Corpus
10 (2011)
Varia

Frédéric Landragin

Une procédure d'analyse et d'annotation des chaînes de coréférence dans des textes écrits

Référence électronique
Frédéric Landragin, « Une procédure d'analyse et d'annotation
des chaînes de coréférence dans des textes écrits », *Corpus* [En ligne], 10 | 2011, mis en ligne le 08 juin 2012, consulté le 16 janvier 2013. URL : http://corpus.revues.org/2010

Éditeur : Bases, corpus et langage - UMR 6039
http://corpus.revues.org
http://www.revues.org

Document accessible en ligne sur : http://corpus.revues.org/2010
Ce document est le fac-similé de l'édition papier.

Une procédure d'analyse et d'annotation des chaînes de coréférence dans des textes écrits

Frédéric LANDRAGIN
CNRS, LaTTICe (Montrouge)[1]

1. Introduction

Cet article présente une étude en cours sur la référence à des personnes et à des groupes de personnes, étude dont l'objectif principal est d'analyser les chaînes de coréférence dans des textes écrits et de mettre en œuvre une méthodologie d'annotation des phénomènes référentiels et coréférentiels, afin de constituer ultérieurement un corpus annoté autour de ces phénomènes. Cette étude est menée dans le groupe de travail « coréférence » du laboratoire LaTTICe[2] et cet article en constitue une première synthèse écrite.

La procédure générale est la suivante : étude de textes narratifs courts (résumés de films) et plus longs (nouvelles, articles de presse) ; repérage des personnages ; délimitation des éléments linguistiques qui participent à la référence et à la coréférence ; annotation manuelle de ces éléments avec des données sémantiques et pragmatiques ; construction des chaînes de coréférence en regroupant les éléments relatifs au même référent ; affectation de relations entre les chaînes obtenues ; annotation des chaînes et des relations avec des données pragmatiques liées notamment à l'interprétation. L'objectif à terme est de constituer un corpus de référence annoté finement en phénomènes référentiels et coréférentiels, de manière à

1 Laboratoire mixte CNRS / ENS / Université Paris 3, UMR 8094.

2 Nous remercions les participants passés et présents de ce groupe de travail que nous animons depuis maintenant deux ans, et nous remercions en particulier Michel Charolles et Bernard Victorri dont les propositions ont largement contribué à faire progresser la réflexion collective.

proposer à la communauté un ensemble de données permettant de tester des hypothèses sur des questions telles que : comment les chaînes de coréférence sont-elles constituées (typologie) ? à quel moment un démonstratif apparaît-il ? quelle est la fréquence d'apparition des noms propres ? quelles sont la nature et la fréquence des expansions dans les expressions référentielles ? quels sont les rôles thématiques privilégiés au début des chaînes de coréférence ? peut-on prévoir des motifs dans la manière dont les chaînes de coréférence se croisent ? y a-t-il une corrélation entre un genre textuel et un type de chaîne ? y a-t-il une corrélation entre la typologie des chaînes et le découpage du texte en paragraphes ? etc. (de nombreuses autres questions ont été imaginées), et, de manière plus personnelle, quels sont les facteurs de saillance qui mettent en avant une chaîne de coréférence parmi d'autres ? Nous sommes conscient de l'aspect chronophage du travail d'annotation, mais aussi de la richesse des données que ce travail permettra d'obtenir et par conséquent de la multitude des hypothèses qui pourront être testées. En attendant, l'objectif à plus court terme est d'affiner la procédure d'annotation et d'optimiser les outils pour l'annotation et l'interrogation d'un tel corpus. Par exemple, la recherche de corrélations entre des données annotées (détermination, rôle thématique) et des données déduites d'annotations (typologie des chaînes) nécessite des outils capables de déduire de nouvelles données à partir de celles annotées, ainsi que des outils permettant aux linguistes de repérer efficacement – de manière visuelle par exemple (Victorri 2011) – des régularités.

Dans cet article, nous nous focalisons sur les questions qui se sont posées lors de la mise en place de notre méthodologie d'annotation, avec deux préoccupations : nous nous sommes demandé d'une part comment exploiter et améliorer les méthodes d'annotation usuelles (Habert *et al.* 1997 ; van Deempter & Kibble 2000) pour qu'elles incluent des aspects sémantiques et pragmatiques fins (Corblin 1995 ; Schnedecker 1997 ; Cornish 1999 ; Kleiber 2001 ; Charolles 2002), et d'autre part comment les méthodologies de la linguistique de corpus influent sur notre façon d'appréhender les phénomènes de

référence, de coréférence, de transition référentielle et de saillance. Parmi les questions qui se sont ainsi posées, les cinq suivantes constituent les sections 2 à 6.

