Conception d'un outil de visualisation et d'exploration de chaînes de coréférences

Frédéric Landragin

Conception d'un outil de visualisation et d'exploration de chaînes de coréférences

Frédéric Landragin

Laboratoire Lattice, CNRS, ENS, Université de Paris 3, Université Sorbonne Paris Cité, PSL Research University – Paris/Montrouge – France

Abstract

Being the result of a manual or an automatic annotation procedure, a coreference chain groups a set of referring expressions that all refer to the same entity. A chain can cover the whole text and therefore contributes to its coherence. Each mention may be annotated with linguistic interpretations. Moreover, relations (that can be annotated, too) may exist between mentions. As a consequence, one can find difficult to apprehend and quickly analyze a chain. In this paper, we present a tool for visualizing coreference chains and for detecting relevant phenomena like patterns in referential transitions. We propose a general methodology for coreference analysis, and we illustrate it with the results of preliminary tests using short texts. Then, we discuss the interests of such a tool for text analytics and textometry, and we identify links with existing works that can lead to a future version of our tool.

Résumé

Qu'elle soit le résultat d'une annotation manuelle ou automatique, une chaîne de coréférence est une structure qui regroupe un ensemble d'expressions référentielles désignant toute la même entité. Une chaîne peut s'étendre tout le long d'un texte et contribuer ainsi à sa cohérence. Chaque expression (ou maillon) peut être enrichie d'annotations linguistiques, et les différents maillons d'une chaîne peuvent être reliés par des relations, elles-mêmes annotées avec des interprétations linguistiques. En conséquence, il est difficile d'appréhender cognitivement une telle structure et d'en tirer directement des analyses. Dans cet article, nous présentons un outil pour visualiser des chaînes de coréférences et pour repérer des phénomènes remarquables tels que des motifs dans les transitions référentielles. Nous proposons une méthodologie pour analyser les coréférences et les transitions référentielles, que nous illustrons avec les résultats de tests effectués sur des textes courts. Nous discutons alors les intérêts de cette méthodologie et de notre outil pour le domaine de l'analyse des données textuelles et de la textométrie. Nous identifions notamment un ensemble de liens avec des travaux existants qui pourront permettre d'envisager des perspectives de recherche ainsi qu'une prochaine version de notre outil.

Key words : Referring expressions, coreference, coreference chains visualization, corpus analysis.

1. Introduction

Considérons une succession de phrases comme celle de l'exemple suivant : « Le village était désert. Il semblait abandonné. La place principale était vide. Elle en paraissait triste. Tout reprendrait vie le lendemain matin : le village s'animerait, la place se remplirait de monde ». Ce n'est certes pas un texte littéraire, mais – comme tout texte – une succession de phrases reliées les unes aux autres, caractérisée ainsi par une certaine **cohérence** (Charolles, 1994 ; Legallois, 2006). Cette cohérence passe par les relations anaphoriques et coréférentes qui relient les **expressions référentielles** (Charolles, 2002) parsemant le texte : la deuxième phrase est reliée à la première grâce au pronom anaphorique « il » qui reprend l'expression référentielle « le village ». Même chose pour les deux phrases suivantes avec « la place principale » reprise par « elle ». Entre ces deux paires de phrases, le lien ou « pont »

correspond à celui existant entre un village et sa place principale (méronymie « partie-tout ») : il s'agit d'une relation associative entre deux référents, que l'on peut considérer comme une anaphore un peu particulière entre les expressions référentielles « la place principale » et « le village » (Kleiber, 2001). Les deux référents ne sont pas identiques ; autrement dit cette relation anaphorique n'est pas coréférente. Au contraire, le lien entre « le village » (dernière phrase) et « le village » (première phrase), ainsi que celui entre « la place » (dernière phrase) et « la place principale » (troisième phrase) sont clairement coréférents. Par ailleurs, nous observons aussi l'expression référentielle « tout », qui reprend d'une certaine façon « le village », voire l'ensemble formé par « le village » et « la place principale » : encore un lien tissé entre les différentes phrases de ce petit texte.

Organiser ces liens peut se faire en considérant des **chaînes** (Corblin, 1995 ; Schnedecker, 1997). Par exemple, si l'on suit une à une toutes les relations anaphoriques et que l'on retient à chaque fois les expressions référentielles concernées, on obtient la chaîne {« le village », « il », « la place principale », « elle », « tout », « le village », « la place »}, dont les éléments (ou maillons) sont indiqués selon leur ordre d'apparition. Il s'agit ici de l'ensemble des expressions référentielles du texte, c'est-à-dire de la **suite des références** (Landragin, 2011). Si l'on décide d'ignorer les relations associatives, on n'obtient que les deux chaînes : {« le village », « il »} et {« la place principale », « elle »}, composées d'anaphores pronominales, et dont le prototype est : {« N », « il », « il », « il »…}. Si l'on intéresse aux coréférences, on considère les **chaînes de coréférences** suivantes : {« le village », « il », « le village »}, {« la place principale », « elle », « la place »}, et éventuellement {« le village » + « la place principale », « tout »}. Selon les approches (Schnedecker, 1997 ; Landragin, 2011), il existe de multiples façons d'appréhender les chaînes anaphoriques et/ou coréférentes d'un texte. Pour être exhaustif, on devrait construire également une chaîne regroupant le pronom « en » (quatrième phrase) et son antécédent (à savoir la troisième phrase), sans oublier des chaînes-singletons pour les référents cités une seule fois : le lendemain matin, le monde.

L'enjeu général est de tenir compte de ces chaînes dans l'analyse linguistique (pragmatique) des textes, et notamment dans l'analyse de leur cohérence. Le cadre scientifique de cet article est donc l'étude de la cohérence textuelle et des chaînes de coréférences, afin de caractériser les moyens mis en œuvre par les rédacteurs pour écrire des textes qui fonctionnent et ne ressemblent pas à des juxtapositions décousues de phrases non liées. Les études relevant de ce cadre prennent un nouveau tournant avec les avancées de la linguistique de corpus outillée et de l'analyse des données textuelles. Le cadre technique de cet article est la conception d'outils d'annotation, de visualisation et d'étude des chaînes, notre objectif à terme étant de contribuer à outiller la linguistique de discours. Dans cet article, nous présentons une méthodologie axée sur la suite des références et sur les chaînes de coréférences, ainsi qu'un outil – plus précisément un nouveau module pour le logiciel ANALEC (Landragin *et al.*, 2012) – module développé pour le projet ANR DEMOCRAT (« Description et Modélisation des Chaînes de Référence : outils pour l'Annotation de corpus et le Traitement automatique », 2016-2020).

