Anaphores et coréférences : analyse assistée par ordinateur

– DRAFT AUTEUR –

Frédéric Landragin
Lattice, CNRS/ENS/Université de Paris 3 Sorbonne Nouvelle

1. Introduction

Si l'annotation, la visualisation et l'interrogation de corpus pour des phénomènes morphosyntaxiques et syntaxiques posent beaucoup de problèmes techniques et requièrent des outils adaptés, il en est de même, et peut-être encore plus, pour des phénomènes sémantico-pragmatiques tels que la référence, l'anaphore et la coréférence. D'une part parce que gérer informatiquement des anaphores et des chaînes de coréférence nécessite des structures de données complexes, couvrant potentiellement un texte entier, d'autre part parce que travailler sur la référence implique de construire le monde des référents, avec notamment une modélisation des liens entre chacun d'eux. Notre objectif dans cet article est d'illustrer quelques apports de la linguistique outillée, et en particulier du logiciel ANALEC (Victorri, 2012), pour des études linguistiques, sur les données structurées que sont les chaînes de coréférence. Nous parlons ici de linguistique outillée, mais on peut, à la suite de Habert (2005), faire la distinction entre linguistique à l'instrument et linguistique outillée : dans le premier cas, le logiciel informatique est conçu pour le traitement de matériau linguistique et permet d'obtenir des représentations transformées ; dans le second cas, le logiciel n'est pas spécifiquement orienté vers le traitement de données linguistiques, mais peut être utile pour en tirer des observations telles que des calculs de fréquences par exemple. Comme nous le verrons, ANALEC sert à la fois d'instrument – il permet de construire les représentations transformées que sont les chaînes de coréférence – et d'outil plus général – il permet d'effectuer un ensemble de calculs statistiques. Comme nous le verrons également, un outil tel qu'un tableur a aussi un intérêt pour les études linguistiques qui nous concernent.

Pour illustrer notre propos, nous avons choisi les chaînes de coréférence parce qu'il s'agit d'une structure qui relie un ensemble de mentions, dans un ordre précis, avec éventuellement des relations anaphoriques entre ces références (Landragin, 2011). Les problèmes techniques posés par l'annotation et l'étude d'anaphores sont donc un sous-

ensemble de ceux posés par l'annotation et l'étude de chaînes de coréférence. Les remarques qui suivent restent par ailleurs valables pour toute structure d'annotation de type « schéma » au sens de Widlöcher & Mathet (2009), c'est-à-dire pour toute structure regroupant des données annotées sous la forme d'« unité » (suite consécutive de caractères, délimitée dans une donnée unitaire que le linguiste va pouvoir enrichir d'annotations), de « relation » (lien entre deux unités) et de « schéma » (ensemble structuré d'unités, de relations, et, de manière récursive, de schémas).

Ainsi, une mention ou expression référentielle correspond à une « unité », qui porte (éventuellement) des annotations sous la forme d'une structure de traits, c'est-à-dire sous la forme de couples attribut-valeur. Une relation anaphorique entre deux mentions peut se construire via une « relation », elle-même pouvant porter des annotations sous la forme d'une structure de traits. Une chaîne de coréférence peut se construire soit *de facto* par un ensemble d'unités et de relations qui forment une liste chaînée, soit via un « schéma » qui groupe alors mentions et relations, et auquel on peut bien entendu adjoindre sa propre structure de traits. C'est la première solution, celle de la liste chaînée, qui est utilisée dans la plupart des travaux informatisés sur les chaînes de coréférence, que ce soit dans les campagnes de type MUC (« *Message Understanding Conferences* ») ou via des outils tels que MMAX et MMAX2 (Müller & Strube, 2006). Comme il suffit de repérer les unités puis de les lier les unes aux autres en suivant l'ordre du texte, elle semble plus simple à mettre en œuvre. C'est la seconde solution que nous avons choisie : elle s'avère plus efficace (un schéma regroupe toutes les mentions d'un même référent, sans qu'on ait besoin de spécifier l'intégralité des relations anaphoriques coréférentielles) et plus logique (elle réserve la « relation » aux anaphores non coréférentielles). Des conversions de l'une à l'autre solution sont toujours possibles, mais il vaut mieux partir d'une représentation adaptée au maximum de phénomènes linguistiques. Cette seconde solution nous permet d'ajouter des annotations à chaque chaîne, et donc à chaque référent. C'est aussi une solution adoptée pour l'annotation d'autres types de structures discursives, par exemple dans le projet ANR Annodis (« Annotation du discours ») à l'origine de l'outil GLOZZ (Mathet & Widlöcher, 2012) dont ANALEC reprend la structure unité-relation-schéma. La figure 1 schématise les deux solutions.

Figure 1 à insérer à peu près ici : Chaîne par relations successives ou par regroupement.

Dans cet article, nous ne donnons pas ou peu de résultats d'analyses linguistiques de chaînes de coréférence dans des textes particuliers. Nous présentons avant tout une méthodologie de linguistique instrumentée et outillée, tournée vers l'analyse d'annotations,

que celles-ci aient été obtenues manuellement, automatiquement, ou par combinaison des deux approches. C'est la voie prise par Bernard Victorri lors de la conception des logiciels ANALOR (Avanzi *et al*., 2008) puis ANALEC, où l'annotation manuelle est prévue aussi bien pour la saisie de données nouvelles que pour la correction ergonomiquement rapide d'annotations existantes. Concernant la combinaison des approches manuelles et automatiques pour la coréférence, il reste un problème posé par la fusion des annotations : la structure d'annotation la plus complexe doit être utilisée comme pivot afin de rendre compte de la finesse des phénomènes. Dans notre cas, il s'agit donc de garder des structures composées d'unités, de relations et de schémas, et de définir des procédures d'intégration d'annotations automatiques dans ce cadre, plutôt que d'essayer de le réduire à des unités et des relations – sans schémas – comme c'est le cas dans MUC ou ailleurs (Van Deempter & Kibble, 2000).

Ce choix n'est pas sans conséquences. Une structure comme un « schéma » s'avère difficile à manier : difficile à appréhender pour un annotateur, et difficile à représenter graphiquement pour un outil. Il existe une panoplie d'outils de visualisation conçus pour des données sémantiques (Venant, 2008). Sur le problème spécifique des chaînes de coréférence, GLOZZ dans sa dernière version est peut-être le plus pertinent. Citons également ANNIS, et d'une manière générale les plateformes citées dans (Chiarcos *et al*., 2008). Visualisation et interrogation vont souvent de pair, et, là aussi, une structure comme un schéma ne facilite pas les outils de recherche. Avec TIGERSEARCH (Albert *et al.*, 2003 ; Lezius, 2002) ou encore VIQTORYA (Steiner & Kallmeyer, 2002), les requêtes parcourent les structures complexes que sont les arbres syntaxiques. Pour les chaînes de coréférence, parcourir des arbres localisés au niveau de la phrase ne suffit pas, et il faut là encore se tourner vers GLOZZ et sa gestion des schémas, avec le langage de requête GLOZZQL (Mathet & Widlöcher, 2012). Quant à ANALEC sur lequel nous allons nous focaliser, ce n'est que récemment qu'il dispose d'une fonctionnalité de représentation graphique de chaînes. Associée aux fonctionnalités existantes telles que les représentations géométriques d'analyses factorielles de correspondance ou le rendu coloré de tests de khi-deux, cette fonctionnalité nous semble importante pour l'appréhension, l'annotation et la vérification de données portant sur des références et des coréférences. L'enjeu est d'exploiter des principes simples de présentation d'information (Gestalt, couleur, saillance visuelle) de manière à offrir à l'utilisateur des moyens efficaces pour repérer des phénomènes intéressants, ceux-ci pouvant correspondre à des erreurs d'annotation, des exemples déviants, ou juste des phénomènes linguistiques remarquables.