Section 2 : compte tenu de la nature sémantique et pragmatique des phénomènes de référence, comment modéliser les biais interprétatifs, les ambiguïtés et les sous-déterminations qui se présentent forcément à ce niveau d'analyse de la langue ? Comment ces modélisations sont-elles influencées par les procédures d'annotation ? Comment adapter ces modélisations pour leur intégration dans un schéma d'annotation ?

Section 3 : comment intégrer dans les annotations les identifiants des référents humains et leur appartenance stricte ou floue à divers groupes, collectifs et institutions dont il peut être question, directement ou indirectement, dans le texte ? Les méthodes d'annotation ont-elles un effet sur notre façon de modéliser les relations entre référents et les relations d'appartenance d'un référent à un autre ?

Section 4 : comment modéliser les éléments non référentiels qui participent malgré tout à la coréférence, notamment certains sujets zéro, certaines constructions pronominales, certaines constructions attributives ? Comment les procédures d'annotation peuvent-elles gérer ce type de phénomènes (éléments sans trace linguistique, par exemple) ? Comment intégrer ces éléments dans un schéma d'annotation compte tenu de leur contribution particulière à la coréférence ?

Section 5 : comment séparer ce qui relève nécessairement d'une annotation manuelle de ce qui peut être déduit automatiquement d'annotations déjà effectuées, en particulier quand on s'intéresse à des phénomènes et des opérations linguistiques sans marqueur direct, notamment la saillance des référents et les transitions référentielles ? Quelles sont les influences des procédures d'annotation existantes dans l'élaboration d'une telle méthodologie ?

Section 6 : comment matérialiser les points précédents, premièrement avec le développement d'un outil adapté à l'annotation de chaînes de coréférence, et deuxièmement avec un schéma et un manuel d'annotation compatibles avec cet outil ? Cette dernière section sur les aspects logiciels nous permettra de

conclure sur la nature et l'efficacité de la chaîne des traitements que nous envisageons actuellement pour l'analyse des chaînes de coréférence d'un texte écrit.

2. Biais interprétatifs, ambiguïtés et sous-déterminations

Si les ambiguïtés et autres phénomènes de flou dans l'interprétation d'un texte par un sujet humain apparaissent à tous les niveaux d'analyse du langage, par exemple en syntaxe, c'est peut-être en sémantique et en pragmatique qu'ils sont les plus subjectifs, et parfois les plus difficiles à détecter. Des études faites par Charolles (2002) avec des résumés de films tels que celui ci-dessous montrent que d'un lecteur à l'autre, et ce même pour des lecteurs linguistes (du moins étudiants en linguistique), un pronom va référer à tel ou tel autre personnage du texte selon l'interprétation réalisée.

Même quand on axe la lecture sur la recherche des ambiguïtés et des sous-déterminations, les listes d'alternatives diffèrent d'un lecteur à l'autre. Attribuer un référent à une expression référentielle peut ainsi s'avérer difficile, voire impossible. Plus que cela, l'attribution réalisée de manière immédiate peut s'avérer arbitraire, éventuellement justifiée du point de vue de la langue, mais fausse du point de vue de la réalité décrite. Ce biais entre le monde des référents qui est décrit et les interprétations réalisées par divers lecteurs a une conséquence majeure : plusieurs annotateurs d'un même texte risquent d'aboutir à des résultats différents, ce qui va entraîner un corpus de mauvaise qualité compte tenu des standards actuels concernant la constitution de corpus – cf. par exemple Habert *et al*. (1997) pour les méthodes de constitution de corpus et les calculs d'accord inter-annotateurs, ou Denis (2008) pour une présentation des métriques d'évaluation utilisées pour les phénomènes de coréférence.

> Eric Masson, un « demi-sel », est devenu l'amant de la belle Solange Mideau, femme d'un graveur raté. Eric veut se servir de Robert Mideau pour monter, à **son** insu, un trafic de fausse monnaie. Il s'associe à Charles Lepicard, tenancier d'une ancienne maison close, et à

> Lucas Malvoisin, l'homme d'affaires de celui-ci. Charles et Lucas n'ont pas grande confiance en Eric, mais Solange **leur** promet **son** concours. Elle souhaite en effet mener la grande vie. Avec l'accord de **ses complices**, Charles contacte Ferdinand Maréchal, dit le Dabe, vieux truand célèbre qui s'est retiré dans une île des Tropiques. Il le décide à venir à Paris. (résumé du film *Le cave se rebiffe* – nous soulignons les noms des personnages et grasseyons les références ambiguës)

