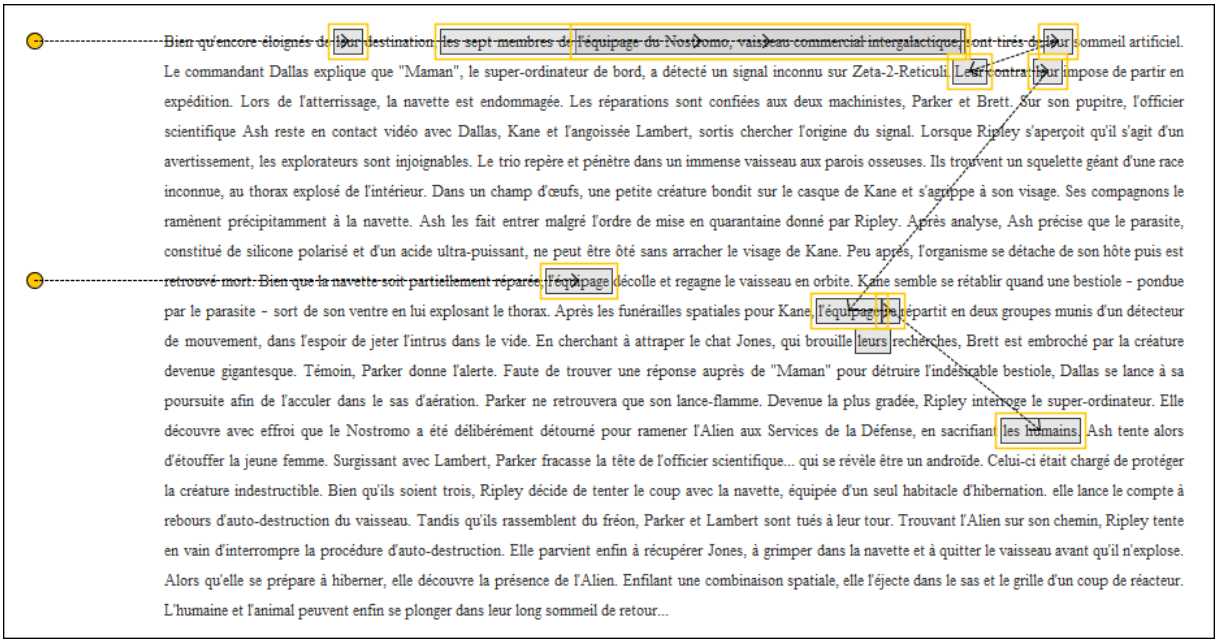


Figure 1 : Visualisation des chaînes de référence relative à l'équipage. Il s'agit d'une copie d'écran issue de l'outil Glozz (Widlöcher & Mathet, 2009), le texte – illisible ici, le but n'étant pas de le lire mais de visualiser la répartition des chaînes – étant exactement celui cité plus haut.

Une modélisation avec deux chaînes s'avère nécessaire lorsque l'on considère des relations de coréférences strictes. L'équipage de la navette n'étant qu'une partie de l'équipage du Nostromo, il n'y a pas identité et deux chaînes distinctes sont donc considérées. Pour être complète, la modélisation doit intégrer une relation d'appartenance du référent de $CR_2$ au référent de $CR_1$. La modélisation visible dans la figure 1 repose cependant sur un choix discutable, celui de ne pas séparer l'équipage avant et après la mort de Kane. Deux arguments s'affrontent ici. Premièrement, le fait que les référents sont différents, puisqu'il s'agit dans un cas d'un groupe de sept personnes, dans l'autre de six personnes. Deuxièmement, le fait qu'on est en présence d'un groupe visant à inclure le maximum de personnes : aucun humain vivant n'est écarté du groupe et, même si celui-ci évolue quelque peu, on peut envisager sa référence comme « l'ensemble des membres vivants de l'équipage ». On observe alors une presque-identité (Recasens *et al.*, 2010) entre le groupe de sept et le groupe de six, ce qui amène à les modéliser par le biais d'une chaîne unique. Un autre argument allant dans ce sens repose sur le contenu lexical des maillons. Un seul nom plein apparaît :

« équipage ». Si le texte avait contenu une expression telle que « le groupe privé de Kane », l'argument n'aurait pas été retenu et la modélisation différente.

Ces considérations peuvent paraître subtiles, voire anodines. Leur discussion s'avère cependant nécessaire, dans la mesure où les choix d'annotation, c'est-à-dire de matérialisation de la modélisation en données analysables, diffèrent nettement. Et la situation se complexifie encore avec l'expression « les humains », qui intervient dans le texte peu après la révélation de la nature robotique de l'un des membres de l'équipage. Dans la figure 1, le choix de modélisation a consisté à maximiser le nombre de maillons de la chaîne relative à l'équipage, et donc d'y inclure l'expression « les humains ». Mais on pourrait objecter que la référence de celle-ci exclut le robot Ash, et correspond donc à un référent de type groupe, inclus dans le groupe « équipage » sans l'atteindre exactement. Encore une question de presque-identité (si l'on considère Ash comme un membre de l'équipage) ou de relation partie-tout (si l'on considère Ash comme un équipement du vaisseau Nostromo, au même titre que le super-ordinateur « Maman »).

A ce stade de la discussion, beaucoup de questions restent ouvertes sur l'identification des expressions référentielles et par conséquent la constitution des chaînes de référence parsemant le texte. Avec un exemple d'apparence banal, nous avons montré que les points de vue sont multiples et conduisent à plusieurs modélisations envisageables. Le but de cet article n'est pas forcément de trancher, mais de montrer en quoi les choix effectués à tel ou tel stade du travail peuvent conduire à des analyses différentes. De fait, plusieurs projets se sont déroulé, avec à chaque fois la nécessaire décision de considérer ou non les sujets zéro comme maillons de chaînes, de considérer ou non la presque-identité comme un facteur de séparation de chaînes, de considérer ou non les indications et révélations (sémantiques) portées par le texte comme des arguments de modélisation des chaînes, de considérer ou non les référents évolutifs comme relevant d'une ou de plusieurs chaînes, et ainsi de suite. Chaque projet apporte son lot d'analyses, qui peuvent dans un second temps nourrir à nouveau les modélisations. C'est ce que nous allons discuter maintenant avec les exemples des projets MC4 et Democrat.

## 3. De la modélisation aux analyses quantitatives

Mettre en place et réglementer au jour le jour un projet collaboratif permet de confronter les points de vue et les arguments, à la fois sur la modélisation et sur la matérialisation de celle-ci sous la forme de recommandations d'annotation d'un corpus. Le projet MC4, « Modélisation contrastive et computationnelle des chaînes de coréférences », a consisté avant tout en discussions regroupant une vingtaine de chercheurs des laboratoires Lattice, LiLPa, ICAR (Landragin, 2018) sur la modélisation des chaînes de référence, avant de se concrétiser en un corpus annoté, corpus prenant spontanément le nom du projet (Landragin, 2015). Le fil directeur du projet consistait à maximiser les annotations manuelles, c'est-à-dire à caser, en quelque sorte, toutes les informations linguistiques potentiellement utiles pour des analyses des chaînes de référence. Le manuel d'annotation ainsi que son versant technique, c'est-à-dire le schéma d'annotation, ont accueilli des considérations morphologiques, morphosyntaxiques, syntaxiques et sémantiques – en plus bien entendu des aspects purement référentiels. La diversité des annotations résultantes a conduit à un corpus protéiforme, incluant des données parfois non comparables, et au final de très petite taille. Le temps de travail des annotateurs restant limité, il n'a pas été possible d'obtenir un corpus à la fois finement annoté et de grande taille. MC4 incarne l'approche expérimentale qui vise à tester le maximum de configurations, mais sans en choisir une seule de manière définitive, pour un passage à l'échelle.

Au contraire, le projet Democrat, « Description et modélisation des chaînes de référence : outils pour l'annotation de corpus (en diachronie et en langues comparées) et le traitement automatique », a procédé au choix inverse, consistant à annoter le plus grand nombre de textes, au prix d'annotations nécessairement minimalistes, et ce malgré la quarantaine de chercheurs participants, issus quasiment des mêmes laboratoires et tutelles que MC4 (Landragin, 2020). Le corpus résultant, même s'il n'atteint pas le million de mots, est ce que l'on peut considérer comme un corpus de grande taille. Il comporte environ 200 000 expressions référentielles délimitées et annotées à la

main, ce qui représente environ 20 000 chaînes de référence. Ces chiffres permettent une multitude d'analyses et d'exploitations, aussi bien linguistiques que statistiques et informatiques, avec dans ce dernier cas la réalisation de deux systèmes de détection automatique de chaînes de référence, utilisables sur du texte tout venant (à partir du moment où il est écrit en français contemporain). Mais, plus que les résultats, ce sont ici les interrogations relatives aux modélisations et aux annotations qui nous intéressent, et sur lesquelles nous allons maintenant nous attarder.

### 3.1. Le corpus MC4 et les outils associés

Le projet MC4 s'est déroulé de 2011 à 2013, avec un financement du CNRS obtenu via l'appel à projets PEPS, « Projets exploratoires premier soutien ». Les principaux résultats ont été publiés dans un numéro thématique de la revue *Langages* (Landragin & Schnedecker, 2014), et le petit corpus annoté par les membres du projet est paru peu après (Landragin, 2015). Nous ne reviendrons pas ici sur ces résultats, ni sur la discussion des avantages et inconvénients des petits corpus en linguistique (Landragin, 2018). Notre objectif est plutôt de montrer en quoi des besoins d'études qualitatives et quantitatives de chaînes de référence ont conduit à mettre en place une première méthodologie d'annotation, que nous appellerons la méthodologie MC4.