Avant d'entrer dans les détails de l'utilisation d'ANALEC, la section 2 décrit les spécificités des chaînes de coréférence et des schémas associés, et présente un ensemble de besoins relevant de la linguistique de corpus outillée pour ce qui concerne l'analyse de ces chaînes de coréférence. En réponse aux plus graphiques de ces besoins, la section 3 présente les visualisations disponibles dans ANALEC. En réponse aux plus numériques de ces besoins, la section 4 décrit quelques procédures de calcul à effectuer sur les chaînes pour obtenir des analyses quantifiées. Nous concluons alors sur les intérêts principaux d'ANALEC et sur les besoins pour lesquels de nouveaux outils informatiques sont toujours attendus.

2. Nature des chaînes de coréférence et problématiques associées

Nous avons au départ deux phénomènes très différents : la référence, qui consiste à identifier par quel mécanisme une entité du discours est désignée par une mention linguistique (Charolles, 2002), et la chaîne de coréférence, qui consiste à lier entre elles des mentions linguistiques relatives à la même entité de discours (Schnedecker, 1997 ; van Deempter & Kibble, 2000). Sans tenir compte de la référence, une chaîne de coréférence n'est qu'un ensemble de liens entre mentions. En tenant compte de la référence, une chaîne de coréférence est une notion complexe qui décrit l'ensemble des moyens mis en œuvre dans un texte pour désigner une entité de discours. Nous nous intéressons ici à la nature de cette notion et à la structure d'annotation retenue pour la décrire. Par ailleurs, nous présentons un ensemble indicatif d'hypothèses linguistiques la concernant, avant d'expliciter les rôles de la linguistique outillée dans le test de ces hypothèses.

2.1. Les chaînes de coréférence et leurs maillons

Selon qu'on étudie des articles de presse ou des romans, les chaînes de coréférence peuvent inclure plus ou moins de mentions. Avec le point de vue consistant à regrouper l'ensemble des mentions désignant une même entité du discours, par exemple un personnage, une même chaîne de coréférence peut couvrir un roman entier. Il est alors difficile de l'analyser tant elle comprend de « maillons ». Si le personnage n'a pas été mentionné pendant plusieurs chapitres, il est par ailleurs peu pertinent de considérer sa réapparition comme un $n^{ième}$ maillon (Schnedecker, 1997). Une solution consiste à considérer des sous-chaînes à l'intérieur d'une même chaîne, en exploitant le découpage en chapitres (voire en paragraphes) pour segmenter la chaîne en sous-chaînes. Une autre solution consiste à construire et annoter d'un côté les chaînes complètes, de l'autre côté les chapitres, sections, paragraphes et autres titres et sous-titres, puis à intégrer les deux lors de la phase de visualisation, c'est-à-dire de

matérialisation graphique des structures annotées pour les appréhender visuellement, et lors de la phase d'interrogation, c'est-à-dire de saisie de requêtes pour explorer les données annotées (Landragin, 2011).

La figure 2 montre sous forme de schéma des matérialisations de ces deux solutions. La partie supérieure de la figure explicite un découpage de chaîne en sous-chaînes sur le critère de changement de paragraphe. Nous avons fait le choix de ne pas procéder à une telle identification, pour la raison que les changements de paragraphe sont détectables automatiquement, et qu'il revient donc à l'outil d'en tenir compte lors de toute requête de visualisation ou de calcul sur une chaîne. La partie inférieure de la figure montre un exemple classique faisant intervenir deux référents, un village et son église. Une chaîne de coréférence est définie pour chacun de ces deux référents, avec détermination des maillons (expressions référentielles). Chaque maillon peut être annoté avec des caractéristiques qui lui sont propres, par exemple « pronom de 3^{e} personne » pour « il » et « elle », « SN défini » pour « le village » et « l'église », etc. Chaque chaîne de coréférence peut également être annotée avec des caractéristiques qui lui sont propres, c'est-à-dire des caractéristiques du référent. Ainsi, « le village » pourra avoir une annotation explicitant qu'il s'agit d'un référent inanimé, concret, singulier. Enfin, l'anaphore associative (qui caractérise l'utilisation du SN « l'église » peu après avoir mentionné « le village », celui-ci servant d'antécédent pour interpréter celui-là) est modélisée sous la forme d'une relation, avec plusieurs possibilités. Soit on spécifie la relation entre les deux chaînes de coréférence, ce qui revient à considérer que la relation vaut lors de l'introduction du deuxième référent. C'est le choix qui a été fait dans (Landragin, 2011), et qui permet de n'avoir qu'au maximum une relation entre deux chaînes données. Soit on spécifie la relation entre deux mentions, ce qui est illustré figure 2 et ce qui permet de considérer plusieurs relations entre deux mêmes chaînes, par exemple à chaque reprise de sous-chaîne, du moins si le texte comporte bien une expression qui peut être interprétée comme une anaphore associative.

Figure 2 à insérer à peu près ici : Chaînes de coréférence, sous-chaînes et relations anaphoriques.

La gestion des chaînes n'est donc pas un problème simple. De plus, quand l'on considère une chaîne comme un ensemble regroupant plusieurs mentions, on entend implicitement que chaque mention est en relation de coréférence avec toutes les autres mentions de la chaîne. Si l'on souhaite annoter les différents types d'anaphores coréférentielles apparaissant au sein d'une chaîne, par exemple les anaphores pronominales et les anaphores non pronominales, il est nécessaire d'ajouter des relations entre certains

maillons et d'enrichir cette relation par une structure de traits. Cette structure peut ainsi inclure un trait décrivant s'il s'agit d'une anaphore pronominale, ou – autre exemple – s'il s'agit d'une anaphore fidèle (l'antécédent et l'expression anaphorique partagent la même tête nominale) ou infidèle (deux têtes nominales différentes). On obtient alors un schéma « chaîne de coréférence » qui inclut des unités « mentions » et des relations « anaphores » entre certaines d'entre elles. La structure se complique donc, et elle peut devenir d'une complexité difficile à appréhender dès que l'on considère les anaphores associatives comme on l'a vu ci-dessus, dans la mesure où cela conduit à gérer une structure arborescente de schémas et de mentions. C'est ce type de structure de données que les outils actuels ne permettent pas de visualiser aisément.

En outre, chaque mention intervient à un endroit du texte qui se caractérise par un certain nombre d'aspects syntaxiques, sémantiques ou discursifs. La nature de la phrase voire de la proposition dans laquelle la mention apparaît (incise, parenthèse, etc.), la nature de la mention elle-même (expression référentielle, sujet zéro d'un infinitif ou d'un participe), ou encore le plan énonciatif (narration, dialogue) sont autant d'arguments qui conduisent à considérer plusieurs niveaux de contribution d'une mention à une chaîne de coréférence. Ces niveaux de contribution peuvent être codés explicitement, avec par exemple une distinction entre maillons forts et maillons faibles (Landragin, 2011), distinction qui permet d'opposer les expressions référentielles (maillons forts) des sujets non exprimés de verbes, voire des marques d'accord en genre et en nombre qui, sans référer, évoquent les référents et contribuent ainsi aux chaînes de coréférence (maillons faibles). Plusieurs niveaux de contribution peuvent être également pris en compte par le biais de l'importance donnée à certains traits présents dans les données annotées, avec par exemple la gestion d'un système de pondération des traits. Dans les deux cas, il est nécessaire de réfléchir à la structure des données annotées, et de répondre aux problèmes suivants :

- De quelles annotations a-t-on besoin pour réaliser une étude sur la référence et la coréférence ? A priori on peut se contenter d'informations telles que la catégorie de l'expression référentielle (GN défini, pronom de 3^{e} personne, cf. plus haut) et la présence ou non de modifieurs, c'est-à-dire des informations qui caractérisent l'expression référentielle elle-même. Cependant, si on veut obtenir des renseignements sur les liens entre référence et rôles thématiques, entre référence et macro-syntaxe (ou discours rapporté, ou toute donnée issue d'une analyse linguistique de la phrase), il est nécessaire de considérer ces données comme des annotations nécessaires (qui peuvent être éventuellement obtenues de manière automatique).

- Quelles annotations va-t-on affecter aux unités « expression référentielle » ?
- Quelles annotations va-t-on affecter aux schémas « chaîne de coréférence » ?
- Quelles unités, relations, voire schémas supplémentaires va-t-on introduire pour autoriser la recherche de liens entre expressions référentielles et données relevant de l'analyse de la phrase (voire du discours) ?
- Quelles vont être les conséquences de ces choix sur les possibilités de visualisation et d'interrogation du corpus ?