L'article comprend deux sections principales : la section 2 présente la méthodologie et se focalise sur les représentations graphiques des chaînes ; la section 3 propose quelques pistes pour des statistiques orientées vers l'étude des chaînes. Nous concluons alors sur les perspectives envisagées pour notre outil. Nous prenons comme support d'illustration la nouvelle *Les Bijoux* de Guy de Maupassant (1883), avec notamment ses référents humains, c'est-à-dire les personnages de l'histoire. Nous aurons ainsi une chaîne de coréférences pour le personnage de Mme Lantin, une autre pour Mr Lantin, une troisième pour le groupe formé par Mr et Mme Lantin (de la même façon que ci-dessus avec « tout »), et ainsi de suite.

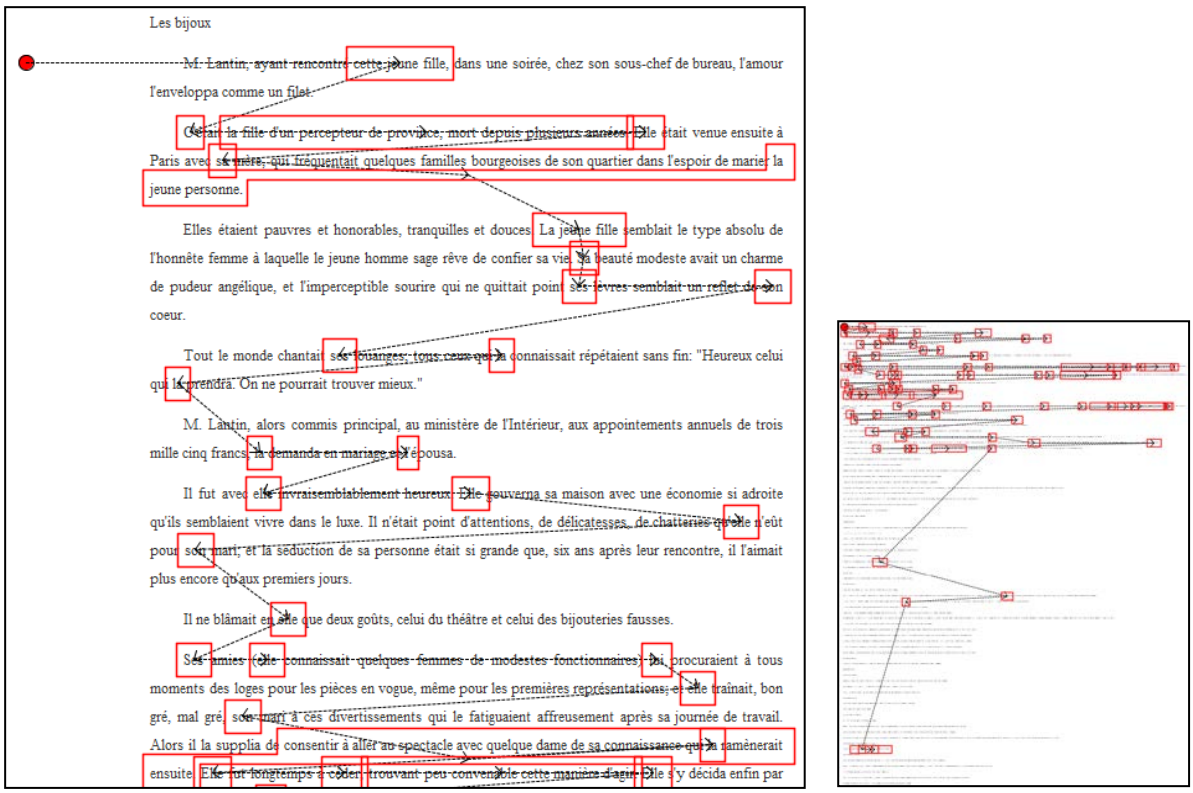

Figure 1 : visualisation d'une chaîne d'annotation dans l'outil GLOZZ (Widlöcher & Mathet, 2009). Il s'agit d'une chaîne de coréférences dans la nouvelle Les Bijoux *de Guy de Maupassant, les autres chaînes de coréférences ayant été supprimées pour ne pas surcharger la visualisation. A gauche se trouve la fenêtre de travail (active) et à droite une visualisation globale (non active) sur le texte.*

2. Annotation et visualisation de chaînes de coréférences

2.1. Contexte et méthodologie

En attendant les avancées du traitement automatique des langues et l'amélioration des logiciels détectant automatiquement les chaînes de coréférences dans des textes écrits en français (Désoyer *et al.*, 2014), la plupart des études réalisées actuellement sur les chaînes de coréférences en français partent de corpus annotés manuellement (Corblin, 1995 ; Schnedecker, 1997 ; Landragin & Schnedecker, 2014).

Des projets comme MC4 (« Modélisation Contrastive et Computationnelle des Chaînes de Coréférences », https://www.ortolang.fr/market/corpora/mc4), ANCOR (« Anaphores et coréférences dans les corpus oraux », http://tln.li.univ-tours.fr/Tln_Corpus_Ancor.html) ou le projet ANR DEMOCRAT déjà cité étudient ainsi les chaînes de coréférences dans des textes et dialogues de genres variés. La méthodologie commence par la délimitation manuelle des expressions référentielles, puis leur regroupement en chaînes. Expressions et chaînes sont annotées (figures 2 et 3) avec des interprétations linguistiques qui varient selon les études : aspects sémantiques et pragmatiques de la résolution de la référence, nature des relations d'anaphore ou de coréférence, etc. L'analyse des données ainsi annotées se confronte à deux enjeux pour lesquels les outils actuels s'avèrent insuffisants : premièrement la visualisation ergonomique des chaînes, sachant qu'une chaîne couvre potentiellement toute la longueur du texte ; deuxièmement l'analyse quantitative de ces chaînes : quels décomptes effectuer ? À quels indicateurs numériques et statistiques faire appel ?

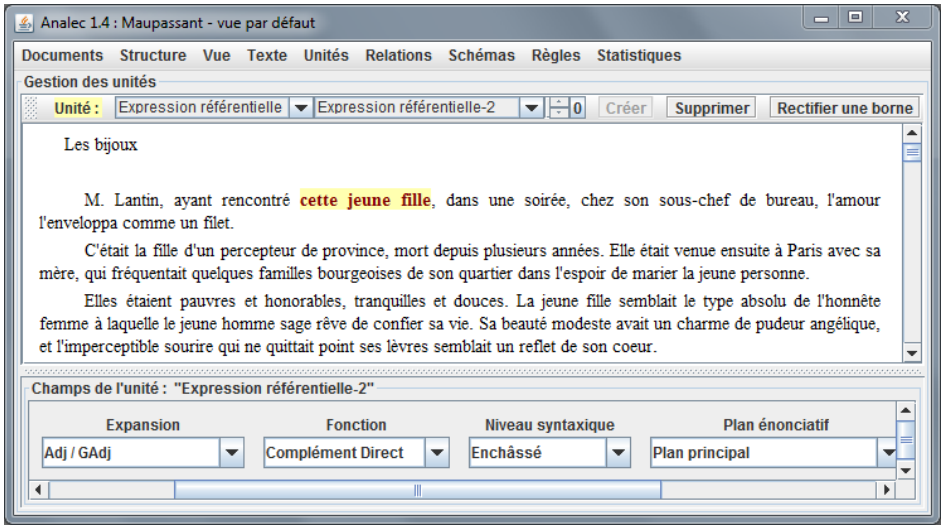

Figure 2 : annotation d'une expression référentielle. Dans cette configuration, quatre propriétés linguistiques sont annotées manuellement, via des menus déroulants comprenant les choix possibles.