Avant le projet MC4, de très nombreux travaux avaient été effectué sur la référence et les chaînes de référence en français (Corblin, 1995 ; Schnedecker, 1997 ; Cornish, 1999). Certaines études visaient à décrire les chaînes observables à l'intérieur d'un genre textuel particulier (Schnedecker, 2005) ; d'autres se focalisaient sur des phénomènes bien choisis de reprise ou d'appel au contexte (Chastain, 1975). Mais il n'existait pas vraiment d'analyse réalisée de manière systématique, sans choisir aucun phénomène ni exemple particulier, faisant appel à des statistiques descriptives « indépendantes » de l'approche linguistique initiale. On trouvait ce type d'approche dans une discipline nommée « statistiques textuelles » (ou « analyse statistique des données textuelles »), qui avait depuis des décennies mis en place une méthodologie très aboutie pour le traitement de textes bruts, non annotés (Lebart & Salem, 1994 ; Landragin & Poudat, 2017), parfois pour le traitement de corpus légèrement annotés (Pincemin, 2004), mais jamais pour celui de corpus annotés en références et en chaînes de référence. Pour faire un premier pas dans ce sens, le projet MC4 a appliqué à l'annotation des maillons et des chaînes la méthodologie de la linguistique de corpus outillée (Habert, 2005 ; Fort, 2012), y compris quand elle s'est intéressée à la coréférence (Van Deemter & Kibble, 2000).

La littérature étant très précise sur les principes à suivre pour obtenir des annotations exploitables, y compris en proposant des métriques pour calculer leur reproductibilité (Fort, 2012 ; Mathet & Widlöcher, 2016), beaucoup de soin a été apporté à la préparation de l'annotation du corpus MC4. Un manuel d'annotation a été rédigé, de manière à indiquer notamment les règles de délimitation des expressions référentielles, ainsi que les différentes valeurs possibles pour chacun des champs (morphosyntaxiques, syntaxiques, sémantiques) constituant une annotation de ces expressions référentielles. Une décision importante a été prise : annoter uniquement les références aux personnages animés.

Au final, chacune des 4 000 expressions référentielles du corpus a été annotée selon onze propriétés, pour un total de 78 étiquettes possibles. Ce sont ces annotations qui ont permis d'effectuer plusieurs analyses sur les constitutions des chaînes, par exemple dans le texte *L'Occupation des sols* de l'écrivain Jean Echenoz (Landragin *et al.*, 2015). Deux constats ont été faits dans les mois qui ont suivi. Premièrement, le nombre de 4 000 expressions annotées est bien trop faible pour en tirer des analyses statistiques significatives, que ce soit la tendance des chaînes à se comporter comme essentiellement pronominales ou avec des redénominations multiples, ou la tendance des premiers maillons de chaîne à prendre la fonction syntaxique de sujet, etc. Deuxièmement, la diversité des exemples entraîne une difficulté à faire émerger des principes ou comportements intéressants. Au vu du corpus MC4, on ne peut pas affirmer qu'une chaîne commence majoritairement par un nom propre ; on ne peut pas affirmer que la chaîne principale d'un texte a des caractéristiques que d'autres chaînes ne possèdent pas, etc.

Finalement, il faut formuler une hypothèse très précise avant d'aller vérifier dans le corpus si cette hypothèse se vérifie ou non. Pour rebondir sur les trois objets d'étude définis dans la section précédente, nous allons détailler un exemple de chacun d'entre eux.

Concernant les références, tous les phénomènes envisagés ont été observés dans le corpus MC4, y compris des exemples d'ambiguïté référentielle, de référence générique ou de regroupement de type partie-tout. L'analyse consiste alors à réaliser des décomptes et des pourcentages, de manière à mieux caractériser les données présentes dans le corpus, et à légitimer la modélisation qui en est à l'origine. L'analyse peut ensuite prendre des chemins qui n'avaient pas été imaginés mais qui sont rendus possibles par l'existence même des données. Ainsi, suite à l'observation que les quinze paragraphes composant *L'Occupation des sols* alternaient narration et description, nous avons formulé une hypothèse sur la fréquence des expressions référentielles : supposées fréquentes dans les paragraphes narratifs, beaucoup moins dans les paragraphes descriptifs. Un traitement spécifique des données permet d'obtenir la « densité référentielle » de chaque paragraphe, c'est-à-dire le nombre d'expressions référentielles divisé par le nombre de mots. On voit ainsi dans la figure 2 que le onzième paragraphe contient une forte densité de références à des humains, alors que les deux paragraphes qui l'encadrent n'en contiennent quasiment pas. Vérification faite, les paragraphes 10 et 12 sont effectivement descriptifs, alors que le paragraphe 11 est narratif. L'analyse permet de valider numériquement une hypothèse. Néanmoins, ce graphique atteint vite ses limites : premièrement, il ne montre pas la longueur des paragraphes (or les paragraphes 10 et 12 sont très courts, ce qui modère quelque peu la significativité de la densité référentielle qui y est mesurée) ; deuxièmement, il reste dans le carcan de la modélisation initiale, qui filtrait les expressions référentielles pour ne retenir que celles portant sur des référents humains. Forcément, on constate des disparités dans le texte, mais celles-ci s'expliquent par la présence de références plus ou moins nombreuses à des référents non humains. À ce stade, il nous faudrait soit recommencer une annotation – et donc une modélisation – qui tienne compte des référents non humains, soit mettre de côté l'ensemble des analyses faisant intervenir la notion de densité référentielle, dans la mesure où celle-ci n'a une validité que partielle.

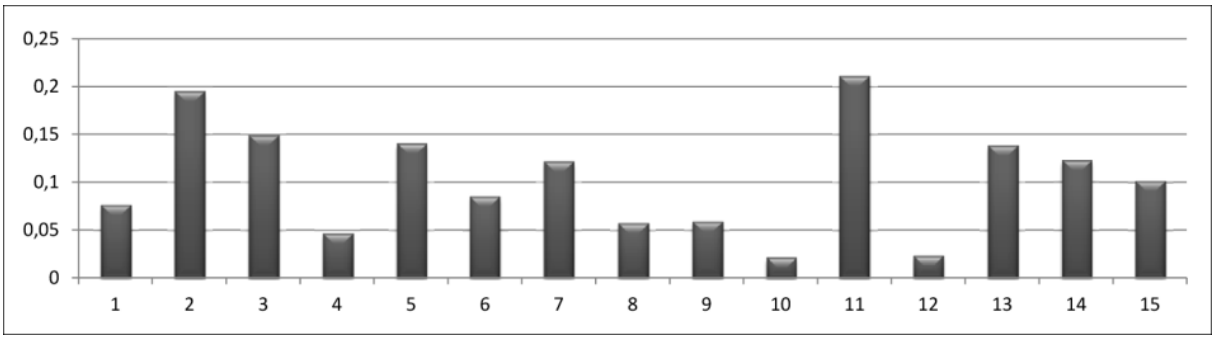


Figure 2 : Visualisation des densités référentielles d'un texte, paragraphe par paragraphe (Landragin *et al.*, 2015).

Passons au deuxième objet d'étude, la suite des références d'un texte. Exploitons les données annotées de la manière suivante : pour chaque expression référentielle, notons 1, 2, 3... les référents (humains) identifiés. Si la première phrase du texte comporte une référence au référent 1, puis une au référent 2 et une dernière de nouveau au référent 1, alors transcrivons cette phrase par la suite « 1 2 1 ». L'ensemble du texte, ou du moins ses références, peut se transcrire selon une succession de nombres. C'est ce qui apparaît en haut à droite de la figure 3. Nous avons introduit un nombre supplémentaire, « 0 », et nous lui avons affecté une signification particulière : celle d'un changement de paragraphe. Avec cet ajout, montrons à quel point la formulation d'une hypothèse linguistique a des incidences sur le traitement des données annotées. L'hypothèse en question est qu'un paragraphe commence souvent par une référence au dernier référent mentionné dans le paragraphe précédent. Impossible ici de prendre en compte un éventuel changement de point de vue (qui

justifierait un tel comportement référentiel), car ce type d'information ne fait pas partie de la modélisation initiale. Le corpus annoté – et donc ses possibilités d'interrogation et d'exploration – restent dans les limites de cette modélisation. En revanche, il est possible de transcrire l'hypothèse en une requête, consistant à compter le nombre d'occurrences de « x 0 x » (où x est un référent) et à le comparer au nombre d'occurrences de « x 0 y » (où x et y sont deux référents distincts). Pour cela, on fait appel à un petit programme qui identifie tous les trigrammes du texte, c'est-à-dire toutes les suites de trois références successives. Ce programme décompte les occurrences de chaque trigramme identifié, et le résultat apparaît à gauche de la figure. On y constate que le trigramme le plus fréquent est « 1 1 1 », ce qui rend compte de la fréquence des continuations sur le référent numéroté 1, visiblement le référent principal du texte. Surtout, le trigramme qui apparaît au deuxième rang est le fameux « 1 0 1 », ce qui donne un argument de poids à l'hypothèse linguistique.