La figure 3 montre un exemple correspondant à un projet d'annotation manuelle avec GLOZZ et ANALEC de textes courts en références et coréférences. L'exploitation interactive de ces deux outils est rendue possible par le fait qu'ils partagent la même structure de données pour représenter les annotations dans des fichiers informatiques. Le choix d'un outil plutôt que l'autre lors des phases d'annotation, de visualisation et d'interrogation repose sur les fonctionnalités complémentaires des deux interfaces graphiques. Dans ce projet, nous avons choisi de maximiser les données annotées, et de toutes les affecter aux unités « expression référentielle ». On se retrouve ainsi avec, dans la colonne de gauche, une unité « Maillon » (expression référentielle, mais pas seulement) qui comporte une structure de traits conséquente : 11 traits, dont 2 sont développées en figure 3 pour en voir les valeurs possibles. C'est le choix qui a été fait, mais nous aurions pu multiplier les types d'unité, avec par exemple une unité « Maillon » réduite à 3 ou 4 traits, et des unités « Phrase », « Enonciation » ou autre, avec chacune leurs propres traits. Si nous avons décidé de tout rassembler dans un seul type d'unité, c'est pour profiter d'une interface d'annotation unique, avec toutes les informations rassemblées au même endroit. Le travail des annotateurs est ainsi facilité. Une autre conséquence de ce choix a trait à l'interrogation des données annotées : il est plus facile de spécifier des requêtes quand toutes les données sont centralisées que quand elles sont dispersées sur plusieurs éléments de la structure d'annotation.

Figure 3 à insérer à peu près ici : Exemple de structure d'annotation.

Enfin, nous avons jusqu'ici évoqué les personnages d'un texte narratif, mais la référence et la coréférence peuvent intervenir pour tout type d'entités du discours, animées ou inanimées, objets ou événements. La coréférence événementielle et la nature des relations entre participants et événements étant des sujets de recherche en soi, la construction d'une structure de données regroupant les différentes chaînes de coréférence d'un texte n'est pas chose aisée ni consensuelle. Par rapport à la structure montrée en figure 3, il faudrait notamment ajouter d'autres types de maillons…

Cette figure 3 est une copie d'écran d'ANALEC, mais nous aurions pu montrer chacune des structures de trait avec des copies d'écran de GLOZZ, les deux logiciels utilisant le même format de codage pour le corpus annoté. C'est d'ailleurs l'interface principale de GLOZZ qui est montrée figure 4, et l'on y retrouve dans un texte (la nouvelle *Les bijoux* de Maupassant), un ensemble de maillons annotés (cadres) et une chaîne de coréférence (représentée sous la forme d'une liste chaînée de maillons). La chaîne de coréférence correspond à la jeune fille, l'un des personnages principaux du texte. A droite de l'écran se trouve l'interface d'annotation avec certains des traits illustrés dans la figure 3. Avec les figures 3 et 4, nous montrons quelques-unes des fonctionnalités d'annotation des logiciels GLOZZ et ANALEC, et nous nous situons pour l'instant dans le périmètre de la linguistique à l'instrument.

Figure 4 à insérer à peu près ici : Interface principale de GLOZZ.

2.2. Unités, relations et schémas pour l'annotation de chaînes

Au-delà d'un ensemble d'unités et de relation comme cela est fait manuellement (MMAX) ou automatiquement (MUC), nous utilisons donc un schéma pour représenter informatiquement des chaînes de coréférence. Au final, suite à notre étude de la référence et de la coréférence et à quelques essais préliminaires, une structure d'annotation complète et idéale se composerait :

- d'unités « maillon » pour les mentions, qu'elles soient explicites (expression référentielle) ou atténuées (sujet zéro d'un infinitif ou d'un participe) ;
- de relations « anaphore coréférentielle » pour affiner les liens entre maillons d'une même chaîne (cette relation n'apparaît pas dans la figure 3) ;
- de schémas « coréférence » incluant des instances des unités et des relations précédentes pour les chaînes de coréférence ;
- de relations « anaphore non coréférentielle » entre maillons de chaînes différentes pour représenter les anaphores associatives ;
- de relations « connexion » pour lier entre elles deux chaînes, c'est-à-dire pour décrire d'une part les cas d'appartenance d'un référent individu à un référent groupe, d'autre part les cas de participation d'un référent à un événement ;
- d'unités « chapitre », « paragraphe » voire « phrase » pour la segmentation textuelle, à partir du moment où – comme c'est le cas avec les paragraphes dans GLOZZ – ces unités sont repérées automatiquement par l'outil et n'apparaissent pas dans l'interface d'annotation ni dans la structure d'annotation (elles ne font qu'apporter des paramètres

supplémentaires lors de la visualisation et de l'interrogation, de manière transparente pour l'utilisateur) ;

- d'unités supplémentaires telles que « mot » ou « chunk », obtenues automatiquement suite à l'utilisation d'un ou de plusieurs analyseurs (morpho)syntaxiques, et visant là aussi à apporter des paramètres supplémentaires ;
- d'unités et de relations (dépendance par exemple) obtenues automatiquement suite à l'utilisation d'un analyseur tel que SYNTEX (ANALEC importe de telles annotations et permet donc de les exploiter ensuite lors de l'interrogation) ;
- d'unités, de relations voire de schémas pour toute autre donnée susceptible d'apporter des informations liées de près ou de loin aux chaînes de coréférence, un exemple exploré dans l'un des groupes de travail du laboratoire Lattice étant les chaînes d'organisateurs textuels.

On le voit, on peut vite se perdre dans des structures complexes, avec des données qui proviennent de sources multiples et qui font éventuellement appel à plusieurs outils en cascade. Nous reviendrons sur ce point dans la conclusion.

2.3. Types d'hypothèses linguistiques concernées

Etudier avec une approche outillée les chaînes de coréférence présente plusieurs intérêts. Il s'agit premièrement de tester des hypothèses sur les typologies des chaînes présentes dans un texte, avec éventuellement des comparaisons entre genres textuels. Notre but dans cet article n'est pas de répondre à ces hypothèses, mais de considérer les possibilités et impossibilités techniques pour le faire. Nous mentionnons à titre d'exemples les idées de calculs suivantes :

- Une chaîne commence-t-elle par un nom propre, par une description indéfinie ?
- Quels sont les seconds maillons les plus probables, compte tenu du genre textuel (certains journaux – voire la presse écrite en général – sont connus pour introduire un référent humain tel qu'une personnalité, premièrement via un nom propre, et deuxièmement via une description définie ou démonstrative qui indique par exemple le métier ou la fonction) ?
- Combien de fois le pronom de 3^{e} personne apparaît-il consécutivement ?
- Après combien de maillons de type pronom réapparaît le nom propre ?
- A quel rang dans une chaîne un démonstratif apparaît-il ?
- Quelles sont les proportions de maillons faibles et de maillons forts ?

Il s'agit deuxièmement de calculer des indices de répartition et de croisement de chaînes de coréférence dans un texte, de manière à apporter des données aux études sur la cohérence et la cohésion :

- Existe-t-il des phrases ne comportant aucun maillon d'aucune chaîne ?
- Les phrases sont-elles reliées les unes aux autres par au moins une chaîne ?
- Y a-t-il une corrélation entre la segmentation en sous-chaînes et le découpage du texte en paragraphes ?
- Quelle est la proportion des continuations sur un même référent (autrement dit des suites d'au moins deux mentions coréférentes) ?
- Quel est le nombre de passage d'un référent à un autre en fonction du nombre total de mentions ?

Troisièmement, nous pouvons envisager des calculs destinés à apporter des éléments aux théories sur l'accessibilité et la saillance des référents (Schnedecker, 2011) :

- Vérifie-t-on au sein d'une chaîne une augmentation de l'accessibilité du référent (il faut pour cela coder l'accessibilité, mais les marqueurs – pour autant qu'ils soient repérés – le font) ?
- Y a-t-il à tout moment du texte un seul centre, au sens de la Théorie du Centrage (Grosz *et al.*, 1995), ou au contraire plusieurs référents saillants ?