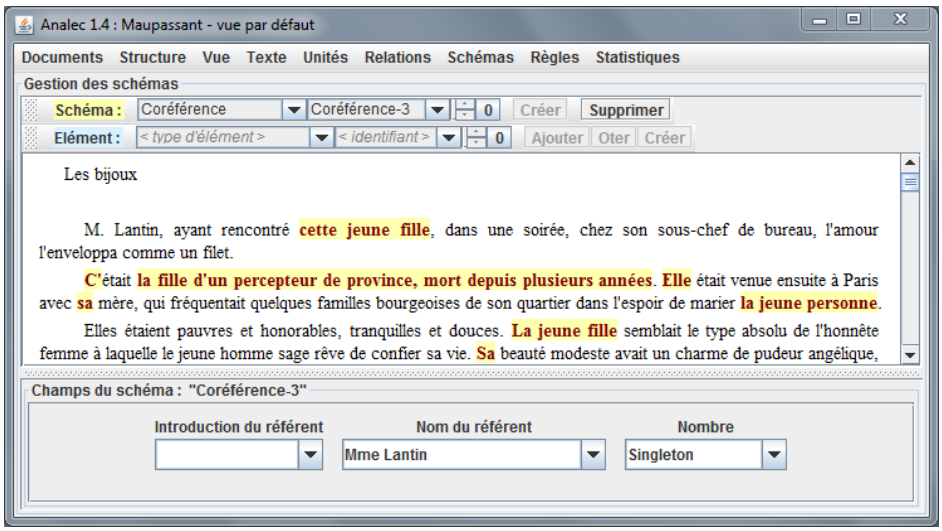

Figure 3 : visualisation et annotation d'une chaîne de coréférences. Ici, les expressions référentielles ont déjà été ajoutées à la chaîne, et l'utilisateur est en train d'annoter celle-ci selon trois propriétés.

(Landragin & Schnedecker, 2014) propose un bilan de quelques études, notamment celles issues du projet MC4. Ce qu'il en ressort, c'est qu'aucun outil de gestion et d'annotation de corpus n'est vraiment adapté à une analyse fine des chaînes de coréférences, c'est-à-dire avec une mise en valeur de considérations sur les liens entre la structure du texte en paragraphes et la répartition des maillons des chaînes, sur les typologies de chaînes (à l'instar du prototype {« N », « il », « il », « il »…} évoqué plus haut), sur la fréquence des maillons ayant une détermination démonstrative (ce qui indique une sorte de recommencement de la chaîne), ou encore sur la nature des successions d'annotations, que ce soit dans le cadre d'une chaîne spécifique ou dans celui de la suite des références du texte. Or ces préoccupations intéressent les linguistes, qui en sont réduits à explorer manuellement leurs corpus.

De fait, il existe une panoplie d'outils de visualisation et d'exploration conçus pour des données morphosyntaxiques, lexicales, syntaxiques (Steiner & Kallmeyer, 2002 ; Lezius, 2002), sémantiques (Venant, 2008), et même discursives : l'outil GLOZZ (Widlöcher & Mathet, 2009) fait partie de cette dernière catégorie et s'avère exemplaire pour l'appréhension

de données couvrant potentiellement la globalité du texte (figure 1). Les outils Annis, MMAX et surtout MMAX 2 (Müller & Strube, 2006 ; Chiarcos *et al.*, 2008) en font aussi partie, mais avec peut-être un approfondissement moindre – comparé à GLOZZ – des problèmes de visualisation. GLOZZ est peut-être l'outil qui propose le plus de solutions d'exploration : le langage de requête GlozzQL a été développé en tenant compte des besoins liés aux chaînes d'annotations, à travers le concept « unités-relations-schémas » à la base de l'outil (Widlöcher & Mathet, 2009). Le logiciel ANALEC sur lequel nous nous appuyons reprend ce concept propice à la modélisation des chaînes. Le module que nous présentons dans cet article (et dont toutes les figures depuis la figure 2 sont des copies d'écran opérationnelles) étend les possibilités de visualisation et d'exploration, sur la base du bilan de (Landragin & Schnedecker, 2014), et donc des préoccupations linguistiques mentionnées ci-dessus.

Notre méthodologie suit trois phases, qui reprennent trois types d'interrogations des linguistes des projets MC4 et DEMOCRAT : la première phase relève de statistiques générales sur le texte et sur la répartition des chaînes dans ce texte ; la deuxième phase relève de l'étude approfondie – visualisation, exploration et calculs statistiques – de la suite des références du texte ; la troisième phase relève de l'étude approfondie – visualisation, exploration et calculs statistiques – des chaînes de coréférences. L'utilisateur est censé suivre ces trois phases dans l'ordre. Celles-ci sont donc proposées dans un menu dédié à l'étude des chaînes d'annotations, et une fenêtre interactive s'ouvre spécifiquement pour chaque phase. Visualisation et statistiques sont regroupées dans chaque fenêtre, dans la mesure où l'on procède à des calculs de fréquences ou de corrélations sur les données que l'on visualise et que l'on sélectionne. Nous allons pour l'instant nous focaliser sur les aspects graphiques et interactifs, avant de revenir, dans la section 3, sur les calculs statistiques proprement dits.