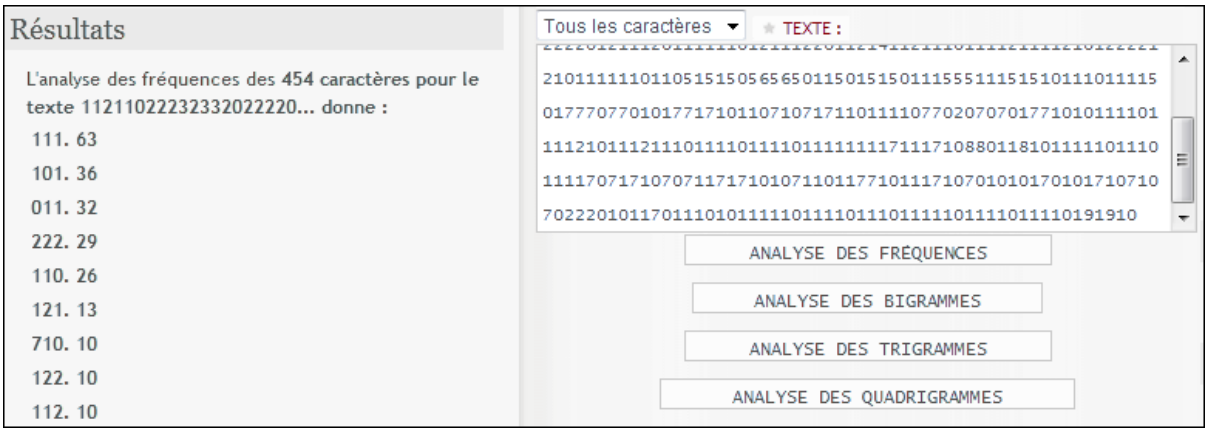

Figure 3 : Visualisation des trigrammes référentiels les plus fréquents.

Certes, le chemin emprunté pour aboutir à cet argument quantitatif n'est pas simple, et nécessite une adaptation des données annotées à l'aide d'un voire de plusieurs outils informatiques. Il illustre en tout cas le travail consistant à traduire en requête informatique une question linguistique, travail qui ne devient opérationnel qu'à partir du moment où nous disposons d'un corpus annoté.

Enfin, abordons l'analyse des chaînes de référence d'un texte. Si l'on reprend le résumé du film *Alien*, l'annotation des expressions référentielles aboutit à un mini-corpus permettant d'appréhender les différentes chaînes qui le composent, en commençant par une ou plusieurs visualisations de ces chaînes. C'est l'objet d'une interface conçue dans le logiciel Analec (Landragin *et al.*, 2012), dans laquelle l'utilisateur peut choisir différents traits d'annotation et visualiser sous la forme d'une succession de points colorés les chaînes d'annotations correspondantes. Dans la figure 4, nous visualisons ainsi la composition des chaînes de référence relative à Ripley et à Kane. Comme nous l'avions rapidement constaté lors de la lecture de ce texte, le nom propre est de rigueur. Il apparaît non seulement dans le premier maillon, mais aussi dans le second, voire le troisième (du moins si l'on néglige l'importance référentielle du pronom réflexif, deuxième maillon de la chaîne relative à Ripley).

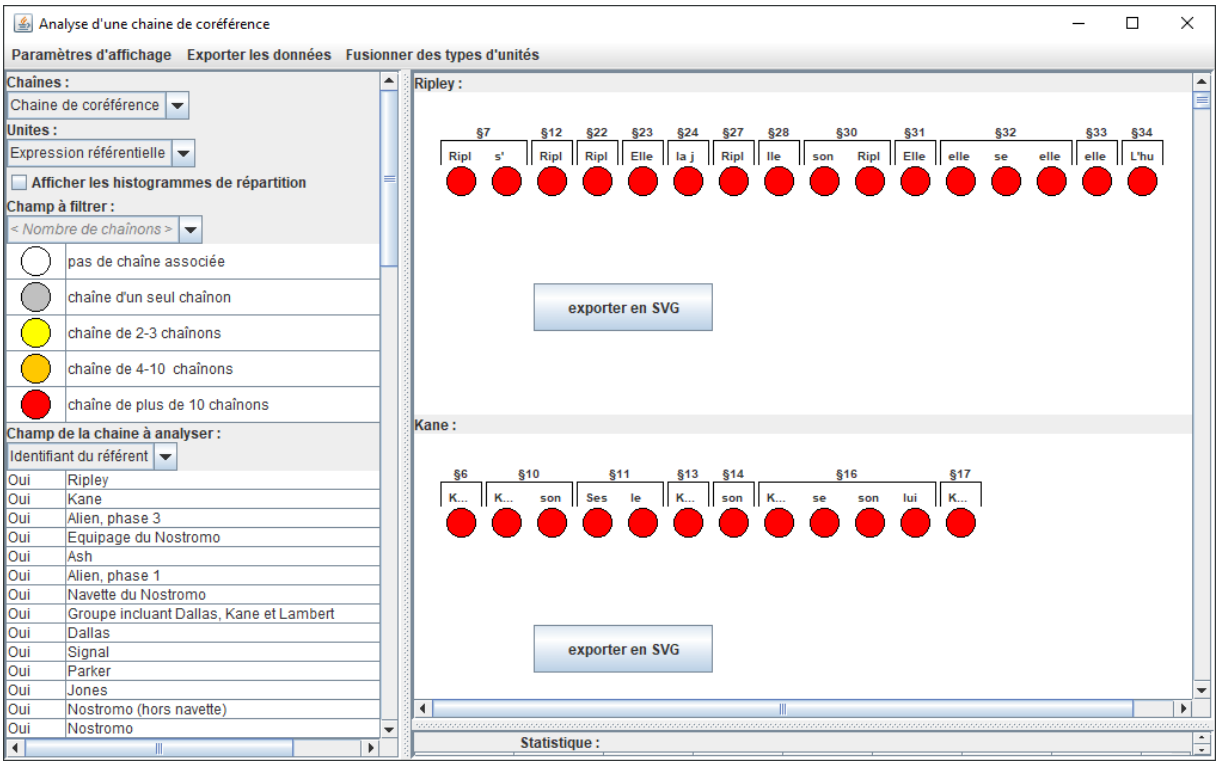


Figure 4 : Visualisation des chaînes de référence relatives à Ripley et à Kane dans le résumé du film *Alien*.

Certes, l'observation ne révèle rien d'exceptionnel. Il suffit quasiment de relire le texte en se focalisant sur l'un des personnages pour appréhender la chaîne de référence qui le caractérise. Néanmoins, cette interface extrait clairement les maillons, leur affecte éventuellement une couleur (pour mieux en repérer certains plutôt que d'autres), permet de comparer deux chaînes – en les affichant en parallèle, comme ici – et a de fait servi de prototype pour l'analyse des chaînes, avant les innovations apportées par le projet Democrat.

### 3.2. Le corpus Democrat et les outils associés

Le projet Democrat s'est déroulé de 2015 à 2020, avec un financement de l'Agence Nationale de la Recherche. Les premiers résultats ont été publiés dans un numéro thématique de la revue *Langue Française* (Schnedecker *et al.*, 2017) ainsi que dans d'autres revues, par exemple *Discours* (Guillot-Barbance & Quignard, 2019 ; Rousier-Vercruyssen & Landragin, 2019). Le corpus annoté par les membres du projet est paru un an avant la fin du projet (Landragin, 2019). Les résultats linguistiques, méthodologiques et relevant du domaine du TAL sont décrits de manière synthétique dans le rapport final du projet (Landragin, 2020).

Par rapport au corpus MC4, le corpus Democrat opère un passage à l'échelle, avec une multiplication des annotations par un facteur de 50. Tous les référents ont été considérés, pas seulement les humains, de manière à compenser par exemple l'analyse toute relative des densités référentielles dans *L'Occupation des sols*. Comme le travail d'annotation manuelle est pour le moins chronophage, il a été décidé de n'annoter aucune information morphosyntaxique, syntaxique ou sémantique : seules comptent la délimitation de l'expression référentielle et l'identification du référent.

Par ailleurs, les efforts ont porté sur la constitution d'un corpus suivant une répartition à peu près équilibrée entre genres textuels (narratifs *versus* non narratifs) et états de langue (français médiéval *versus* français moderne et contemporain). La constitution s'est faite en identifiant des extraits de texte d'environ 10 000 mots chacun, taille choisie selon différents critères : adéquation par rapport à la diversité des phénomènes attendus, par rapport à la longueur souhaitée des chaînes,

mais aussi par rapport aux extraits exploités dans d'autres projets, de manière à favoriser les compatibilités.

L'annotation a suivi l'approche désormais usuelle en linguistique de corpus outillée (Fort, 2012) : (a) décisions collectives concernant les phénomènes linguistiques à annoter, ainsi que les consignes pour ce faire ; (b) écriture collective d'un manuel d'annotation, dont la faisabilité est vérifiée à l'aide de plusieurs expérimentations chronométrées ; (c) annotation proprement dite, soit purement manuelle, soit aidée par l'exécution de macros spécifiques permettant de repérer les expressions susceptibles d'être annotées (repérage des pronoms, par exemple) et d'identifier rapidement des erreurs d'annotation (doublons, par exemple) ; (d) évaluation de la reproductibilité des annotations réalisées, en procédant à une double annotation de 10% du corpus et un calcul de l'accord inter-annotateurs ; (e) homogénéisation des annotations avant : (f) publication du corpus.

Le corpus Democrat, tel qu'il a été conçu, représente un pendant pour le français écrit du corpus d'oral transcrit ANCOR (Muzerelle *et al.*, 2014), dans lequel les annotations portent toutefois sur les relations anaphoriques plutôt que sur les coréférences. Plus que sur la nature des annotations, la comparaison avec ANCOR repose sur le nombre d'annotations – environ 125 000 anaphores annotées pour ce dernier.