Cette liste de questions et d'hypothèses liées à l'étude de la référence et de la coréférence montre qu'une étude approfondie de la coréférence passe par beaucoup de questions scientifiques, de tests, de calculs, qui vont au-delà de ce que proposent les méthodes de repérage automatique de coréférences telles qu'elles sont actuellement développées en traitement automatique des langues. Pour de telles préoccupations, des outils d'annotation manuelle et de traitement des données annotées sont indispensables. Ces préoccupations requièrent en fait des fonctionnalités qui ne sont pas incluses dans la majorité des outils : les données structurées liées aux chaînes de coréférence sont complexes et les procédures de calcul trop spécifiques. Notre approche consiste à exploiter dans des outils tels que GLOZZ et ANALEC toutes les possibilités de représentation graphique et de génération de tableaux de chiffres, de manière à ce que l'utilisateur puisse naviguer rapidement de l'un à l'autre en se reposant sur les capacités humaines de perception visuelle pour détecter les phénomènes intéressants. Il s'agit donc d'exploiter des techniques de présentation graphique d'information. C'est un aspect particulièrement développé dans ANALEC, avec des calculs de fréquences qui apparaissent sous la forme classique d'histogramme et des calculs de

corrélations qui mettent en œuvre un code couleur pour indiquer la significativité. Nous ne nous étendrons pas ici sur ces aspects qui ne sont pas spécifiques aux chaînes de coréférence et qui sont présentés par ailleurs (Landragin *et al.*, 2012). Par contre, nous nous intéresserons à une fonctionnalité de visualisation d'ANALEC particulièrement pertinente pour les chaînes de coréférence.

3. Visualisation de chaînes de coréférence

Les phénomènes de coréférence peuvent être visualisés dans des outils tels que MMAX ou GLOZZ sous la forme de listes chaînées, comme dans la figure 4. GLOZZ ajoute la possibilité de visualiser l'intégralité du texte (figure 5), de manière à repérer visuellement les zones les plus concernées par une chaîne de coréférence. Cette fonctionnalité a été développée à l'origine pour la visualisation des structures discursives et argumentatives, mais s'avère totalement adaptée aux chaînes de coréférence. L'un des intérêts, non visible sur la figure 5, est d'exploiter, en même temps que des distances réduites à l'échelle de la taille du texte, un code couleur : si par exemple chaque maillon de la chaîne est coloré selon la catégorie du maillon (les noms propres en rouge, les pronoms de 3^{e} personne en bleu, etc.), alors l'utilisateur voit immédiatement, dans une seule représentation, la proportion de rouge et de bleu pour la chaîne à l'échelle du texte.

Figure 5 à insérer à peu près ici : Visualisation de chaînes de coréférence sur toute la longueur du texte avec GLOZZ.

De son côté, ANALEC propose des fonctionnalités similaires, avec deux exemples d'utilisation dans les figures 6 et 7. La figure 6 présente l'interface principale d'ANALEC pour l'annotation de corpus, interface qui permet de travailler sur des expressions référentielles et des chaînes de coréférence de manière souple, avec des raccourcis entre les deux. L'interface comprend, outre une barre de menus, trois zones principales, décrites ici du haut vers le bas : une première zone regroupe un ensemble de contrôles (boutons, choix déroulants) pour gérer les unités, les relations et les schémas définis dans la structure d'annotation ; une deuxième zone affiche le texte et utilise un code typographique (couleur, taille, gras, etc.) pour mettre en relief les annotations déjà réalisées ; et une troisième zone comprend l'interface d'annotation, générée automatiquement à partir de la structure d'annotation et comprenant, pour chaque trait, un menu déroulant permettant à l'annotateur de choisir parmi les valeurs possibles. Dans la capture d'écran de la figure 6, l'annotateur se trouve dans la situation suivante. Toutes les expressions référentielles ont été annotées. Le texte étant un résumé du roman *Les trois mousquetaires*, un exemple de référent est d'Artagnan, d'autres exemples étant Athos,

Milady, ou encore le groupe des quatre mousquetaires (groupe désigné notamment par « ces quatre hommes » dans la deuxième ligne du texte). Toutes les formes atténuées de référence ont également été annotées, notamment les sujets zéro de verbes à l'infinitif ou au participe. Par conséquent, tous les futurs maillons de chaînes de coréférence sont prêts. La figure ne porte que sur un seul référent, d'Artagnan. En rouge sur fond jaune apparaissent ainsi les formes explicites et les formes atténuées de référence à d'Artagnan, par exemple sur la première ligne du texte : « un gascon désargenté de 18 ans, d'Artagnan », et « faire », repérage choisi pour le sujet zéro de ce verbe à l'infinitif. Si seuls les maillons concernant d'Artagnan apparaissent, c'est parce que la chaîne de coréférence de ce référent est en cours de construction, et que le code couleur « rouge sur fond jaune » correspond à un maillon affecté à une chaîne de coréférence. L'annotateur est ainsi en train de vérifier, voire de compléter, une chaîne de coréférence déjà partiellement spécifiée. C'est pourquoi la première zone de l'interface correspond aux contrôles dédiés à la gestion d'une chaîne : création d'un schéma (chaîne), suppression d'un schéma, ajout d'un élément d'un schéma, suppression d'un élément d'un schéma, création d'un nouvel élément de schéma. Le schéma en cours de traitement est bien, comme cela apparaît figure 6, le schéma de type « coréférence » concernant le référent repéré en tant que « d'Artagnan ». L'annotateur n'a plus qu'à choisir un nouveau maillon – s'il en a oublié un – ou à choisir de retirer un maillon de la chaîne – s'il a par inadvertance intégré une expression référant à Athos ou Porthos. Enfin, les caractéristiques de la chaîne en question apparaissent dans l'interface d'annotation, où l'on peut voir par exemple que le trait « nombre » du référent se voit affecter la valeur « singleton », d'Artagnan étant un individu unique et non un groupe d'individus. Les menus déroulants permettent aussi de corriger ces traits caractérisant la chaîne de coréférence, et ANALEC autorise à tout moment non seulement un choix parmi les valeurs possibles telles que définies dans la structure d'annotation, mais aussi la saisie d'une nouvelle valeur, qui va alors mettre à jour automatiquement cette structure d'annotation. C'est utile quand on indique le nom du référent, ce trait, contrairement à tous les autres, dépendant du texte étudié.

Figure 6 à insérer à peu près ici : Interface d'annotation d'une chaîne dans ANALEC.

De plus, ANALEC propose (depuis peu) une fonctionnalité supplémentaire consistant à représenter la suite des mentions d'un texte par une succession de points auxquels sont affectés d'une part un chiffre correspondant au référent, d'autre part une couleur correspondant à l'une des données d'annotation. L'utilisateur choisit cette donnée, et donc le code couleur. Il peut ainsi confronter diverses représentations graphiques, les superposer, les exporter en tant qu'images dans des fichiers informatiques exploitables par ailleurs, etc. La

figure 7 présente l'interface de visualisation des chaînes, avec un code couleur (à droite de la copie d'écran) lié à la catégorie des mentions, et un affichage (en haut) mettant en perspective le découpage en paragraphes du texte. Il est à noter qu'en plus des données d'annotation, d'autres données ou calculs peuvent être sélectionnés en tant que code couleur. C'est notamment le cas du nombre de maillons des chaînes. Cette possibilité a fait apparaître lors d'une pré-étude d'un texte une corrélation significative entre découpage en paragraphe et types de personnages mentionnés : certains paragraphes ne faisaient apparaître quasiment que les personnages principaux, alors que les autres paragraphes ne faisaient apparaître quasiment que les personnages secondaires. C'est la perception immédiate des couleurs qui a permis de s'en rendre compte.

Figure 7 à insérer à peu près ici : Représentation de chaînes dans ANALEC.