En regardant l'exemple ci-dessus, on constate après plusieurs lectures et un certain temps de réflexion que le possessif de « à son insu » peut référer aussi bien à Robert Mideau qu'à Solange Mideau (il faut quasiment voir le film pour vérifier qu'il s'agit de Robert Mideau, ce qui correspond généralement à la première intuition de lecture). Même remarque pour le possessif de « son concours », complètement ambigu entre Solange Mideau et Eric Masson. Au niveau des références plurielles, on constate qu'il est parfois difficile de délimiter un groupe de personnes : « leur » inclut forcément Charles Lepicard et Lucas Malvoisin, mais peut aussi inclure Eric Masson ; « ses complices » inclut forcément Lucas Malvoisin, mais d'une part peut aussi inclure Solange Mideau, et d'autre part pourrait inclure d'autres personnages, éventuellement des personnages non cités dans le résumé.

Si l'on veut annoter les références d'un texte, ces questions doivent être posées. A chaque personnage mentionné plusieurs fois dans le texte correspond une chaîne de coréférence. Pour spécifier le référent d'une expression référentielle, on la fait appartenir à la bonne chaîne de coréférence. Les chaînes se construisent ainsi, par une succession d'appartenances. Or nous constatons qu'une expression référentielle peut pointer non pas sur un référent unique et bien déterminé, mais d'une part sur une alternative entre plusieurs référents possibles (ambiguïté), d'autre part sur un ensemble non délimité qui inclut un ensemble de référents clairement identifiés. A ces deux modélisations s'en ajoute une troisième, suscitée par l'étude du texte de presse suivant :

> L'ancien président de la République de Côte d'Ivoire, Henri Konan Bédié et son épouse ont reçu à dîner l'ancien Premier ministre Alassane Dramane Ouattara et son épouse, le 23 septembre. La rencontre très médiatisée avait un objectif, celui de montrer que **les héritiers du premier président de Côte d'Ivoire** peuvent se retrouver pour reconquérir le pouvoir. Les deux leaders ont l'habitude de se voir et de s'appeler depuis le déclenchement, le 19 septembre 2002 de la rébellion en Côte d'Ivoire. A Paris, à Abidjan, à Accra, les deux hommes se côtoient, mais dans des cadres formels. Leur rencontre en soi n'est donc pas un événement, sauf qu'ils ont voulu donner à cette entrevue un cachet particulier. Les retrouvailles autour d'un même idéal politique que commande la mémoire du « Vieux » dont ils se réclament. […] Mais après que tout le monde ait perdu le pouvoir, en faveur d'un autre héritier, le général Robert Guéi, par un coup d'Etat en décembre 1999, la gestion du pays semble échapper aux « enfants ». (nous soulignons les mentions de certains personnages et grasseyons la référence problématique)

L'expression référentielle « les héritiers du premier président de Côte d'Ivoire » inclut forcément les deux personnages dont il est question dans ce communiqué de presse, à savoir HKB (Henri Konan Bédié) et ADO (Alassane Dramane Ouattara). De manière beaucoup moins évidente, elle peut également inclure les épouses de ces deux hommes, puisqu'il en est fait mention dans le même temps, à l'aide de coordinations (indices de groupes). Nous passerons sur l'ambiguïté artificielle de la coordination « L'ancien président de la République de Côte d'Ivoire, Henri Konan Bédié et son épouse » qui pose une alternative entre deux ou trois référents. Lors de la première lecture de ce texte, l'interprétation est globalement dirigée vers la construction d'un groupe qui comprend deux « héritiers », HKB et ADO, notamment avec les expressions « les deux leaders » et « les deux hommes ». Or, quand on continue la lecture, on arrive à l'expression référentielle « un autre héritier, le général Robert Guéi » qui remet en question la constitution

de ce groupe. Même pour un lecteur averti qui dispose de toutes les connaissances encyclopédiques nécessaires, l'expression « les héritiers du premier président de Côte d'Ivoire » reste problématique en raison du style même du texte.