Les corpus Democrat et ANCOR, de par leur grande taille, permettent de nombreuses applications, et notamment des applications de TAL. Depuis quelques années, des compétitions internationales sont organisées autour de la détection automatique des chaînes de référence. Les participants développent désormais des techniques d'apprentissage artificiel. Or ces techniques se nourrissent de corpus annotés, et des corpus comme Democrat et ANCOR deviennent extrêmement utiles pour le TAL. Jusqu'à présent, les grandes campagnes d'évaluation internationales telles que SemEval et CoNLL se sont focalisées sur la langue anglaise (Lassalle, 2015 ; Clark & Manning, 2016 ; Lee *et al.*, 2017), sur l'espagnol (Recasens, 2010) ou le polonais (Ogrodniczuk *et al.*, 2015). Le projet Democrat contribue à inciter ces compétitions à intégrer la langue française, ce que le projet MC4 ne pouvait absolument pas faire. Nous ne nous étendrons cependant pas plus sur ces aspects TAL, dans la mesure où – pour le moment du moins – ils n'ont pas permis de retour significatif sur les travaux de modélisation linguistique.

Or Democrat repose sur encore une autre ambition. Dans le domaine de la linguistique de corpus outillée, il existe de multiples outils permettant la gestion, l'annotation et l'interrogation de corpus (Fort, 2012), mais aucun qui permette à l'utilisateur linguiste d'appréhender des chaînes de référence dans leur globalité, ni qui lui fournisse des outils adaptés – statistiques descriptives, graphiques – pour les interroger.

Un premier pas avait été fait avec l'outil Glozz (Widlöcher & Mathet, 2009) dont le projet Democrat s'est largement inspiré, notamment au niveau de la structuration des annotations selon le modèle URS (unité, relation, schéma). Le corpus ANCOR cité plus haut avait justement été annoté à l'aide de Glozz. Mais cet outil présente un inconvénient majeur, à savoir la nécessité de découper le corpus en petits extraits, dans la mesure où il s'exécute difficilement sur des extraits de grande taille.

Un deuxième pas a été fait avec l'outil Analec qui a permis de générer la figure 4 évoquée plus haut. Analec présente lui aussi des inconvénients techniques, en premier lieu sa limitation à la gestion simultanée d'un seul texte – et non d'un corpus entier.

Un enjeu technique important était d'intégrer ces premiers pas dans une plateforme performante pour la gestion de corpus, et le choix de Democrat s'est porté sur la plateforme TXM (Heiden *et al.*, 2010). Le projet Democrat a apporté à TXM les fonctionnalités d'annotation qui lui manquaient. Mieux, il a permis de tester et de mettre en œuvre dans TXM de nouvelles possibilités de visualisation des chaînes de référence. C'est un point essentiel, qui cherche à satisfaire des besoins exprimés par les linguistes étudiant les chaînes de référence, et nous allons en montrer maintenant quelques exemples.

La figure 5, copie d'écran de l'outil Glozz, montre les chaînes de référence du résumé du film *Alien*. L'annotation a suivi les principes de Democrat, donc tous les référents sont annotés, pas

seulement les personnages animés. La conséquence apparaît immédiatement : l'affichage simultané de l'ensemble des annotations conduit à un graphique totalement illisible. Pour les besoins de l'exploration de corpus et donc de l'analyse, l'utilisateur doit pouvoir filtrer, et c'est ce qu'illustre, sur le même texte, la figure 6. On y voit de manière bien plus lisible toutes les expressions référant au fameux Alien. Comme il s'agit d'un référent évolutif, ces expressions sont regroupées selon plusieurs chaînes de référence, ancrées dans la marge gauche par un petit rond coloré qui sert – entre autres – à la sélection d'une chaîne pour en visualiser les annotations.

Figure 5 : Visualisation de toutes les expressions référentielles et de toutes les chaînes.

Ces visualisations effectuées dans l'outil Glozz ne peuvent pas être éditées, ni copiées dans un traitement de texte – ou alors sous la forme d'une image non modifiable. En complément, l'outil Analec propose son propre traitement de texte. La figure 7 montre ainsi le résumé de film dans une version éditable. Afin de tirer parti des italiques, gras, soulignements et couleurs variées, une feuille de style a été utilisée. Elle est totalement paramétrable par l'utilisateur. Ici, le soulignement est réservé aux expressions désignant Ripley, l'italique aux différents équipages et les caractères gras à l'Alien. Il devient ainsi possible de copier-coller des exemples annotés, par exemple pour illustrer un

article de recherche, ou encore pour une exploitation ultérieure par un outil de mise en forme ou d'exploration de texte.

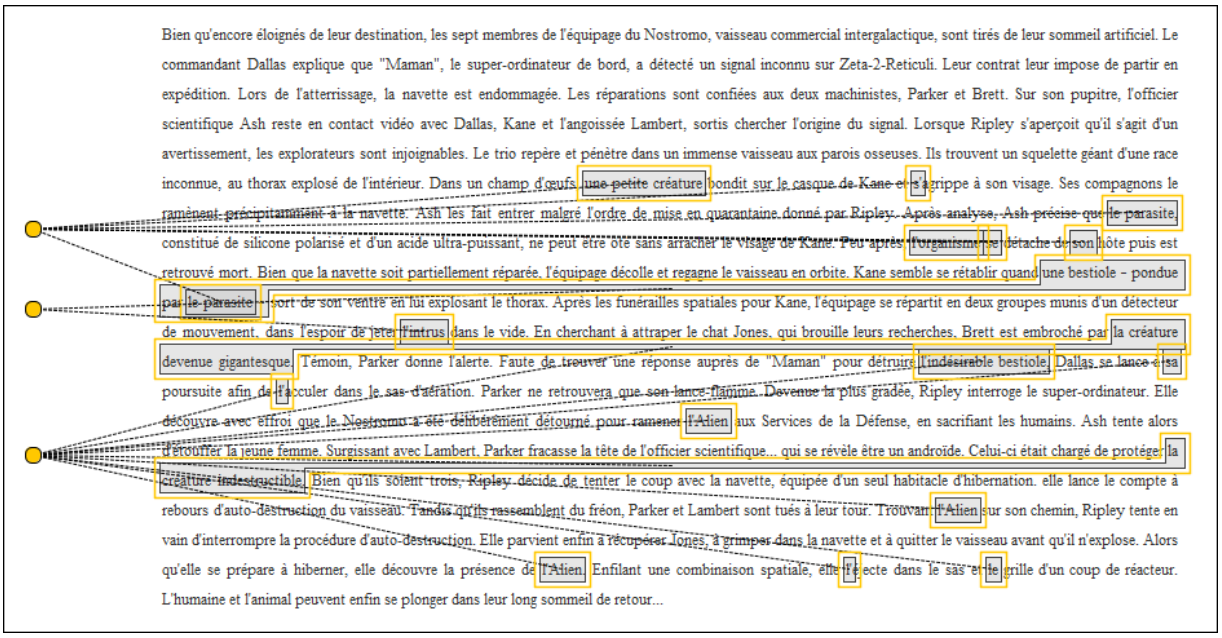


Figure 6 : Visualisation des chaînes relatives à l'*Alien*, en trois sous-chaînes, une par phase d'évolution de l'extraterrestre. Se reporter également à la figure 1 pour une visualisation similaire.

En étendant le panel de fonctionnalités de la plateforme TXM, le projet Democrat a fait émerger plusieurs autres manières de visualiser les annotations en chaînes de référence. Parmi elles, citons les diagrammes de progression. Il s'agit de courbes croissantes, calculées pour montrer l'évolution – au fil du texte – des références. À chaque chaîne de référence est attribuée une courbe, qui augmente d'un cran à chaque fois qu'un maillon apparaît dans le texte. Toutes les courbes commencent ainsi au point d'origine du graphique, en bas à gauche. Plus une chaîne comporte de maillons, plus sa courbe croît. Les pentes élevées caractérisent des parties de texte très riches en maillons, et les pentes nulles (ou « plateaux ») les parties dépourvues de maillons – pour la chaîne considérée. Ce graphique permet d'appréhender d'un coup d'œil si telle ou telle chaîne a la propriété de couvrir l'intégralité du texte ou seulement un extrait, si telle ou telle chaîne présente beaucoup de plateaux (auquel cas le référent disparaît pendant plusieurs parties du texte, et peut être qualifié de « ponctuel »), si telle ou telle chaîne a une pente constante (auquel cas le référent est mentionné très régulièrement, dans tout le texte, et peut donc être qualifié de « régulier »), et ainsi de suite. Les possibilités d'interprétation s'avèrent nombreuses et les diagrammes de progression une bonne synthèse visuelle de la répartition des références et des chaînes dans un texte.