Plusieurs représentations visuelles peuvent être générées, de manière à multiplier les possibilités de détection rapide de phénomènes intéressants. Par ailleurs, bien que cela ne soit pas très pratique dans ANALEC (alors que c'est le cas dans GLOZZ), on peut prévoir des filtres sur les mentions, de manière à ne faire apparaître qu'un sous-ensemble de phénomènes. Avec ces exemples, nous illustrons l'éventail des possibilités qui s'offre au linguiste pour optimiser ses recherches sur corpus. De plus en plus, les outils vont vers une diversification des rendus graphiques, de manière à s'adapter à un maximum d'utilisateurs, et vers une utilisation de plus en plus souple. A titre d'exemple, il est possible dans l'interface de la figure 7 de cliquer sur n'importe quel rond coloré (en haut) représentant un maillon de chaîne, et l'expression correspondante apparaît alors en gras dans le texte, ce qui permet – par un simple clic également – d'ouvrir directement l'interface d'annotation pour cette expression particulière. Autrement dit, les allers et retours entre l'annotation et la visualisation sont possibles à tout moment, pour ne pas dire encouragés.

4. Calculs sur les chaînes de coréférence

Cette section aborde la question des calculs réalisables sur des suites de données telles qu'elles apparaissent dans le haut de la figure 7, c'est-à-dire des suites de mentions (expressions référentielles et/ou formes atténuées) qui se caractérisent chacune par un référent et une propriété. Le référent est identifié par un nombre, qui correspond à l'ordre d'apparition des référents annotés : 1 pour le premier référent annoté, 2 pour le suivant, etc. La propriété est repérée à l'aide d'une couleur. Dans l'exemple de la figure 7, l'utilisateur regarde la fin de la nouvelle *Les bijoux* de Maupassant, plus exactement les paragraphes 59 à 68 (cf. le nombre indiqué tout en haut), qui font intervenir uniquement deux personnages : Mr. Lantin (chiffre

1) et un bijoutier (chiffre 7). La propriété observée est la catégorie des mentions, avec un code couleur qui affecte par exemple la couleur jaune aux possessifs et la couleur rouge aux GN définis. La suite de points colorés montrée figure 7 est donc une succession de chiffres corrélée avec une succession de couleurs, qui pourrait aussi se représenter sous la forme suivante : (1, GN défini) ; (7, possessif) ; (7, GN défini) ; (1, possessif) ; etc. Cette fois, dans la mesure où notre objet d'étude est un code simple, composé uniquement d'étiquettes d'analyse (et non de matériau linguistique), nous nous plaçons dans le cadre de la linguistique outillée et non plus dans celui de la linguistique à l'instrument.

Au premier abord, il existe deux façons très simples de gérer une telle suite codée. La première consiste à ne tenir compte que du nombre correspondant au référent, et on étudie alors la suite des références du texte, dans l'ordre dans lequel elles apparaissent. La seconde consiste à fixer un référent, donc à ne tenir compte par exemple que du chiffre 7, et à s'intéresser à la succession des catégories des mentions.

Plus précisément, concernant la première façon de procéder : avec l'exemple de la figure 7, on obtient la suite 1 7 7 1 1 7 1 7 1 1 7 1… Cette suite, qui peut paraître un peu abstraite au premier abord, peut en fait s'avérer riche d'informations sur les continuations et les transitions référentielles. Elle peut permettre notamment de détecter les parties d'un texte qui se focalisent sur un seul référent, les parties comme c'est le cas ici où deux référents alternent continuellement, ou encore les configurations telles que 1 1 1 7 1 7 7 7 où un référent disparaît au profit d'un autre. Il s'agit en fait d'un message écrit dans un alphabet où chaque lettre correspond à un référent. Comme tout message, celui-ci peut être analysé, et notamment analysé en termes de N-grammes. Avec la valeur 3 pour N, l'analyse revient à regarder toutes les successions de 3 lettres consécutives, c'est-à-dire 1 7 7, puis 7 7 1, puis 7 1 1, etc. On décompte chaque triplet observé, et on obtient une caractérisation du texte en triplets représentatifs. Par exemple, le triplet le plus fréquent peut être 1 1 1, ce qui montre alors que les continuations sur le personnage de Mr. Lantin sont les phénomènes référentiels les plus fréquents, ce qui apporte un argument quantitatif au fait que Mr. Lantin est le personnage principal. Nous n'irons pas plus loin dans cette voie ici, mais nous voulons mettre en avant deux points : d'une part, qu'une fois que l'on dispose de données annotées, il est possible de multiplier les approches mathématiques ou statistiques sur ces données (et c'est au linguiste à faire le lien entre les résultats chiffrés obtenus et des notions linguistiques telles que les continuations ou transitions référentielles) ; d'autre part que, ici aussi, la linguistique de corpus outillée est riche en perspectives… Les calculs les plus simples ne nécessitent aucun outil particulier, mais ne doivent pas pour autant être oubliés : nombre de maillons de la

chaîne en cours d'analyse, répartition des maillons par chapitre ou paragraphe, taille moyenne des intervalles entre deux maillons, etc. Il s'agit de comptages sur la structure même de la chaîne de coréférence, sans tenir compte des annotations, comptages qui peuvent être pris en compte en tant que paramètres dans tout type de calcul, ne serait-ce qu'en tant que pondérations.

De même, concernant la deuxième façon de procéder, c'est-à-dire en sélectionnant un référent, on se retrouve alors à étudier une chaîne de coréférence en particulier, et à regarder de quoi elle se compose : proportion de possessifs, proportion de GN définis, alternance de telle ou telle catégorie de mention et ainsi de suite. C'est ce qui est montré en figure 8, dans laquelle on voit une représentation graphique, produite par MS Excel, de la répartition des types d'expressions référentielles pour les personnages les plus cités dans *Les bijoux*, à partir d'un tableau de corrélation généré dans ANALEC et exporté en CSV (format texte compatible avec un tableur). On peut voir par exemple que les deux bijoutiers sont mentionnés beaucoup plus souvent à l'aide de GN définis que les personnages principaux de la nouvelle. On pourrait également appliquer aux représentations codées des chaînes de coréférence des calculs de N-grammes tels que vus précédemment, l'alphabet étant cette fois composé de codes correspondant aux différentes catégories de mentions, ou encore aux différentes fonctions syntaxiques, aux différents rôles thématiques, etc.

Figure 8 à insérer à peu près ici : Analyse de chaînes avec Excel.

Conclusion

Nous avons montré comment l'étude de la référence, de l'anaphore et de la coréférence pouvait être opérationnalisée grâce aux avancées de la linguistique de corpus, qu'il s'agisse de linguistique à l'instrument ou de linguistique outillée. Il est possible de déterminer à partir des nombreuses études linguistiques portant sur de tels phénomènes, un ensemble vaste et hétérogène de questions, de tests, d'idées de calculs statistiques. Pour trouver des réponses, nous disposons de quelques logiciels performants : des instruments tels que GLOZZ et ANALEC, des outils tels que les tableurs et les analyseurs statistiques. En général, ce n'est pas un logiciel mais la combinaison de plusieurs logiciels qui permettra de procéder à un maximum de calculs et de répondre à un maximum de questions. Comme nous l'avons évoqué dans la section 2.2 et montré ensuite avec des allers et retours constants entre GLOZZ, ANALEC et Excel, la solution ne se trouve que rarement dans l'exploitation d'un logiciel unique, mais plutôt dans la mise en œuvre d'une cascade de logiciels.

Certes, les résultats des quelques analyses que nous avons présentés ne sont pas révolutionnaires. Alternance des mentions des personnages d'un texte narratif, pertinence du découpage en paragraphes, matérialisation des notions de continuation et de transition référentielle : il n'y a rien là de très surprenant. L'important, c'est d'offrir avec un ensemble d'instruments et d'outils des arguments objectifs, chiffrés ou non, pour étayer les hypothèses linguistiques formulées sur les expressions référentielles et sur les chaînes de coréférence. C'est ce que nous avons voulu montrer avec l'ensemble des représentations graphiques illustrant cet article, et c'est la voie que nous avons choisie pour contribuer au développement d'ANALEC (Landragin *et al.*, 2012).