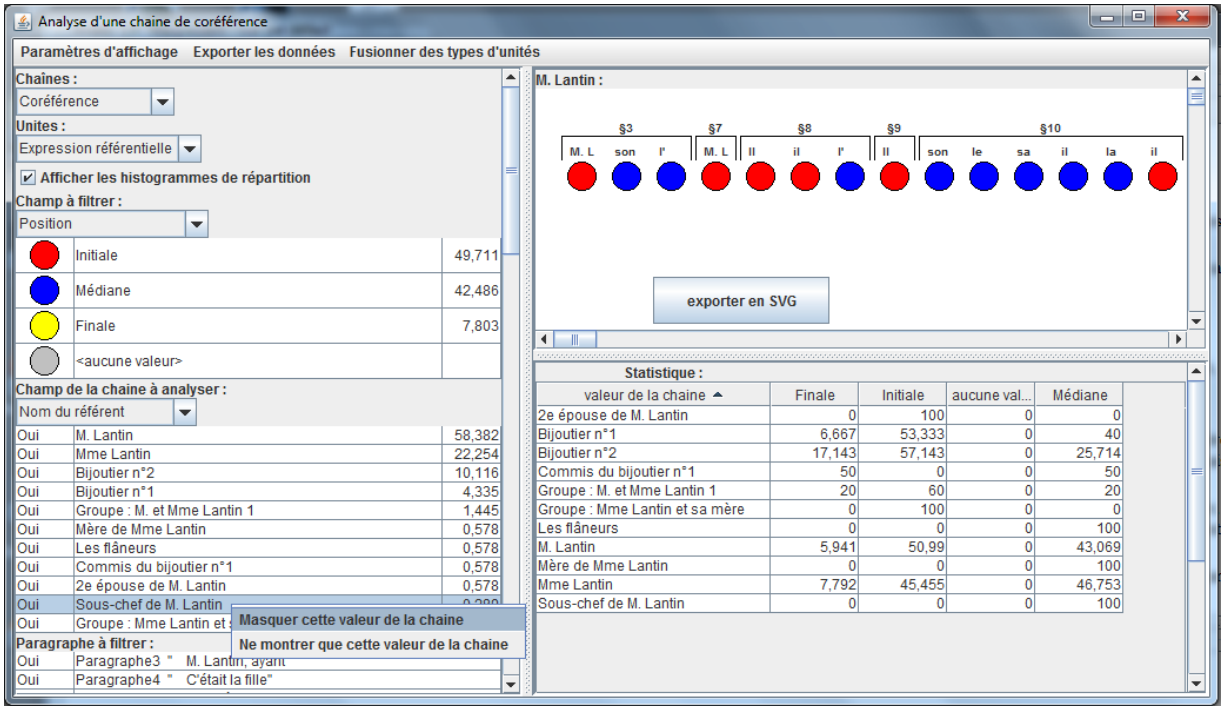

Figure 4 : visualisation des chaînes de coréférences et des répartitions (par chaîne) des valeurs des propriétés annotées. Ici, la chaîne affichée est celle de Mr Lantin, et le tableau des répartitions concerne la propriété « position du maillon dans la phrase », qui est aussi à l'origine du code couleur dans la représentation graphique de la chaîne.

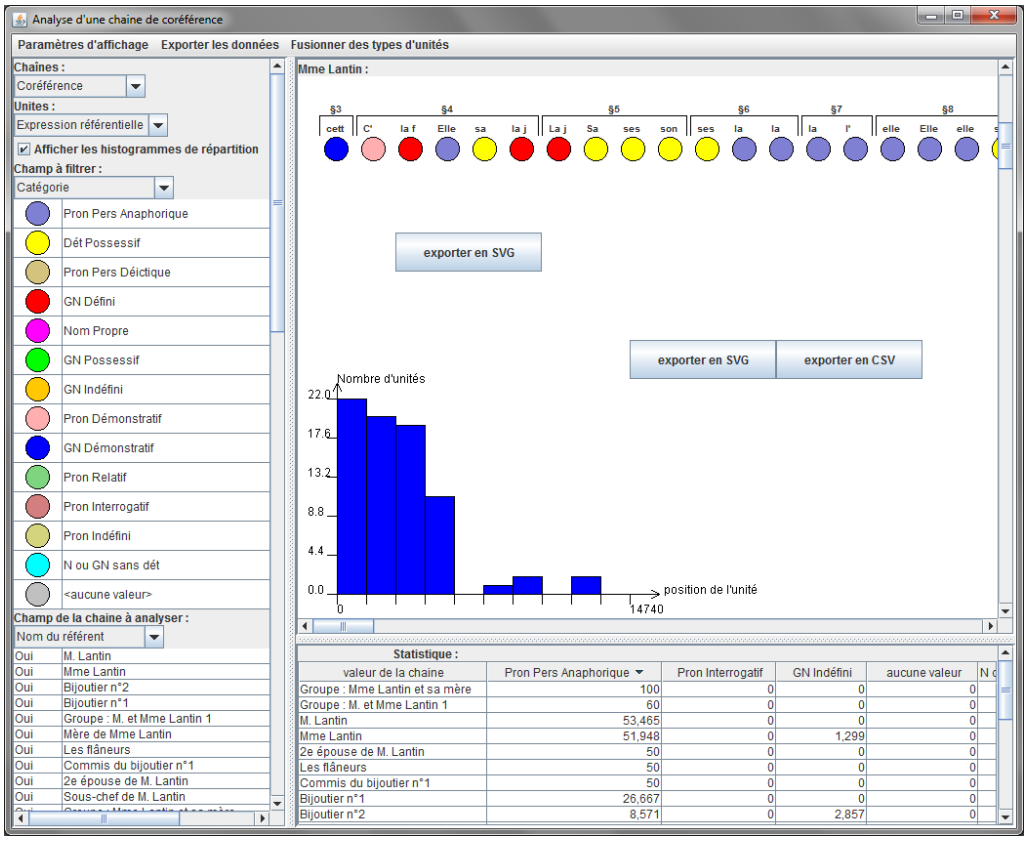

Figure 5 : autre visualisation, cette fois de la chaîne de Mme Lantin : on a ajouté un diagramme en bâtons montrant la répartition des maillons de cette chaîne le long du texte, et la propriété d'étude est ici la catégorie d'expression référentielle (défini, indéfini, démonstratif, possessif, etc.).

2.2. Représentations graphiques de chaînes de coréférences

On peut qualifier la visualisation des chaînes de coréférences dans GLOZZ d'« intégrée » : des cadres rouges se superposent au texte pour permettre à l'utilisateur de repérer où sont les chaînes et leurs maillons (figure 1). Dans notre outil (intégré à ANALEC en tant que module), nous avons choisi d'explorer la voie d'une visualisation « stand-off » : le texte n'apparaît plus, mais uniquement les maillons des chaînes, et ce sous une forme graphique simple (des ronds colorés, cf. le haut de la figure 4). Dans la configuration la plus basique, chaque rond représente un maillon, et l'utilisateur choisit une donnée annotée pour en faire un code couleur. L'enjeu est d'exploiter les théories de présentation d'information (Gestalt, saillance visuelle) de manière à offrir à l'utilisateur des moyens efficaces pour repérer des phénomènes intéressants, qu'il s'agisse de phénomènes linguistiques remarquables ou d'erreurs d'annotation – pour recommencer tout de suite celle-ci, cf. l'aspect cyclique de l'annotation décrit dans (Landragin *et al.*, 2012). Dans une configuration plus avancée, l'utilisateur choisit plusieurs codes graphiques. Il peut par exemple affecter un code couleur selon les valeurs d'une des propriétés annotées, affecter un code de taille selon les valeurs d'une deuxième propriété, ainsi qu'un code de forme géométrique selon une troisième propriété. Tout ceci afin de diversifier les types de visualisation et de faciliter la détection de phénomènes remarquables. Cette technique permet d'avoir un regard complémentaire de celui du lecteur naïf qui parcourt le texte avec une attention particulière pour les expressions référentielles.