Un lecteur peut ainsi remettre en question l'attribution d'un référent au fur et à mesure de la lecture du texte, autrement dit nous assistons à des phénomènes comparables à celui des référents évolutifs. Dans le cas de notre exemple, l'évolution de la référence du groupe des héritiers correspond cependant plus à des faits linguistiques et interprétatifs qu'à des faits réels comme ceux du poulet qu'on découpe en morceaux avant de le cuire. Dans notre cas, il n'est pas pertinent d'ajouter à chaque annotation un repère temporel de manière à modéliser les évolutions. En fait, trois stratégies d'annotation sont possibles :

1 Stratégie linéaire (ou « linguistique ») : on se focalise sur les formes linguistiques, sans tenir compte des éventuelles ré-interprétations ultérieures. Cette stratégie amène à considérer les héritiers HKB et ADO.

2 Stratégie « réaliste » : on se focalise sur les concepts, et on n'annote qu'après avoir compris l'ensemble du texte et calculé toutes les références. Cette stratégie amène à considérer les héritiers HKB, ADO et Robert Guéi dès la mention « les héritiers du premier président […] ».

3 Stratégie adaptée à l'exploitation de corpus : on part des concepts et on élargit ponctuellement aux diverses interprétations possibles, de manière à rendre compte des biais interprétatifs et des effets stylistiques. Cette stratégie amène à procéder à une double annotation de l'expression « les héritiers […] » : une annotation qui décrit la référence au groupe HKB, ADO et Robert Guéi, et une annotation qui décrit l'interprétation initiale (et temporaire) aux seuls HKB et ADO.

Les procédures d'annotation classiques, en morphologie par exemple, n'offrent pas de solution face à ce problème : les phénomènes dont elles traitent sont majoritairement locaux et ne font pas autant appel aux connaissances et à la réflexion de l'annotateur. Avec les linguistiques de corpus, nous sommes orientés par la nécessité de délimiter des unités textuelles et

d'attribuer des traits (ou propriétés, ou annotations) à ces unités. Notre modélisation est influencée par cette nécessité, et, pour ne pas perdre l'essence de la discussion présentée dans cette section, nous adoptons la troisième stratégie. Cette stratégie présente deux avantages : celui de décrire avant tout la réalité des références (ce qui sera utile pour les traitements automatiques ultérieurs, la construction d'ontologies, etc.) et celui de tenir compte de la façon dont le texte est écrit et dont les personnages sont introduits. Elle présente un inconvénient majeur : celui de rendre la procédure d'annotation complexe, en décrivant dans un manuel d'annotation quels sont les cas pour lesquels l'annotation qui décrit l'interprétation temporaire est importante, et en demandant à l'annotateur de lire l'intégralité du texte avant de commencer à l'annoter…

3. Les référents et leur appartenance à des groupes

Les pluriels et les groupes d'individus impliqués construisent des entités du discours au même titre qu'une expression référant à un individu unique. Annoter un pluriel consiste donc à ajouter un pointeur vers un groupe caractérisé par un identifiant. Dans nos précédents exemples, en plus des personnages, on avait ainsi d'autres référents qui étaient : l'ensemble des interlocuteurs de Solange Mideau (pour « leur » dans « Solange leur promet ») ; l'ensemble des complices de Charles (pour « ses complices ») ; et dans le texte de presse, l'ensemble des héritiers du premier président de Côte d'Ivoire. Nous avons choisi cette méthode de manière à construire efficacement les chaînes de coréférence relatives à ces groupes d'individus. Pour le texte de presse en particulier, considérer l'ensemble des héritiers comme un référent en soi permet de lier les trois expressions « les héritiers […] », « les deux leaders » et « les deux hommes » en une même chaîne de coréférence, qui s'ajoute aux chaînes de coréférence spécifiques à HKB ou à ADO, permettant ainsi de modéliser de manière exhaustive les transitions référentielles d'un individu au groupe auquel il appartient et inversement.

Un individu qui appartient à deux groupes est ainsi pris en compte dans trois chaînes de coréférence : celle qui le

concerne individuellement et les deux correspondant aux deux groupes. Pour compléter cette manière de modéliser, il est nécessaire de considérer des relations d'appartenance. Pour ne pas multiplier ces relations, on considère qu'elles sont transitives. Si dans les deux groupes que nous venons d'évoquer le premier est inclus dans le second, on aura une relation d'appartenance entre le premier groupe et le second (dans ce sens), plus une relation d'appartenance entre l'individu et le premier groupe.