Concernant les aspects purement techniques, notre étude de la coréférence a fait apparaître un certain nombre de besoins qui ne sont pas encore complètement satisfaits avec les outils existants. Entre autres aspects, nous mentionnerons ici quelques exemples portant sur ANALEC : développement d'une boîte à outils pour la gestion dynamique des schémas ; multiplication des visualisations possibles, avec à chaque fois la possibilité d'éditer n'importe quel élément visuel afin de modifier les annotations des extraits de corpus correspondants ; extension des fonctionnalités d'importation et d'exportation de corpus pour faciliter la mise en œuvre d'outils en cascade. La multiplication des visualisations possibles passe par l'augmentation du nombre de paramètres ayant une incidence sur la représentation visuelle, par la diversification des codes visuels (pas seulement un code couleur, mais aussi un code de taille, de forme, de teinte, de saturation, etc.), et implique également des fonctionnalités d'exportation (format CSV pour les codes eux-mêmes, formats EMF et SVG pour les images correspondant aux représentations graphiques). On disposera alors de possibilités d'analyse plus nombreuses et plus variées, et, accessoirement, d'illustrations plus claires et plus ciblées pour des articles de recherche.

Remerciements : trois personnes en particulier ont contribué à rendre possible ce travail. Il s'agit de Michel Charolles, qui a porté à nos yeux un ensemble de problématiques pertinentes sur la coréférence, de Bernard Victorri, qui a développé pendant plusieurs années le logiciel ANALEC avec un souci constant des retours d'expérience d'utilisateurs linguistes, et de Noalig Tanguy qui, à travers une tâche d'annotation de textes en références et coréférences, a permis de donner un nouvel élan à nos réflexions méthodologiques dans le domaine de la linguistique de corpus. Plus largement, ce travail est aussi issu des réflexions et des avancées du projet PEPS « MC4 : Modélisation Contrastive et Computationnelle des Chaînes de Coréférence ».

Références bibliographiques

Albert, S., Anderssen, J., Bader, R., Becker, S., Bracht, T., Brants, S., Brants, T., Demberg, V., Dipper, S., Eisenberg, P., Hansen, S., Hirschmann, H., Janitzek, J., Kirstein, C., Langner, R., Michelbacher, L., Plaehn, O., Preis, C., Pußel, M., Rower, M., Schrader, B., Schwartz, A., Smith, G., & Uszkoreit, H. (2003). TIGER Annotation Schema. Technical report, Universities of Saarbrücken, Stuttgart, and Potsdam.

Avanzi, M., Lacheret-Dujour, A., & Victorri, B. (2008). ANALOR, A Tool for Semi-Automatic Annotation of French Prosodic Structure, In: *Proceedings of the 4th Conference on Speech Prosody*.

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

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. *Traitement Automatique des Langues*, 49(2), 217-248.

Cunningham, H., Maynard, D., Bontcheva, K., & Tablan, V. (2002). GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. *Proceedings of the 40th Meeting of the Association for Computational Linguistics*, 168-175.

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

Grosz, B., Joshi, A., & Weinstein, S. (1995). Centering: a Framework for Modeling the Local Coherence of Discourse. *Computational Linguistics*, 21(2), 203-225.

Habert, B. (2005). Portrait de linguiste(s) à l'instrument. *Revue en ligne Texto!*, 104, www.revue-texto.net/Corpus/Publications/Habert/Habert_Portrait.html.

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

Landragin, F., Poibeau, T., & Victorri, B. (2012). ANALEC: A New Tool for the Dynamic Annotation of Textual Data. *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, Istanbul, 357-362.

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

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

Mathet, Y., & Widlöcher, A. (2012). Glozz Annotation Platform. http://www.glozz.org/.

Morton, T., & LaCivita, J. (2003). WordFreak: An Open Tool for Linguistic Annotation. *Proceedings of Human Language Technology (HLT) and North American Chapter of the Association for Computational Linguistics (NAACL)*, 17-18.

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

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

Schnedecker, C. (2011). La notion de saillance : problèmes définitoires et avatars. In O. Inkova (Ed.), *Saillance. Aspects linguistiques et communicatifs de la mise en évidence dans un texte* (pp. 23-43). Besançon : Presses Universitaires de Franche-Comté.

Steiner, I., & Kallmeyer, L. (2002). VIQTORYA – A Visual Query Tool for Syntactically Annotated Corpora. *Proceedings of the Third International Conference on Language Resources and Evaluation*, Las Palmas, 1704-1711.

Van Deempter, K., & Kibble, R. (2000). On Coreferring: Coreference in MUC and Related Annotation Schemes. *Computational Linguistics*, 26(4), 629-637.

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

Victorri, B. (2012). ANALEC : logiciel d'annotation et d'analyse de corpus écrits. Logiciel téléchargeable sur le site Web du laboratoire Lattice, http://www.lattice.cnrs.fr/ANALEC.

Widlöcher, A., & Mathet, Y. (2009). La plate-forme Glozz : environnement d'annotation et d'exploration de corpus. *Actes de la 16e Conférence sur le Traitement Automatique des Langues Naturelles*, Senlis.

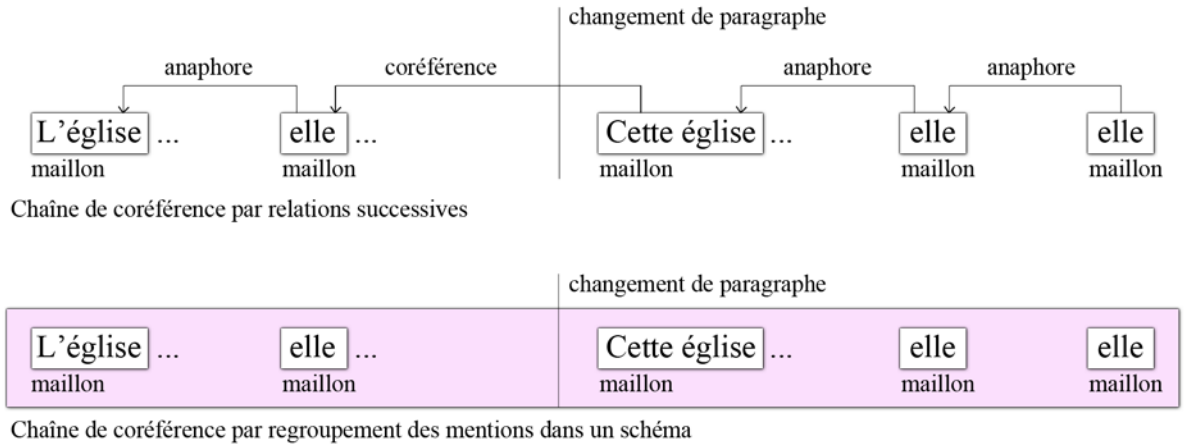

Figure 1. Chaîne par relations successives ou par regroupement.

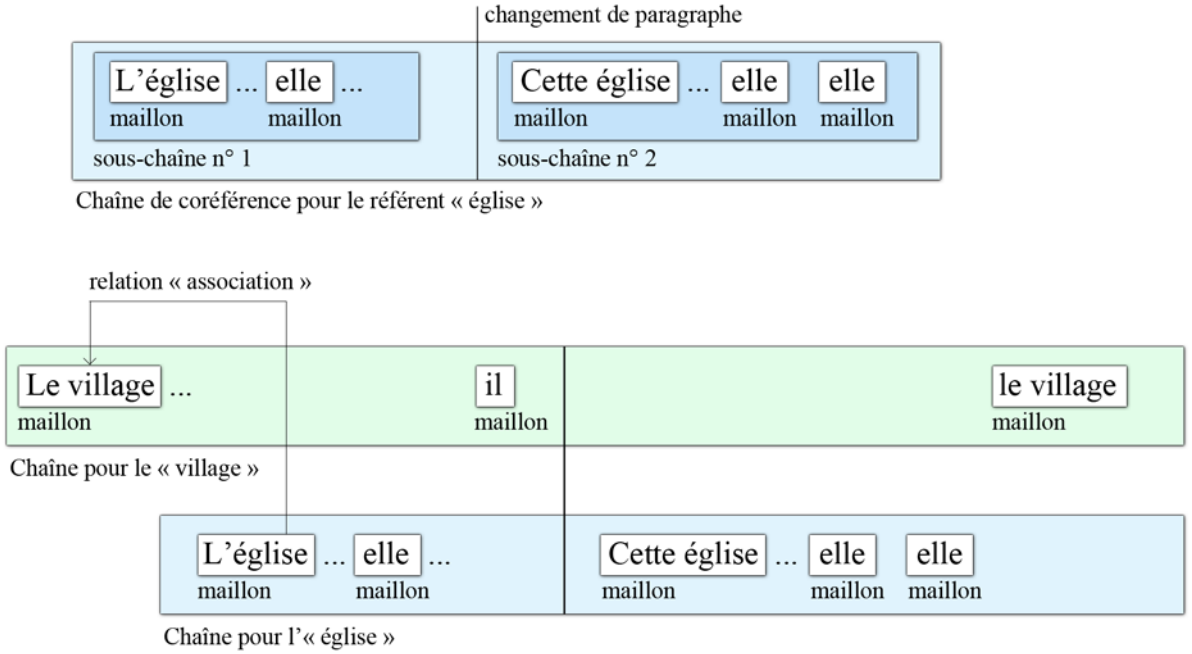

Figure 2. Chaînes de coréférence, sous-chaînes et relations anaphoriques.