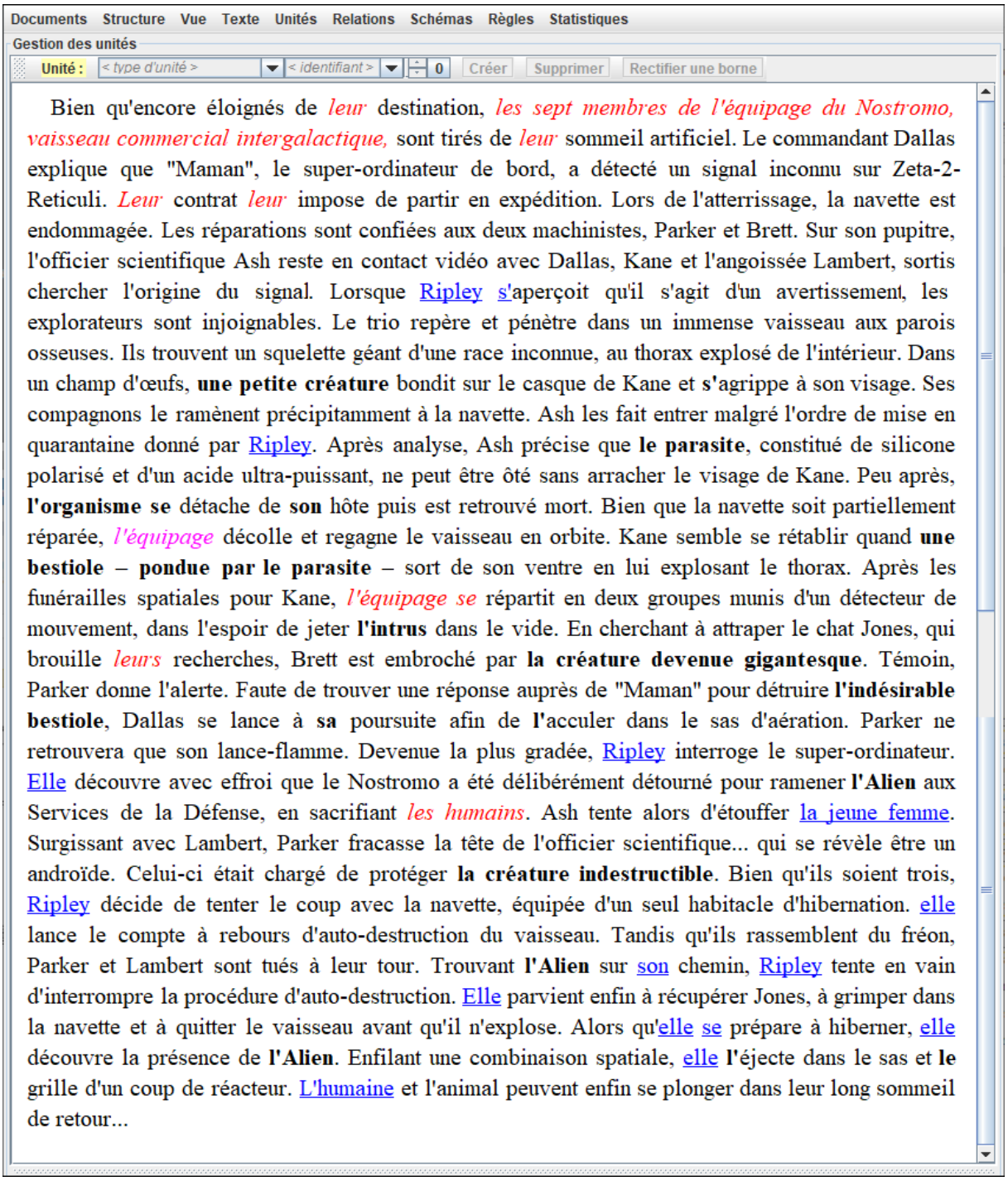

Figure 7 : Chaînes stylisées dans le traitement de texte de l'outil Analec.

Comme tout moyen graphique de visualisation, ces diagrammes, de même que les autres solutions graphiques présentées ici, souffrent de limitations. Pour ne prendre qu'un exemple, on ne peut pas visualiser à la fois la répartition des chaînes et leur constitution en termes de catégories d'expressions référentielles. Par ailleurs, une phase d'apprentissage s'avère nécessaire avant de pouvoir utiliser efficacement l'ensemble de ces outils. Et ceux-ci peuvent procurer, après l'enthousiasme initial, une certaine déception face à leurs limites. Retenons néanmoins que le projet Democrat a ouvert la voie à de nouvelles méthodes pour appréhender et analyser les chaînes de référence. Seul l'avenir nous dira lesquelles, parmi ces méthodes, deviendront incontournables, au même titre que le classique décompte du nombre de maillons.

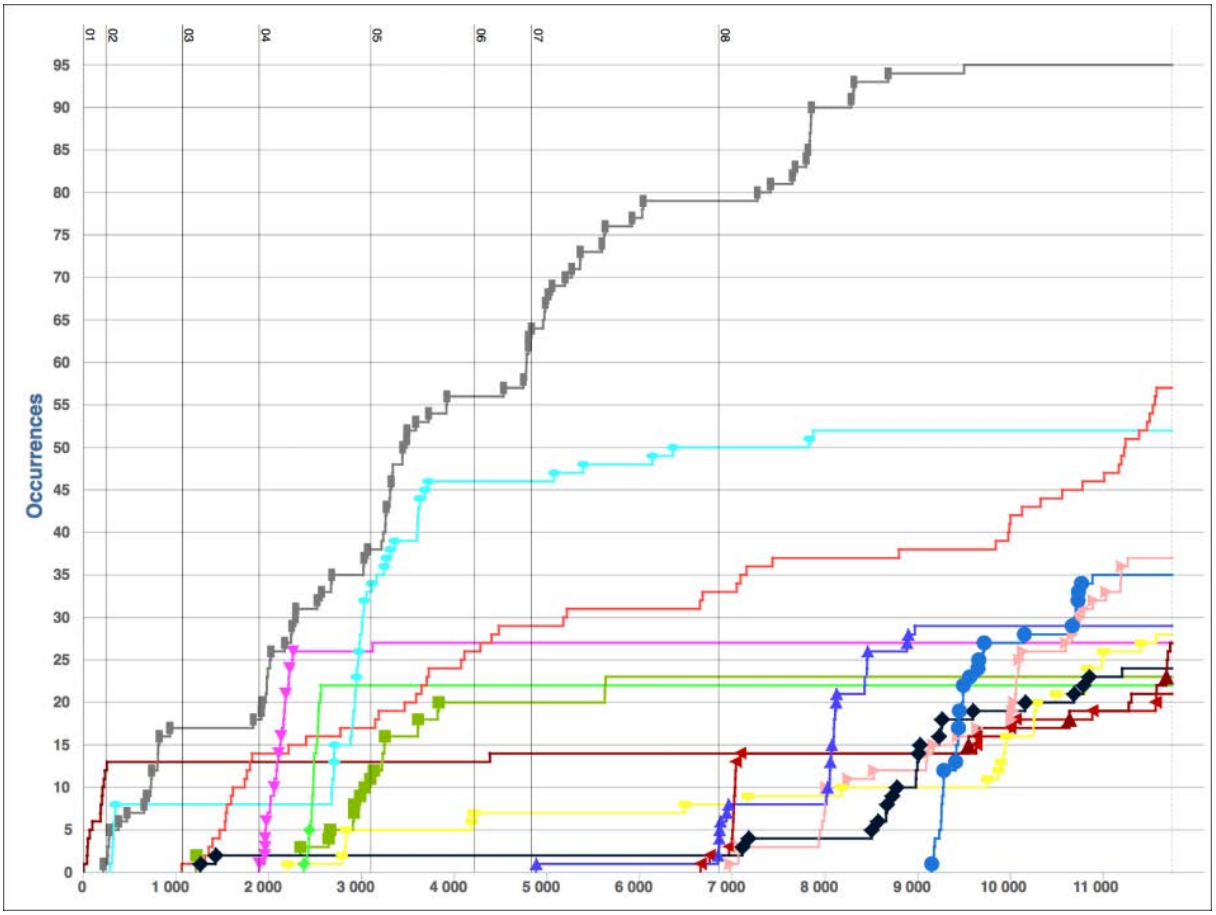

Figure 8 : Visualisation d'un diagramme de progression dans l'outil TXM (Quignard *et al.*, 2018).

Terminons cette section par une comparaison rapide des projets MC4 et Democrat. Dans les deux cas, une grande partie des efforts fournis a consisté en l'annotation manuelle d'un corpus, qui demeure la tâche la plus chronophage, donc celle requérant la meilleure préparation possible, avec la méthodologie la plus affûtée. Le projet MC4 a mis en œuvre des annotations très complètes et très fines reflétant l'ensemble des préoccupations linguistiques de ses participants. Revers de la médaille, le corpus résultant ne comporte que 4 000 expressions référentielles annotées. Le projet Democrat a choisi de ne retenir que les aspects essentiels. Les annotations de son corpus, réduites, ne permettent pas d'intégrer aux visualisations et explorations les aspects syntaxiques ou sémantiques retenus dans MC4. Mais les 200 000 expressions référentielles annotées représentent un passage à l'échelle réussi, ouvrant de nombreuses possibilités d'analyses statistiques et d'entraînements de systèmes de TAL. Il ne sera jamais possible d'annoter à la main 200 000 expressions avec le jeu d'annotations de MC4. En revanche, il est désormais possible d'annoter automatiquement du texte tout-venant avec des annotations, certes minimalistes, mais qui ont le mérite d'exister et de permettre, grâce à l'interopérabilité, de multiples croisements avec des analyses syntaxiques automatiques (utilisation d'un parseur) et des analyses sémantiques (détection automatique des rôles thématiques, par exemple). De futurs projets de recherche devraient œuvrer dans ce sens, afin de permettre aux linguistes de disposer de données plus riches et plus nombreuses que celles disponibles actuellement.

## 4. Des analyses quantitatives vers la modélisation

En attendant de futures exploitations des données du corpus Democrat, intéressons-nous aux retours qu'opèrent les nouveaux moyens de visualisation et d'exploration vers la modélisation

linguistique elle-même. Nous allons voir pourquoi utiliser les interfaces proposées par Glozz, Analec puis TXM, et pourquoi effectuer des calculs de densité référentielle ou de trigrammes fréquents permet d'enrichir la modélisation. Bien entendu, nous discuterons aussi de l'intérêt et du bienfondé de ces retours.