Tous les éléments affichés sont interactifs : une info-bulle apparaît si on laisse la souris recouvrir un rond coloré (afin de présenter l'intégralité des informations caractérisant le

maillon correspondant) ; l'interface d'annotation s'ouvre directement sur l'expression référentielle si on clique sur un maillon ; des menus contextuels sont attachés à l'ensemble des données de paramétrisation, permettant notamment de masquer une partie des données annotées (figure 4, en bas à gauche).

Enfin, et il s'agit là aussi d'un principe décliné sur toutes les fenêtres de l'interface, chaque visualisation (ainsi que chaque résultat de calcul statistique) est enregistrable dans un fichier. La figure 6 montre ainsi une visualisation enregistrée dans un format graphique éditable, permettant de l'intégrer par exemple comme illustration dans un article de recherche.

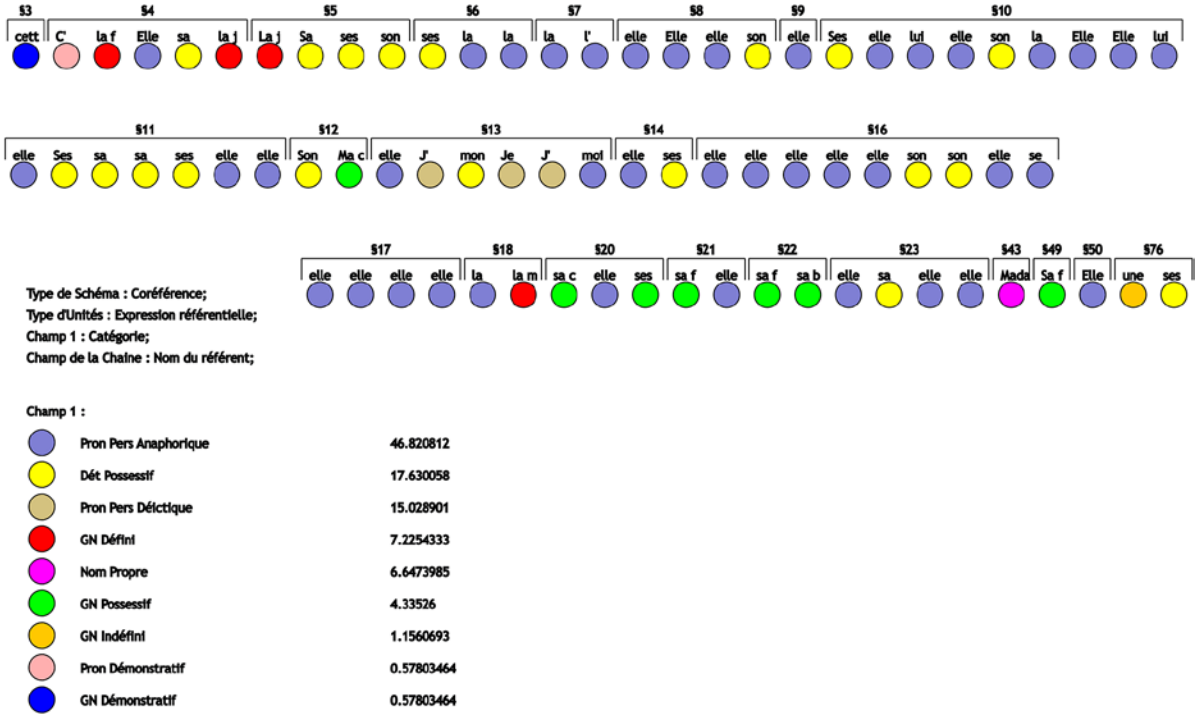

Figure 6 : exportation en format SVG de la représentation graphique de la chaîne, ce qui a permis ici de faire apparaître l'intégralité des maillons (édition – vectorielle – faite avec Adobe Illustrator).

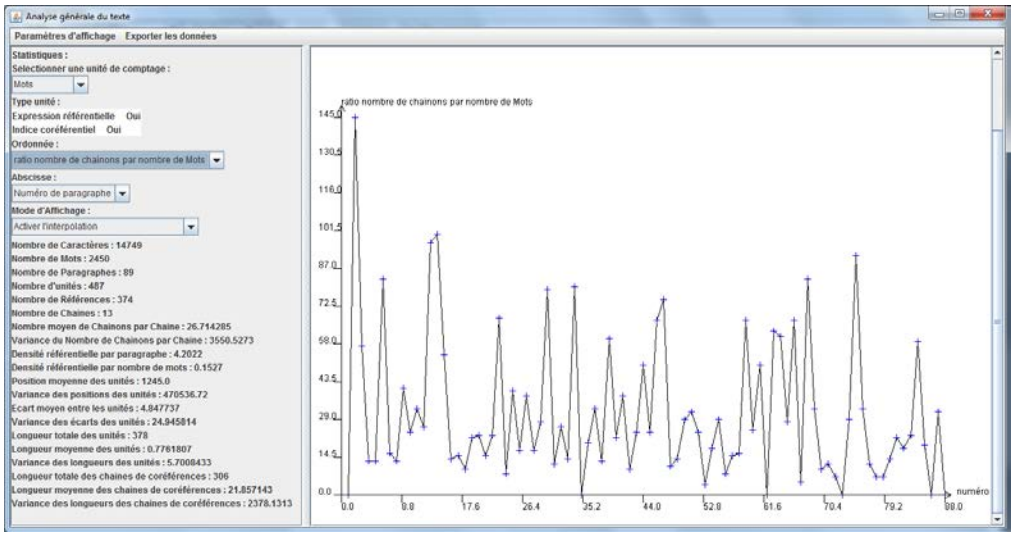

Figure 7 : statistiques générales sur le texte et ses annotations : nombre de chaînes de coréférences, nombre moyen de maillons, densités référentielles… et affichage d'un graphique parmi une dizaine possibles, ici le ratio « nombre de maillons » par « nombre de mots » (une valeur par paragraphe).

3. Statistiques sur les chaînes de coréférences

3.1. Statistiques générales

La première fenêtre de notre module présente des statistiques générales sur le texte et sur la répartition des chaînes dans ce texte (figure 7). Il s'agit de calculs très simples, qui ne font pas appel à des statistiques textuelles intéressantes, mais qui donnent néanmoins des repères utiles concernant le volume des annotations (des chaînes, donc) ainsi que leur répartition et leur densité tout au long du texte. C'est une étape nécessaire avant celle ciblée sur les chaînes.