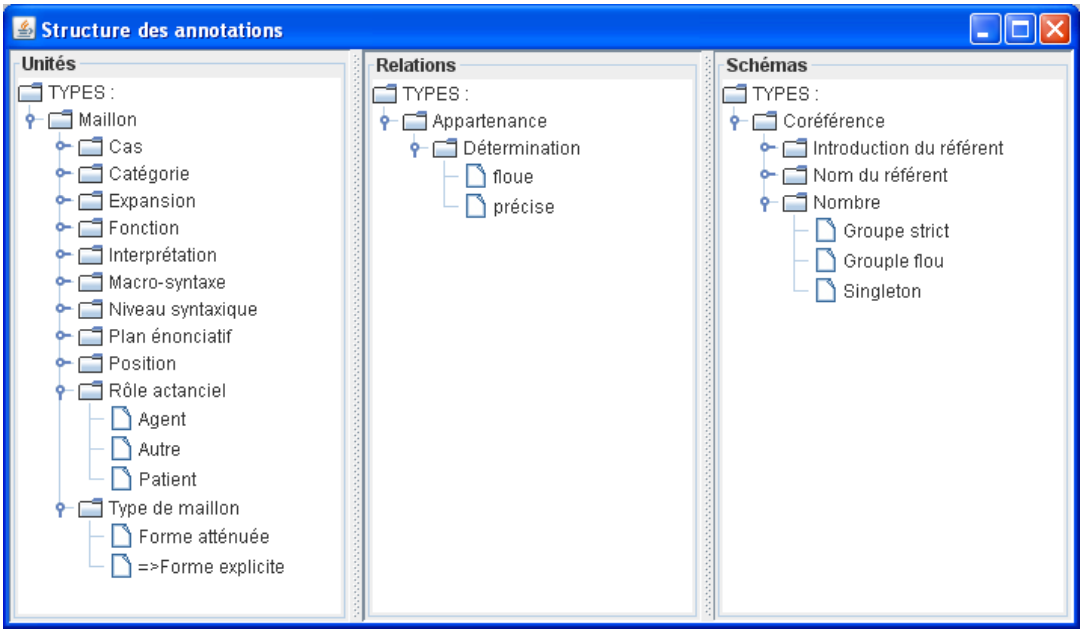

Figure 3. Exemple de structure d’annotation.

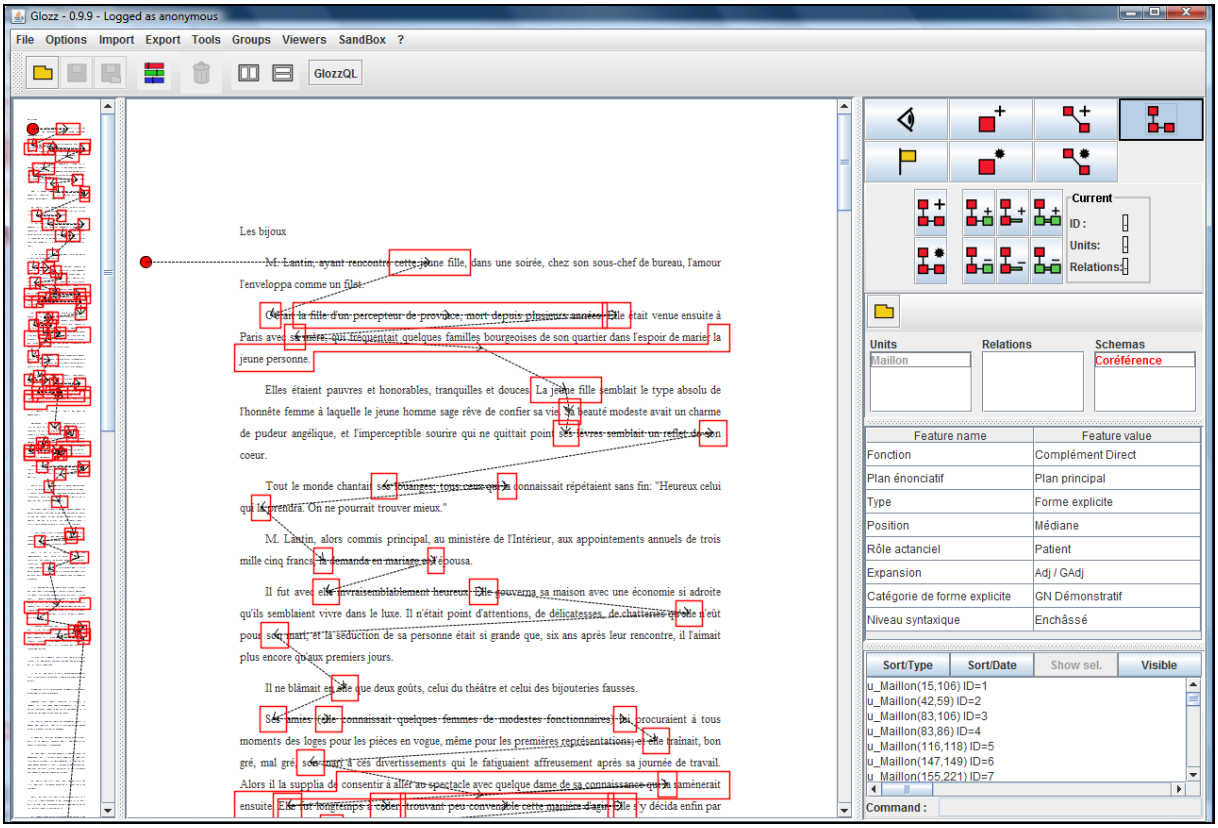

Figure 4. Interface principale de GLOZZ.

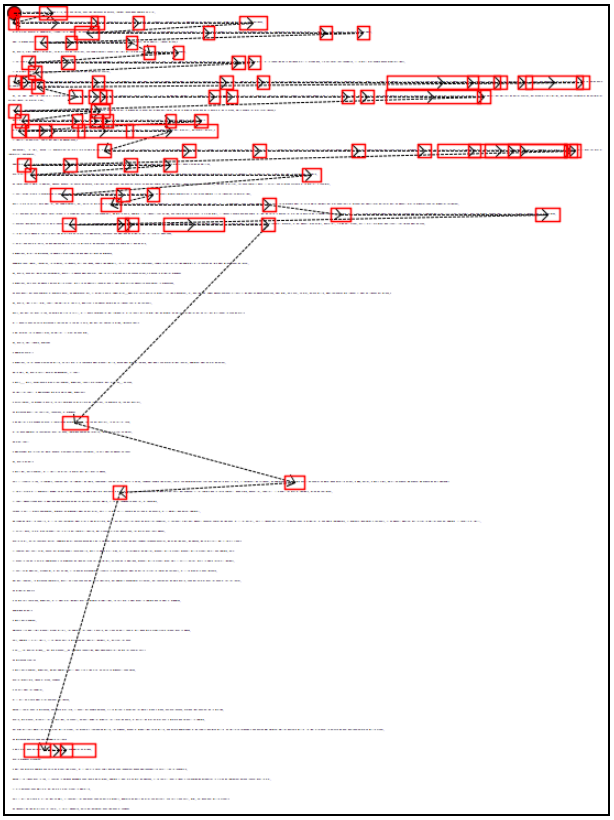

Figure 5. Visualisation de chaînes de coréférence sur toute la longueur du texte avec GLOZZ.