### 4.1. Diversification des modes de visualisation des chaînes

L'objectif d'une interface de visualisation est de proposer un moyen d'appréhender les chaînes de référence autrement qu'en lisant le texte, linéairement. Tous les codes graphiques – taille, couleur, forme, texture, relief, etc. – peuvent être exploités pour ce faire. Dans les figures 1, 5 et 6, les expressions référentielles sont encadrées et les chaînes de référence sont matérialisées par des flèches reliant les expressions coréférentes. C'est un moyen simple, naturel, pour matérialiser une « chaîne », et la plupart des outils l'ont adopté. Dans la figure 7, ce sont les caractères du texte eux-mêmes qui font l'objet d'une colorisation ou d'une mise en valeur. Enfin, dans la figure 4, et encore plus dans la figure 8, l'outil oublie le texte pour ne faire ressortir que les chaînes elles-mêmes, en exploitant les possibilités offertes par ce que l'on appelle des « métaphores graphiques », en l'occurrence des points colorés et des lignes brisées. De fait, on pourrait imaginer bien d'autres façons de rendre graphiquement des chaînes de référence (visualisation 3D, empilements, nuages de points). On pourrait ainsi pallier certaines limites.

Car toute visualisation s'accompagne de contraintes techniques et par conséquent de limites. Par exemple, utiliser des flèches entre expressions coréférentes, en tout cas tel que cela apparaît dans Glozz, ne permet pas de rendre compte visuellement de deux degrés d'appartenance d'une expression à une chaîne. Or c'est justement un aspect de la modélisation, du moins quand elle distingue les expressions référentielles des sujets non exprimés. Autant dire que ces derniers passent à la trappe lors de la visualisation : n'ayant aucune forme de surface, ils restent totalement invisibles. À moins d'encadrer le verbe concerné, mais cette solution entraîne une mise en valeur potentiellement perturbatrice du verbe, que l'on pourrait croire référentiel.

En outre, la présence d'une métaphore graphique peut entraîner un comportement particulier. L'utilisateur peut vouloir agrandir une partie du graphique (« zoom »), faire apparaître des informations complémentaires, changer un code couleur, ou tout simplement revenir au texte lui-même. Proposer un mode de visualisation peut ainsi faire émerger un besoin supplémentaire, qui n'avait pas été identifié au départ par les concepteurs de l'outil. Multiplier les possibilités de visualisation présente un avantage, celui de multiplier les façons de concevoir – cognitivement – une chaîne de référence. Très bien, mais une métaphore graphique n'est jamais exhaustive et peut inciter le linguiste à réclamer d'autres métaphores et d'autres fonctionnalités. En quelque sorte, la boucle n'est jamais bouclée, ce qui peut générer de la frustration.

Tel est le statut d'un outil de visualisation : intéressant au premier abord, vite lassant ensuite, du fait de l'arrêt brutal imposé à l'élan du linguiste qui, entrevoyant un nouvel angle d'approche, cherche à aller au-delà des limites inhérentes à l'outil. Prenons l'exemple du filtrage par l'utilisateur des chaînes à visualiser dans l'interface d'Analec de la figure 4. Conçu pour mettre en parallèle des chaînes de référence, ce filtrage permet de visualiser les chaînes dans un certain ordre, notamment de la plus longue à la moins longue (en nombre de maillons). Ou par ordre alphabétique du nom du référent. Mais c'est tout. L'outil n'a pas été conçu pour que l'utilisateur puisse choisir lui-même un critère d'ordonnancement ou de filtrage, par exemple n'afficher que les chaînes de personnages féminins, ou celles relatives aux seuls objets abstraits, ou à des groupes comportant entre trois et dix humains. Or le linguiste qui explore un texte et en connaît les moindres facettes serait justement tenté de procéder à de tels affichages, qui vont dans le sens de ses recherches. Or le temps requis par le développement informatique, comme celui de l'annotation manuelle, est compté. Surtout, la fonctionnalité supplémentaire pourrait ne s'avérer utile que pour l'étude d'un seul texte, et superflue pour la très grande majorité des cas. Elle serait vouée à disparaître…

Retenons que la diversification des modes de visualisation apporte des possibilités variées pour appréhender cognitivement une chaîne, et pour s'entraîner à repérer des exemples remarquables de chaînes en explorant un corpus comme le corpus Democrat. Dans l'état, les outils restent limités. Ils

permettent surtout de se rendre compte de l'étendue des chaînes (localement dans le texte, tout au long de celui-ci, ponctuellement ou par « rafales »), des entrecroisements entre chaînes, ou encore des densités référentielles, sans même avoir recours à des calculs (la figure 5 en donne une image assez marquante). Ce sont des avancées réelles, qui permettent de remettre en question quelques aspects de la modélisation linguistique, par exemple les grands prototypes de chaînes dont elle a énoncé les caractéristiques.

Mais nous ne disposons pas encore d'outils de visualisation entièrement pilotés par l'utilisateur, où chaque paramètre serait contrôlable et permettrait non seulement un filtrage actif comme en figure 4, mais aussi un filtrage intelligent, tenant compte de la nature des données elles-mêmes et pas uniquement de paramètres universels comme la longueur et le cardinal – seuls paramètres « génériques » envisageables lors de la conception de l'outil.

En attendant une visualisation et une exploration pilotées, le linguiste dispose pour le moment d'un panel d'outils « figés » : à lui de choisir celui qui lui semble le plus adapté face à ses propres besoins et préoccupations de recherche. Une sorte d'équivalent de la sélection naturelle interviendra alors. Les outils les plus utilisés feront l'objet d'une attention croissante de la part des concepteurs, et les autres ne seront plus maintenus et finiront par disparaître, devenus incompatibles avec les nouveaux corpus.

### 4.2. Diversification des mesures et métriques

Nous l'avons vu avec le calcul de la densité référentielle : l'analyse requiert parfois une métrique (ou mesure). Le projet Democrat a exploré et discuté un certain nombre de métriques. Par exemple, le calcul de la distance inter-maillonnaire, ou distance en nombre de mots entre deux expressions coréférentes, a servi à quantifier la modélisation de la répartition des maillons d'une chaîne dans le texte. Quand une dizaine de maillons s'avèrent distants d'environ dix mots, alors que les suivants s'espacent d'une centaine de mots, les chiffres révèlent un comportement référentiel potentiellement intéressant, pour lequel on peut même trouver des qualificatifs : « chaîne ramassée » pour ce qui concerne les dix premiers maillons ; « chaîne diffuse » ensuite.

C'est d'ailleurs en testant l'implémentation de ce calcul de distance inter-maillonnaire dans TXM que la notion de diagramme de progression a été imaginée – preuve s'il en est que le passage au corpus annoté et aux analyses quantitatives permet de progresser dans la conceptualisation des chaînes de référence.

Décomptes, pourcentages, distances, indices de densité et de stabilité : de nombreuses mesures viennent agrémenter les analyses quantitatives des chaînes de référence (Rousier-Vercruyssen & Landragin, 2019). La diversité des mesures vient-elle enrichir la modélisation linguistique ? Sans doute, car elle permet de mettre une notion quantitative sur un phénomène perçu jusque-là de manière purement qualitative, voire intuitive. Ainsi, le coefficient de stabilité, qui décrit la propension des maillons d'une chaîne à répéter – de manière « stable » – les mêmes lexèmes, permet de mieux appréhender ce qu'est une chaîne essentiellement pronominale (forte stabilité) *versus* une chaîne avec des redénominations variées (faible stabilité). En interrogeant les apports des approches quantitatives pour les notions de coréférence et de chaîne de référence, (Schnedecker, 2019) est l'exemple même d'un retour des métriques vers la modélisation linguistique.

Bien entendu, tout chiffre doit être interprété en bonne connaissance des modalités du calcul concerné et des données sur lesquelles celui-ci est effectué. Typiquement, des analyses des corpus MC4 et Democrat montrent que l'ambiguïté référentielle n'y est pas présente (0%). Forcément, puisque les manuels d'annotation des deux corpus imposaient à l'annotateur de faire un choix, de même qu'était imposé le principe consistant à ne faire appartenir une expression référentielle qu'à une seule chaîne. Les textes comportaient très probablement quelques cas d'ambiguïtés référentielles, mais ces cas ont été ignorés par la procédure retenue. Cet exemple un peu caricatural, que l'on peut d'ailleurs considérer comme une critique des manuels et schémas d'annotation inévitablement « simplificateurs », ne remet pas forcément en cause la méthodologie de la linguistique de corpus outillée. Il vise surtout à souligner la prudence qui est de mise lors des analyses. L'ambiguïté référentielle ayant été écartée lors du passage de la modélisation vers sa

matérialisation en corpus, aucun retour vers la modélisation n'est possible à propos de ce phénomène.

Venons-en à la notion de typologie : lorsque l'on étudie les références d'un texte, on en vient rapidement à dresser une typologie des référents. C'est d'ailleurs ce que nous avons implicitement fait quand nous avons évoqué les référents humains et les objets abstraits. De fait, le TAL a produit depuis des décennies un grand nombre de typologies relatives à qu'il appelle les « entités nommées », par exemple les personnes, les organisations, les lieux et les dates. Lorsque le projet Democrat a commencé à réfléchir à une typologie des référents (typologie qui de fait n'a pas été exploitée), il a semblé naturel de se tourner vers ces typologies issues du TAL, et notamment vers celle du projet Quaero (Grouin *et al.*, 2011), telle qu'elle apparaît sur la figure 9.