3.2. Statistiques sur les chaînes de coréférences

Comme on l'a vu dans les figures 4 et 5 – dans la zone située en-dessous de la représentation graphique – la deuxième fenêtre de notre module présente des statistiques spécifiques aux chaînes de coréférences : à partir du moment où l'utilisateur choisit une propriété annotée, la répartition des valeurs prises par cette propriété dans les différents maillons de la chaîne étudiée est présentée. Il s'agit de simples décomptes, mais les possibilités de filtrage sont nombreuses : on peut par exemple ne considérer qu'un sous-ensemble des paragraphes du texte, qu'un sous-ensemble des valeurs elles-mêmes, etc. L'intérêt est de multiplier les possibilités d'interrogation des données annotées. Pour les statistiques qui s'y prêtent, notamment les densités référentielles – c'est-à-dire, pour un paragraphe donné, le nombre de maillons présents dans le paragraphe ramené au nombre de mots de celui-ci – un diagramme en bâtons est également affiché (figure 5).

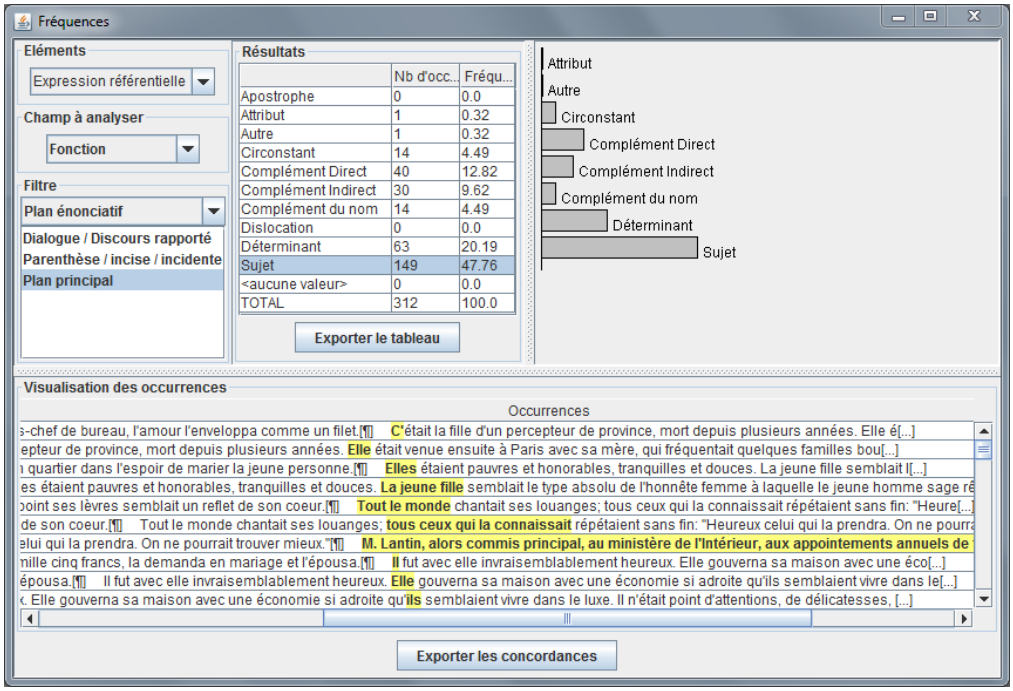

Figure 8 : statistiques sur les annotations : l'utilisateur a choisi de regarder l'une des propriétés annotées des expressions référentielles (la fonction grammaticale) et de calculer les fréquences des différentes valeurs en filtrant selon la valeur d'une autre propriété (le plan énonciatif). Il a ensuite choisi d'afficher le concordancier pour la fonction grammaticale « sujet ».

Par ailleurs, les statistiques qu'ANALEC est capable de calculer sur tous les types d'annotations restent accessibles et adaptables aux chaînes de coréférences. La figure 8 montre de simples calculs de fréquences avec un concordancier associé, mais des

fonctionnalités plus avancées sont également utilisables, notamment les tests de corrélations entre données annotées (chi-deux) et les analyses factorielles des correspondances (Landragin *et al*., 2012), elles-mêmes paramétrables, par exemple en sélectionnant quelques chaînes.

3.3. Statistiques sur la suite des références : bi-grammes et tri-grammes

A la suite de la représentation des chaînes par une succession de points colorés, nous avons étendu ce principe de code pour en faire le cœur de la troisième fenêtre d'exploration : quand on choisit un code, cela revient à considérer la suite des annotations comme un message écrit dans un alphabet particulier. Or ce message peut être analysé à l'aide de techniques classiques telles que le comptage des N-grammes, utilisées habituellement sur les mots et non sur les annotations. Comme le montre le bas de la figure 9 (qui par ailleurs affiche une visualisation combinant trois codes visuels), l'interface propose de compter les bi-grammes et les tri-grammes, et d'afficher ceux-ci dans un tableau. L'intérêt de cette information est d'apporter des arguments chiffrés à des questions de continuité référentielle. Entre autres exemples, on peut ainsi quantifier le fait qu'un paragraphe commence préférentiellement (ou non) avec le même référent que le dernier référent mentionné dans le paragraphe précédent.