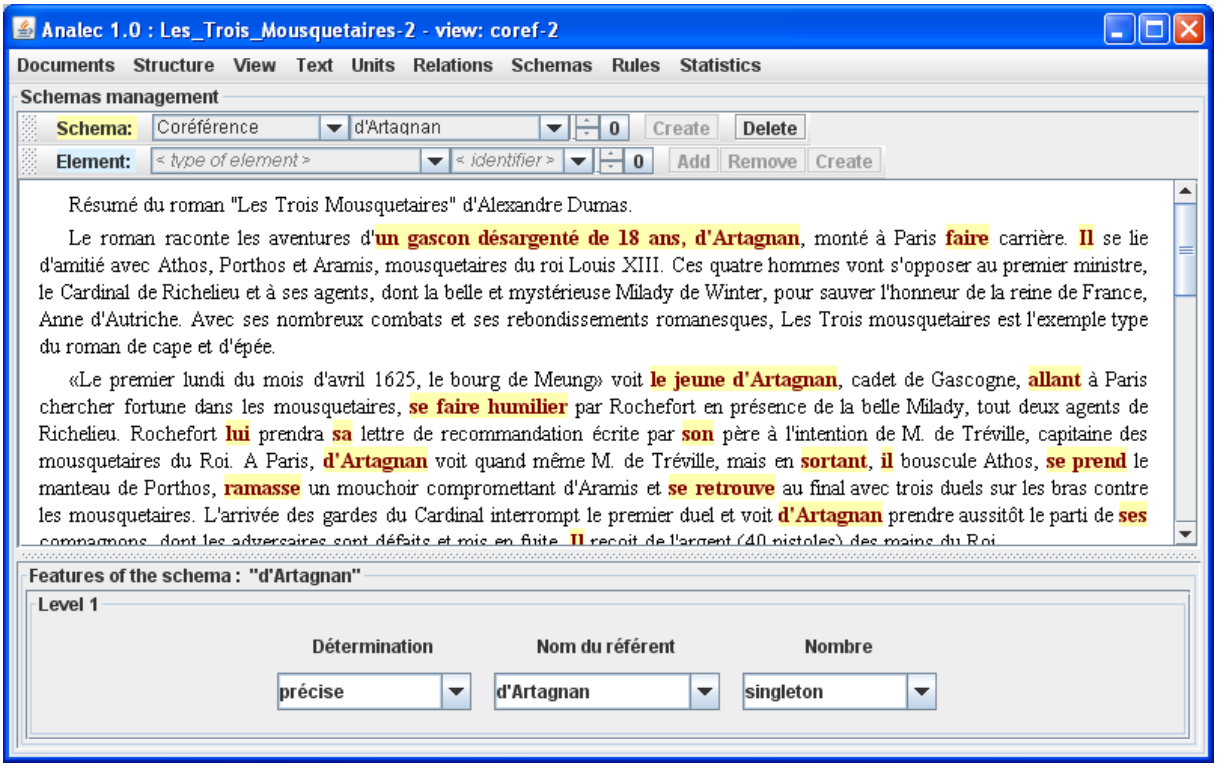

Figure 6. Interface d'annotation d'une chaîne avec ANALEC.

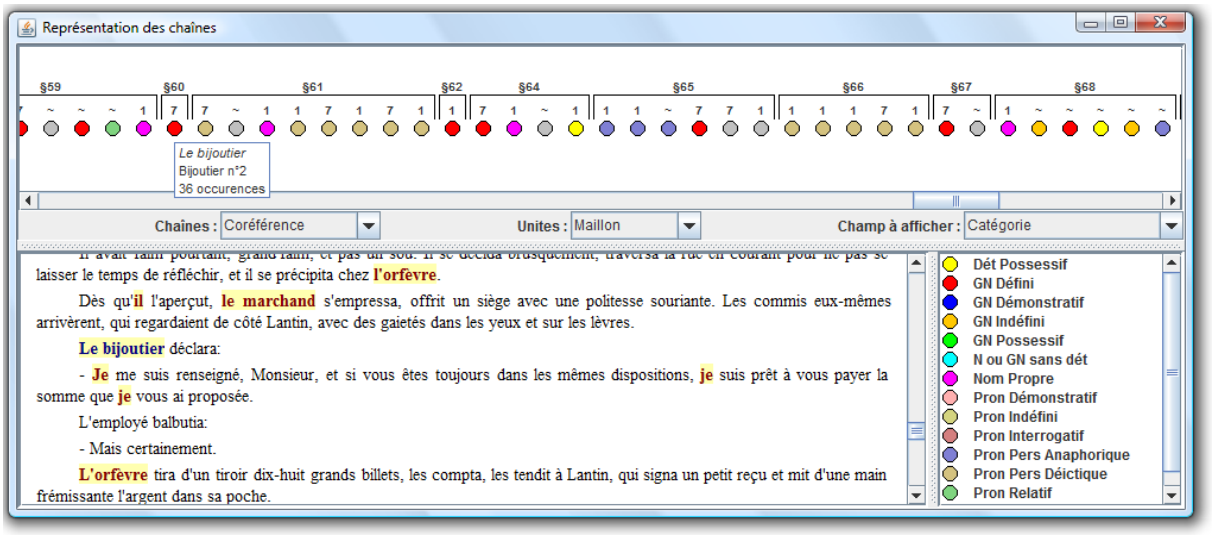

Figure 7. Représentation de chaînes dans ANALEC.

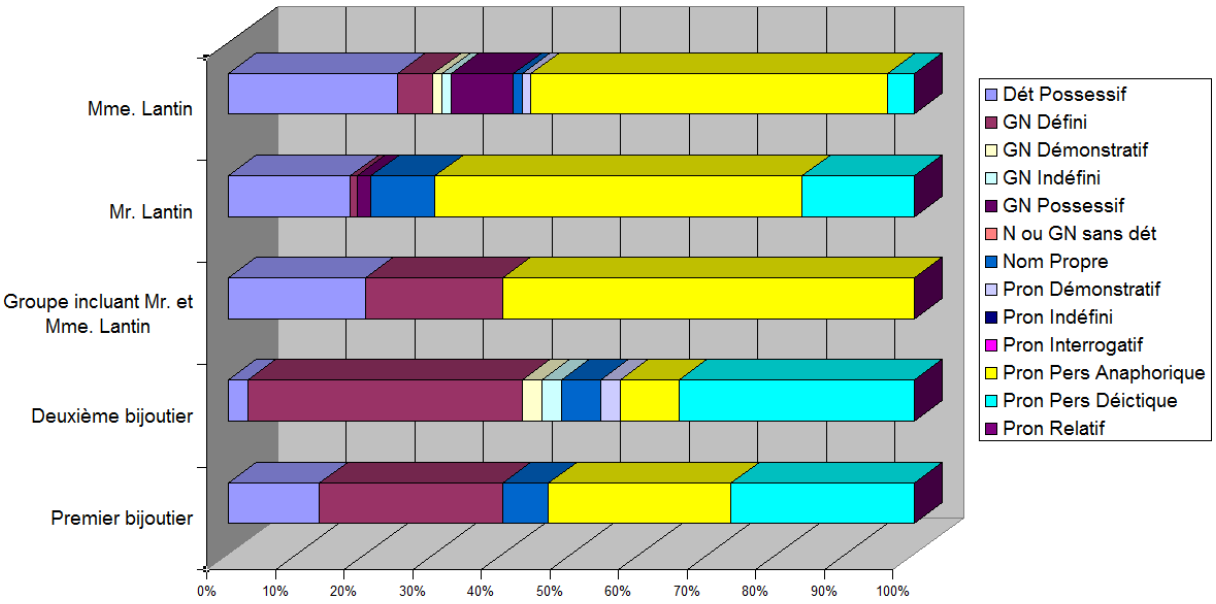

Figure 8. Analyse de chaînes avec Excel.