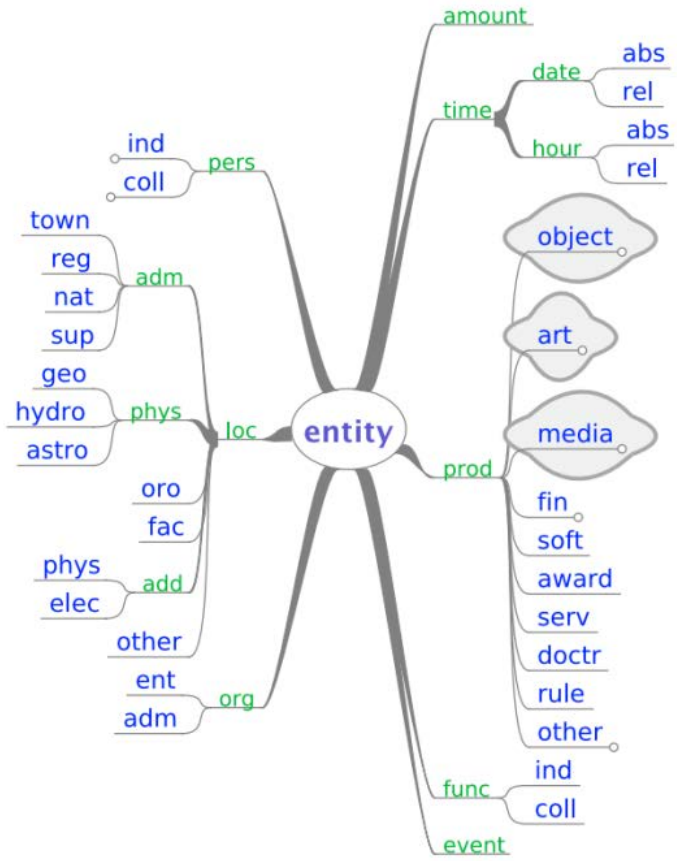


Figure 9 : Typologie proposée par le projet Quaero pour répertorier les différents types de référents (Grouin *et al.*, 2011).

On identifie sur cette figure une quarantaine de types de référents, classés selon une arborescence. Par exemple, les référents de type « personne » se décomposent en deux sous-types, « individuel » et « collectif ». Lorsque l'on annote, on étiquette les chaînes de référence avec un identifiant qui décrit le chemin parcouru dans l'arborescence, par exemple « personne + individuel », « personne + collectif », ou tout simplement « personne » si, dans un souci de simplification, on décide de couper l'arbre au bout d'une seule branche. L'intérêt de maximiser le chemin est de pouvoir orienter les analyses ultérieures vers des requêtes consistant à extraire du corpus toutes les chaînes relatives à des collectifs, par exemple ; et si l'on souhaite toutes les chaînes relatives à des personnes, il suffit de s'arrêter à l'étiquette « personne », sans ajouter de précision. Autrement dit, on peut dans la requête elle-même couper l'arborescence où on le souhaite.

Un autre intérêt est de rééquilibrer des pourcentages favorisant par exemple les personnes aux autres types de référents. Si « personne » apparaît dans 90% des cas et s'oppose à d'autres types,

comme « date » (disons 5% des cas) et « quantité » (2%), alors on peut vouloir mettre sur le même plan « individuel », « collectif », « date » et « quantité » afin que les chiffres rentrent dans une marge de comparaison plus naturelle.

Cette remarque sur les pourcentages a son importance : en annotant un corpus à l'aide d'une dizaine de catégories préalablement identifiées, il arrive parfois que l'une d'entre elle n'apparaisse que dans 1% des cas, voire moins. Il est alors tentant de revoir la modélisation des catégories, et par exemple de fusionner les moins fréquentes en une seule, afin de « gonfler » le pourcentage résultant. Mais il ne faut pas oublier que les phénomènes linguistiques les plus intéressants sont parfois les plus rares, et l'approche fondée sur corpus trouve là ses limites. Autrement dit, autant certains calculs peuvent opérer un retour bénéfique vers la modélisation, autant d'autres peuvent conduire à ignorer des phénomènes pourtant intéressants d'un point de vue théorique, comme l'ambiguïté référentielle ci-dessus.

Le bienfondé de la procédure repose sur un bon choix de métriques, et sur un ensemble de précautions quant à leur interprétation (Landragin & Poudat, 2017).

### 4.3. Chaînes de référence et aspects discursifs

Les diagrammes de progression et autres modes de visualisation des chaînes de référence d'un texte nous semblent importants à plusieurs titres : d'une part pour l'appréhension des chaînes en corpus, avec les aspects cognitifs impliqués (on ne peut pas voir toutes les chaînes en même temps, il faut soit filtrer les informations, soit les organiser en une représentation graphique adaptée) ; d'autre part pour l'étude des relations entre chaînes de référence et aspects discursifs.

En étudiant la suite des références d'un texte aussi bien qu'en interprétant les diagrammes de progression, l'idée est de rendre compte de la progression textuelle, du moins de proposer de nouveaux modes d'analyse qui vont dans ce sens. Nous avons détaillé l'exemple des trigrammes « x 0 x » et « x 0 y » dont les fréquences relatives permettent d'apporter un indice supplémentaire à l'étude de la cohérence d'un texte. Typiquement, c'est ce genre de méthodologie qui peut opérer un retour utile vers la modélisation linguistique, sachant que toute la difficulté réside dans la formulation d'une hypothèse précise, parfois pointue, susceptible d'être transcrite en requête utilisable dans un outil informatique. On aurait pu aller plus loin et étudier également les fréquences relatives des trigrammes « x x x » et « x y z ». On aurait peut-être alors quantifié la continuation sur un même référent *versus* la compétition entre plusieurs référents, et intégrer la notion de bifurcation d'un référent « x » vers un référent « y ». En outre, on aurait pu appliquer ces mêmes outils non pas à la suite des références du texte, mais à une chaîne en particulier, « x », « y » et « z » devenant alors des catégories de maillons, par exemple « nom propre », « groupe nominal » et « pronom ». Des observations potentiellement intéressantes auraient alors pu se faire sur la constitution des chaînes.

En fait, une fois que l'on commence à quantifier, de nombreuses possibilités se font jour, au point que l'on peut s'y perdre quelque peu. L'important est d'explorer les possibilités qui sont le plus susceptibles de faire écho à des préoccupations de modélisation linguistique. Inutile de préciser que le travail à fournir est conséquent.

Le projet Democrat n'a pas exploré toutes les voies ouvertes. De même qu'annoter manuellement un corpus demande des moyens humains importants, mettre au point une méthodologie d'analyse des chaînes de référence en demande également…

## 5. Conclusion et perspectives

Cet article s'est focalisé sur les interactions entre modélisation linguistique et analyses quantitatives en corpus, par le biais de notions telles que la bifurcation ou la progression d'une chaîne dans un texte, et par le biais de métriques telles que la distance inter-maillonnaire et la densité référentielle. Ces notions donnent une idée de ce à quoi pourrait ressembler une méthodologie d'analyse de chaînes de référence. Elles constituent les premiers éléments d'un panel d'outils à la fois conceptuels et informatiques.

Surtout, nous avons montré à quel point il était à la fois souhaitable et délicat de suivre une approche linéaire, comportant les étapes suivantes : (a) identifier l'objet d'étude et l'idée associée ; (b) formuler une hypothèse précise ; (c) déterminer les outils conceptuels et informatiques nécessaires pour confirmer ou infirmer cette hypothèse ; (d) en fonction des analyses obtenues, préciser la signification d'une notion, et lui associer une métrique spécifique ; (e) tenir compte de cette notion dans la modélisation de l'objet d'étude. Ce cycle vertueux s'avère bien difficile à suivre : les notions se précisent au fur et à mesure des recherches, et ne se clarifient parfois qu'après des quantifications ; elles ne se concrétisent pas toutes en hypothèses clairement formulées, ni en métriques calculables, mais reposent au contraire parfois sur une certaine intuition qui résiste à la formalisation ; les métriques et outils présentent des limites qui en réduisent la portée ; les analyses qualitatives aussi bien que quantitatives, pour être réalisables, nécessitent d'écarter des paramètres qui s'avèrent difficiles à réintégrer *a posteriori* (on pensera dans notre cas à l'ambiguïté référentielle), et ainsi de suite.

Les chaînes de référence se révèlent des objets linguistiques difficiles à gérer en corpus, comme l'ont montré les expériences qu'ont été les projets MC4 et Democrat, pour lesquels une approche pluridisciplinaire s'est avérée intéressante, pour ne pas dire indispensable. En l'occurrence, la pluridisciplinarité ne se contente pas de rapprocher linguistique et TAL, mais intègre également la conception d'interfaces humain-machine ergonomiques, la conception de métaphores graphiques aux effets cognitifs maîtrisés, l'analyse statistique de données textuelles (ou textométrie), ainsi que l'ensemble des méthodes de la linguistique de corpus outillée. On comprendra dès lors que les quatre années que dure un projet de recherche tel que Democrat ne suffisent pas à en explorer toutes les facettes, ni même les dix années qui sont déroulées depuis le tout début du projet MC4.