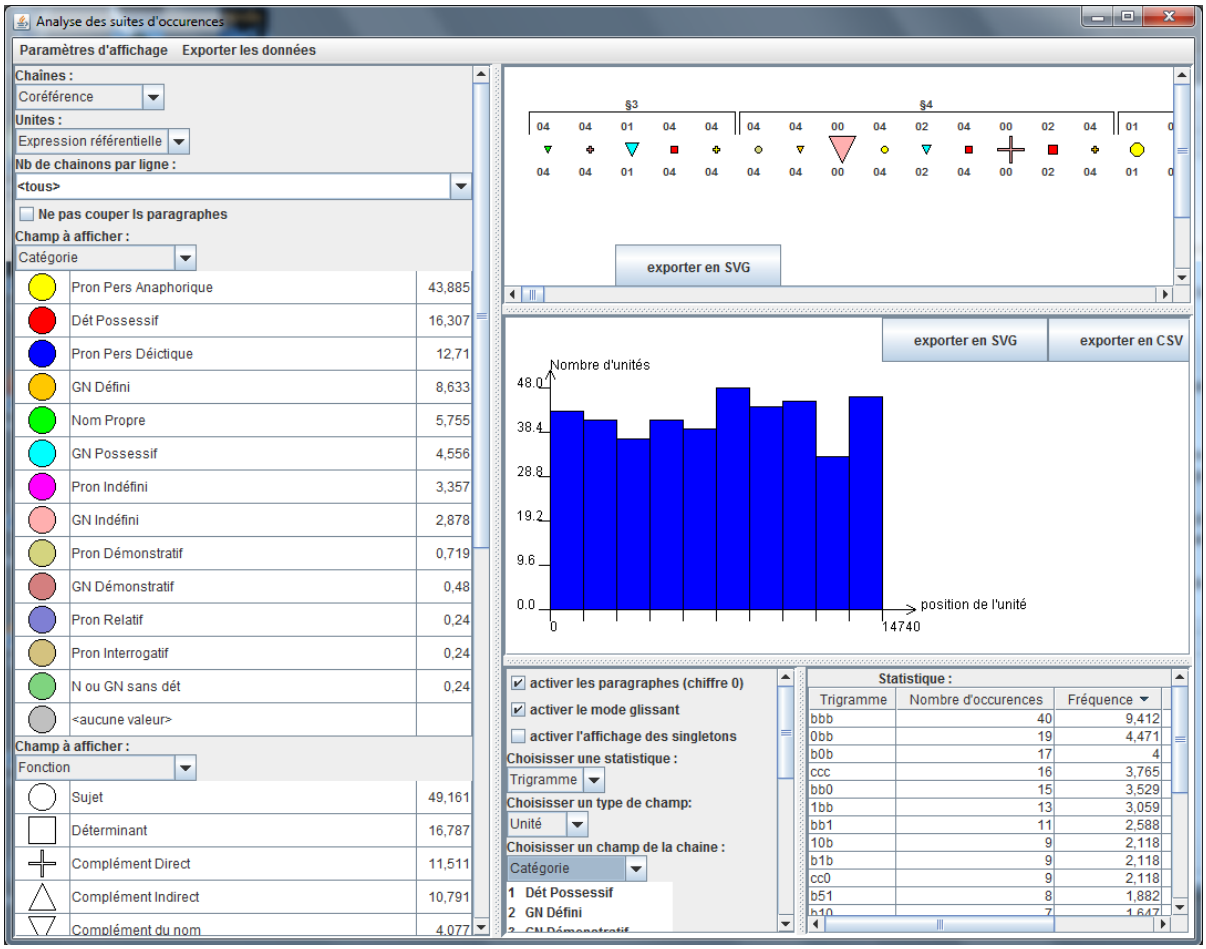

Figure 9 : statistiques plus complexes sur la suite de toutes les références du texte, avec : a. une visualisation graphique qui fait appel à plusieurs codes visuels (couleur, taille et forme) ; b. l'affichage de la densité référentielle ; et c. la paramétrisation d'un calcul de tri-grammes (ou de bi-grammes) avec les fréquences de ceux-ci (du plus au moins fréquent, la copie d'écran étant tronquée pour en permettre la lecture).

3.4. Discussion : intérêts et perspectives pour l'analyse de données textuelles

Comme le montrent les copies d'écran des figures précédentes, nous avons fait le choix d'une interface graphique simple – sans langage de requête ni langage de script – pour accéder au texte et aux données annotées : tous les choix de l'utilisateur se font avec des cases à cocher et des menus déroulants. Les concordanciers, tableaux et graphiques générés restent classiques, de même que les statistiques générées (décomptes, fréquences, n-grammes). De fait, notre outil est un prototype dont la conception a été dirigée par les besoins de linguistes participant au projet MC4 cité plus haut. Avec comme objectif d'aller directement aux informations essentielles face à un objet d'étude donné (les chaînes de coréférences), nous avons développé notre prototype en dehors d'un cadre ADT. Maintenant que le prototype est opérationnel et, inévitablement, montre ses limites, il est possible de remettre en question ce choix et d'identifier des voies d'amélioration pour la suite du projet DEMOCRAT.

D'un point de vue conceptuel, notre gestion des chaînes de coréférences peut se confronter à plusieurs approches de l'ADT. En premier lieu, on peut rapprocher notre étude à des travaux relevant de l'analyse de discours outillée par la textométrie, portant sur des objets proches du nôtre, et notamment quand il s'agit de marques de référence. C'est le cas du travail sur le pronom « nous » de (Née *et al.*, 2014) : comme pour nos expressions référentielles, il s'agit tout d'abord d'un travail de repérage et d'annotation manuelle, avec une finesse linguistique que le traitement automatique des langues ne permet pas encore. Il s'agit ensuite d'un travail de nature textométrique, avec des calculs de fréquences (sur les valeurs annotées aussi bien que sur le contexte d'utilisation de « nous »), des ventilations dans les différents sous-corpus annotés, des distributions différentielles et des recherches d'associations cooccurrentielles. Ces analyses – ainsi que la procédure d'annotation – sont permises par le logiciel Trameur (Fleury, 2013). Si elles se focalisent sur la contextualisation immédiate de chaque occurrence de « nous », elles négligent en revanche les successions de plusieurs occurrences et l'analyse de telles successions. Or c'est ce qui nous intéresse prioritairement avec les chaînes de coréférences, qui posent des problèmes différents, aussi bien au niveau de la contextualisation qu'à celui de la prise en compte de l'ordre d'apparition des différentes occurrences.

Deuxièmement, on peut rapprocher notre gestion des chaînes de coréférences de l'approche topologique des textes, qui apporte justement des réponses potentielles à ces problèmes. (Mellet & Barthélemy, 2007) montre que l'approche topologique – par nature multidimensionnelle comme l'est la nôtre – permet de tenir compte de la linéarité du texte et d'articuler le local et le global dans les analyses. La notion de motif semble s'adapter à la suite des références : transposer les exemples de (Mellet & Barthélemy, 2007) ou de (Longrée *et al.*, 2008) en codes tels que les tri-grammes de la section précédente se fait facilement. On peut même avancer que l'analyse de la suite des références d'un texte est une application possible des motifs, avec les mêmes avantages et les mêmes questions ouvertes : comment tenir compte des séquences textuelles qui ne comportent pas d'expression référentielle ? quelle place accorder à la ponctuation ? comment distinguer les motifs entièrement libres des motifs contraints par les structures de la langue (utilisation des pronoms anaphoriques, par exemple) ? L'approche semble compatible, nonobstant quelques ajustements de détails. Cependant, le rapprochement s'arrête avec le typage des relations anaphoriques et la construction des chaînes de coréférences, deux opérations qui nous éloignent des exemples à l'origine de la notion de motif (et des propriétés vérifiées par celui-ci). Nous retiendrons néanmoins qu'une approche topologique nous permettrait d'étudier les variations de la densité référentielle tout au long du texte, avec des analyses bien plus ciblées que les nôtres. C'est donc une perspective d'amélioration intéressante de notre travail, au niveau conceptuel.

Ces considérations nous conduisent à un troisième rapprochement, avec les travaux qui tendent à réduire le texte à une suite de codes. Nous retiendrons notamment la lexicométrie sur corpus étiquetés telle que décrite et discutée par (Pincemin, 2004) : le texte n'est plus une suite de mots, mais aussi et en même temps une suite d'annotations, référentielles pour ce qui nous concerne. L'approche est là encore multidimensionnelle, avec des dimensions de codage, des dimensions élémentaires, des dimensions d'analyse (anaphores et coréférences) et éventuellement des dimensions d'affichage, pour lesquelles on pourrait dépasser le cadre de (Pincemin, 2004) et imaginer des dimensions propres à l'affichage des chaînes de coréférences. La modélisation en « positions » et en « dimensions » nous semble adaptée à notre objet d'étude, même si, comme l'avait déjà souligné (Habert & Salem, 1995), il est souvent nécessaire de retravailler l'étiquetage soumis aux calculs afin d'obtenir des observations plus intéressantes. Rationaliser ce « retravail » pour les chaînes de coréférences et l'inclure dans l'interface graphique de notre outil constitue une perspective d'amélioration.

D'un point de vue technique, (Pincemin, 2004) présente trois façons suivies par les logiciels de lexicométrie pour s'ouvrir aux corpus annotés : a. la façon directe, « sans rien changer », consistant à considérer les étiquettes comme des mots, et donc à les inclure de manière brutale dans les explorations et calculs textométriques ; b. la façon « multiples vues dans une même base », qui ne permet pas le croisement des informations ; et c. la façon – prônée par l'auteur et que l'on retrouve dans la conception du logiciel TXM (Heiden *et al.*, 2010) – consistant à redéfinir la modélisation du corpus en lexicométrie pour la généraliser. Les positions et les étiquetages sont pris en compte, ce qui permet une gestion adaptée des analyses. Autrement dit, l'amélioration de notre outil pourrait se faire dans une voie similaire à celle de TXM.

4. Conclusion

Nous terminons ainsi notre tour d'horizon des fonctionnalités de notre nouvelle interface dédiée à l'étude des chaînes de coréférences : grâce à un ensemble de visualisations et de calculs, nous espérons apporter aux utilisateurs linguistes des informations utiles à leurs recherches sur les chaînes et la cohérence. À l'usage, notre propre constat est qu'une complémentarité reste nécessaire entre l'utilisation de cette interface et une lecture attentive du texte : l'outil ne permet pas l'émergence « magique » d'analyses parfaitement ciblées ; par contre, il apporte des confirmations à des tendances observées à la lecture, et des illustrations graphiques et quantifiées de ces tendances. Plusieurs perspectives concernent l'analyse des chaînes : au-delà de simples N-grammes, un enjeu consiste par exemple à rechercher des motifs plus souples et robustes quant au bruit généré par la multiplicité des référents.

Remerciements : *ce travail est issu des réflexions et des avancées du projet CNRS PEPS « MC4 : Modélisation Contrastive et Computationnelle des Chaînes de Coréférences » (appel INS2I-INSHS de 2011, merci notamment à Bernard Victorri et à Michel Charolles). Il a été réalisé avec le soutien de l'ANR dans le cadre du projet DEMOCRAT – ANR-15-CE38-0008 – et grâce aux compétences en Java de Marc Chataigner : merci à lui pour sa patience et pour son souci du détail.*

References

Charolles M. (1994). Cohésion, cohérence et pertinence du discours, *Travaux de linguistique*, 29 : 125-151.

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

Chiarcos C., Dipper S., Götze M., Leser U., Lüdeling A., Ritz J., Stede M. (2008). A Flexible Framework for Integrating Annotations from Different Tools and Tag Sets. *TAL*, 49 (2): 217-248.

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

Désoyer A., Landragin F., Tellier I., Lefeuvre A., Antoine J.-Y. (2014). Les coréférences à l'oral : une expérience d'apprentissage automatique sur le corpus ANCOR. *Traitement Automatique des Langues*, 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 (février 2013)*, 53 pages.

Habert B, Salem A. (1995). L'utilisation de catégorisations multiples pour l'analyse quantitative de données textuelles, *Traitement Automatique des Langues*, 36(1-2) : 249-275.

Heiden S., Magué J.-P., Pincemin B. (2010). TXM : Une plateforme logicielle open-source pour la textométrie – conception et développement, *Statistical Analysis of Textual Data – Proceedings of 10th International Conference Journées d'Analyse statistique des Données Textuelles (JADT 2010)*, Edizioni Universitarie di Lettere Economia Diritto, Rome, Italie, pp. 1021-1032.

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

Landragin F. (2011). Une procédure d'analyse et d'annotation des chaînes de coréférence dans des textes écrits. *Corpus*, 10, http://corpus.revues.org.

Landragin F., Poibeau T., Victorri B. (2012). ANALEC: a New Tool for the Dynamic Annotation of Textual Data. In: *LREC 2012*, Istanbul, Turquie, pp. 357-362.

Landragin F., Schnedecker C. (Eds.) (2014). *Les chaînes de référence*. Numéro 195 de la revue Langages, Armand Colin, Paris.

Legallois D. (Ed.) (2006). *Organisation des textes et cohérence du discours*, Numéro thématique de la revue CORELA, http://edel.univ-poitiers.fr/corela.

Lezius W. (2002). TIGERSearch – Ein Suchwerkzeug für Baumbanken. In: *Proceedings der 6. Konferenz zur Verarbeitung natürlicher Sprache (KONVENS)*, Saarbrücken, pp. 107-114.

Longrée D., Luong X., Mellet S. (2008). Les motifs : un outil pour la caractérisation topologique des textes, *Actes des 9e Journées internationales d'Analyse statistique des Données Textuelles (JADT 2008)*, Vol. 2, Presses universitaires de Lyon, Lyon, pp. 733-744.

Mellet S., Barthélemy J.-P. (2007). La topologie textuelle : légitimation d'une notion émergente, *Lexicometrica*, 7 (*Topographie et topologies textuelles*), en ligne, 14 pages.

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*, Peter Lang, Frankfurt, Germany.

Née E., Sitri F., Fleury S. (2014). L'annotation du pronom « nous » dans un corpus de rapports éducatifs. Objectifs, méthodes, résultats, *Actes des 12e Journées internationales d'Analyse statistique des Données Textuelles (JADT 2014)*, Paris, 495-506.

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

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

Steiner I., Kallmeyer L. (2002). VIQTORYA – A Visual Query Tool for Syntactically Annotated Corpora. In: *Proceedings of LREC*, Las Palmas, Gran Canaria.

Venant F. (2008). Semantic Visualization and Meaning Computation. Demonstration at: *22nd International Conference on Computational Linguistics (COLING)*, Manchester, England.

Widlöcher A., Mathet Y. (2009). La plate-forme Glozz : environnement d'annotation et d'exploration de corpus. In : *Actes de la 16e Conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, Senlis. Cf. aussi le site *web Glozz Annotation Platform*, http://www.glozz.org/.