Les perspectives de recherche relatives à cet article se confondent avec celles du projet Democrat, et donc avec les objectifs d'un futur projet collaboratif pluridisciplinaires. À court terme, les perspectives théoriques et de modélisation concernent les suites de trois journées d'étude organisées dans le cadre du projet : (1) clarifier les liens entre chaînes de référence et structures textuelles (et par exemple croiser des annotations des deux sortes) ; (2) multiplier les études contrastives, en comparant les chaînes de référence en français avec les chaînes dans d'autres langues, romanes ou non ; (3) déterminer un ensemble de mesures pour étudier les chaînes de référence de manière claire et rationnelle, quasiment standardisée. Ce dernier point conduit à une autre perspective, cette fois à beaucoup plus long terme : faire un pas significatif vers des statistiques textuelles adaptées aux textes annotés en références et en coréférences. Plus que cela, un enjeu de recherche consiste à intégrer à la méthodologie de la textométrie une prise en compte efficace des annotations et, qui plus est, d'annotations discursives comme le sont les chaînes de référence. Enfin, même si nous n'en avons pas parlé ici, le corpus Democrat regroupe des textes écrits entre le XI$^{e}$ et le XXI$^{e}$ siècles, et relevant de plusieurs genres textuels – pour moitié narratifs, pour moitié techniques, juridiques, encyclopédiques ou journalistiques. Les analyses réalisées dans le cadre du projet peuvent donc jouer avec deux types de variation, diachronique et inter-genres. Une perspective de recherche à long terme est la constitution d'un corpus de grande taille avec des facteurs de variation supplémentaires : inter-langues, productions orales (transcrites) voire issues des nouvelles formes de communication, ou encore productions pathologiques (personnes âgées, Alzheimer, etc.). De nouvelles notions et de nouvelles métriques seront alors à prévoir, entraînant le développement de nouvelles fonctionnalités pour les outils de gestion et d'exploration de corpus, orientant et motivant ainsi les recherches vers plus d'interactions entre modélisation et analyses quantitatives.

## 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 MC4. Il a bénéficié de réflexions et de discussions avec des chercheurs des laboratoires Lattice, LiLPa, ICAR et IHRIM : merci à eux.

## Bibliographie


Achard-Bayle Guy, 2001 : *Grammaire des métamorphoses. Référence, identité, changement, fiction*, Bruxelles : Duculot.

Artstein Ron & Poesio Massimo, 2008 : "Inter-Coder agreement for Computational Linguistics", *Computational Linguistics*, vol. 34 : 555-596.

Bilger Mireille, éd., 2000 : *Corpus. Méthodologie et applications linguistiques*, Paris : Honoré Champion.

Charolles Michel, 2001 : « Référents évolutifs et évolution de la référence ». In : De Mulder Walter & Schnedecker Catherine (éds.), *Les référents évolutifs entre linguistique et philosophie*. Paris : Klincksieck, pp. 39-97.

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

Charolles Michel & Le Goffic Pierre, éds., 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 Charles, 1975 : "Reference and context". In: Gunderson Keith (éd.), *Language, Mind, and Knowledge*. Minneapolis : University of Minnesota Press.

Clark Kevin & Manning Christopher, 2016 : "Improving Coreference Resolution by Learning Entity-Level Distributed Representations", In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, Berlin, Germany, pp. 643-653.

Corblin Francis, 1987 : *Indéfini, défini et démonstratif. Constructions linguistiques de la référence*, Genève et Paris : Droz.

Corblin Francis, 1995 : *Les formes de reprise dans le discours. Anaphores et chaînes de reference*, Rennes : Presses Universitaires de Rennes.

Cornish Francis, 1999 : *Anaphora, Discourse, and Understanding. Evidence from English and French*, New York : Oxford University Press.

Fort Karën, 2012 : Les ressources annotées, un enjeu pour l'analyse de contenu : vers une méthodologie de l'annotation manuelle de corpus. Thèse de doctorat, Université Paris 13.

Godbert Elisabeth & Favre Benoît, 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.

Grouin Cyril, Rosset Sophie, Zweigenbaum Pierre, Fort Karën, Galibert Olivier & Quintard Ludovic, 2011 : « Proposal for an extension of traditional named entities: From guidelines to evaluation, an overview », In: *Proceedings of the Fifth Linguistic Annotation Workshop (LAW 5)*, Portland, Oregon, pp. 92-100.

Guillot-Barbance Céline & Quignard Matthieu, 2019 : « Chaînes de référence et structure textuelle dans les *Essais sur la peinture* de Diderot », *Discours* 25 : 3-26, article publié en ligne à l'adresse suivante : https://journals.openedition.org/discours/10421.

Habert Benoît, 2005 : *Instruments et ressources électroniques pour le français*, Paris : Ophrys.

Heiden Serge, Magué Jean-Philippe & Pincemin Bénédicte, 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 Frédéric, 2011 : « Une procédure d'analyse et d'annotation des chaînes de coréférence dans des textes écrits », *Corpus*, vol. 10 : 61-80.

Landragin Frédéric, éd., 2015 : Corpus MC4, ressource librement disponible sur Ortolang, https://hdl.handle.net/11403/mc4.

Landragin Frédéric, 2018 : « Étude de la référence et de la coréférence : rôles des petits corpus et observations à partir du corpus MC4 », *Corpus*, vol. 18 : 1-20, https://journals.openedition.org/corpus/3422.

Landragin Frédéric, éd., 2019 : Corpus Democrat, ressource librement disponible sur Ortolang, https://hdl.handle.net/11403/democrat.

Landragin Frédéric, 2020 : « Democrat : description et modélisation des chaînes de référence, outils pour l'annotation de corpus (en diachronie et en langues comparées) et le traitement automatique », rapport final du projet, validé par l'ANR, http://fred.landragin.free.fr/fr/publis.htm.

Landragin Frédéric, Poibeau Thierry & Victorri Bernard, 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 Frédéric & Schnedecker Catherine, 2014 : *Les chaînes de référence. Langages, vol. 195*, Paris : Larousse.

Landragin Frédéric, Tanguy Noalig & Charolles Michel, 2015 : « Références aux personnages dans *L'occupation des sols* : apport de la linguistique outillée », *Revue Sciences/Lettres*, n° 3 (Beaucoup de sens en si peu de mots. *L'Occupation des sols* de Jean Echenoz), https://journals.openedition.org/rsl/816.

Lassalle Emmanuel, 2015 : Structured Learning with Latent Trees: a joint approach to coreference resolution, Thèse de l'Université Paris Diderot.

Lebart Ludovic & Salem André, 1994 : *Statistique textuelle*, Paris : Dunod.

Lee Kenton, He Luheng, Lewis Mike & Zettlemoyer Luke, 2017 : "End-to-end Neural Coreference Resolution", In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)*, Copenhagen, Denmark, pp. 188-197.

Mathet Yann & Widlöcher Antoine, 2016 : « Évaluation des annotations : ses principes et ses pièges », *Traitement Automatique des Langues*, vol. 57, n° 2 : 73-98.

Müller Christoph & Strube Michael, 2006 : "Multi-Level Annotation of Linguistic Data with MMAX2". In: Braun Sabine, Kohn Kurt & Mukherjee Joybrato (éds.), *Corpus Technology and Language Pedagogy. New Resources, New Tools, New Methods*. Frankfurt : Peter Lang, pp. 197-214.

Muzerelle Judith, Lefeuvre Anaïs, Schang Emmanuel, Antoine Jean-Yves, Pelletier Aurore, Maurel Denis, Eshkol Iris & Villaneau Jeanne, 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.

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

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

Pincemin Bénédicte, 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 Céline & Landragin Frédéric, 2017 : *Explorer un corpus textuel. Méthodes, pratiques, outils*, Louvain-la-Neuve : De Boeck Supérieur.

Quignard Matthieu, Heiden Serge, Landragin Frédéric & Decorde Matthieu, 2018 : « Textometric Exploitation of Coreference-annotated Corpora with TXM: Methodological Choices and First

Outcomes », In: *Proceedings of the Fourteenth International Conference on the Statistical Analysis of Textual Data (JADT 2018)*, Roma, Italy, pp. 610-615.

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

Recasens Marta, Hovy Eduard & Martı M. Antònia, 2010 : "A Typology of Near-Identity Relations for Coreference (NIDENT)", *Proceedings of the Seventh International Conference on Linguistic Resources and Evaluation (LREC 2010)*, Valletta, Malta, pp. 149-156.

Rousier-Vercruyssen Lucie & Landragin Frédéric, 2019 : « Interdistance et instabilité au sein des chaînes de référence : indices textuels ? », *Discours* 25 : 3-32, article publié en ligne à l'adresse suivante : https://journals.openedition.org/discours/10522.

Schnedecker Catherine, 1997 : *Nom propre et chaînes de référence*, Paris : Klincksieck.

Schnedecker Catherine, 2005 : « Les chaînes de référence dans les portraits journalistiques : éléments de description », *Travaux de linguistique* 51 : 85-133.

Schnedecker Catherine, 2019 : « De l'intérêt de la notion de chaîne de référence par rapport à celles d'anaphore et de coréférence », *Cahiers de praxématique* 72 : 1-19.

Schnedecker Catherine, Glikman Julie & Landragin Frédéric, 2017 : *Les chaînes de reference en corpus. Langue Française, vol. 195*, Paris : Armand Colin.

Van Deemter Kees & Kibble Rodger, 2000 : "On Coreferring: Coreference in MUC and related annotation schemes", *Computational Linguistics*, vol. 26, no 4 : 629-637.

Widlöcher Antoine & Mathet Yann, 2009 : « La plate-forme Glozz : environnement d'annotation et d'exploration de corpus », *Actes de la conférence TALN 2009*, Senlis.