Enfin, pour prendre en compte les limites souvent non précisées de certains groupes (celui des complices comme celui des héritiers), nous intégrons la notion de flou, d'une part pour la délimitation des groupes (groupe strict *versus* groupe flou), d'autre part pour la relation d'appartenance d'un individu (ou groupe) à un groupe (appartenance stricte *versus* floue). La modélisation suit celle de la Théorie des Ensembles Flous, avec pour ce qui concerne la délimitation des groupes :

- groupe strict : pour un groupe construit par une coordination, par exemple « Charles et Lucas » dans le résumé de film, on est en présence d'un groupe strict, au cardinal connu et définitif ;
- groupe flou : pour un groupe construit par une dénomination telle que « ses complices » ou « les héritiers […] » et dans le cas où il est difficile d'identifier les référents dont il est question, on est en présence d'un groupe flou, dont le cardinal peut varier selon les interprétations, peut évoluer au long du texte, ou peut rester impossible à chiffrer.

Et pour ce qui concerne la relation d'appartenance :

- relation stricte : quand le référent HKB, ADO ou Robert Guéi est clairement établi comme un membre du groupe en question, la relation d'appartenance est stricte (pour Robert Guéi, le décalage stylistique est pris en compte par ailleurs, dans l'annotation de l'expression référentielle « les héritiers […] ») ;
- relation floue : quand le lecteur ne peut pas savoir si le référent appartient ou non au groupe en question, mais que la mention linguistique choisie laisse entendre que

> c'est possible, la relation d'appartenance est floue. C'est le cas de deux référents dans le texte de presse étudié : l'épouse de HKB et celle de ADO.

Au final, notre modélisation de la référence et de la coréférence s'appuie sur une modélisation du monde des référents avec exploitation de la Théorie des Ensembles Flous pour les groupes et les relations entre groupes. Notre modélisation est également influencée par les structures des schémas d'annotation utilisés actuellement. Si nous avions réalisé ce travail il y a dix ans, nous aurions probablement enrichi le texte d'annotations portées exclusivement par les expressions référentielles : ce sont ces expressions qui mènent aux référents, et donc aux individus et groupes d'individus pour la description desquels nous aurions déterminé un ensemble de champs. Suite aux avancées de la linguistique de corpus et notamment à celles du projet Annodis avec ses schémas d'annotation structurés en « unités », « relations » et « schémas » (Widlöcher & Mathet 2009), nous avons pu appréhender une chaîne de coréférence comme un « schéma ».

Une expression référentielle reste une unité textuelle qui est délimitée par l'annotateur et enrichie d'annotations (détermination, structure de l'expression, genre, etc.). Compte tenu des possibilités explorées avec Annodis et offertes par l'outil Glozz, nous avons choisi de ne pas mettre l'identifiant du référent dans ces annotations. Ainsi, il n'y a aucun risque de se retrouver avec deux identifiants différents pour deux expressions qui réfèrent à distance au même référent. Nous avons choisi au contraire de construire un schéma pour chaque référent dont il est fait mention au moins deux fois (individu ou groupe). L'identifiant unique du référent constitue une des annotations de ce schéma. Pour identifier tous les maillons de la chaîne de coréférence, l'annotateur n'a plus qu'à associer chaque expression référentielle concernée au schéma. Non seulement il n'y a pas duplication d'information, mais de plus la procédure d'annotation peut se faire en deux phases distinctes : l'annotation des expressions référentielles (figure 1, partie gauche), puis l'affectation des référents (figure 1, partie droite).

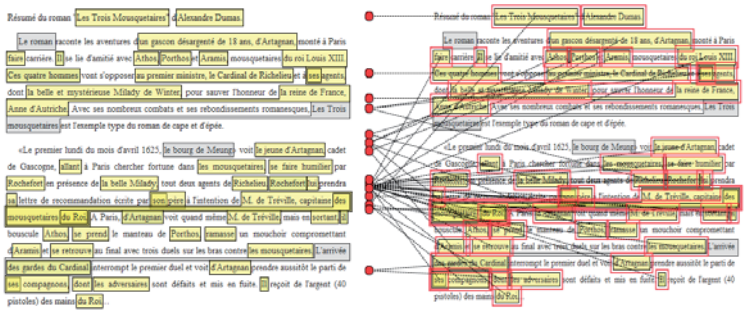

Figure 1. Délimitation des expressions référentielles (« unités ») puis des chaînes de coréférence (« schémas ») dans Glozz

Lorsque l'on crée un schéma, autrement dit lorsque l'on construit une chaîne de coréférence, on est amené à préciser certaines caractéristiques : un identifiant bien sûr, mais aussi le genre pour un individu, et, suite à ce qui précède, le type de groupe (strict ou flou) pour un groupe d'individus. Enfin, la possibilité d'avoir des relations et de pouvoir ajouter des annotations à chaque relation nous a permis de matérialiser nos relations d'appartenance : ici aussi, il suffit d'ajouter un champ avec deux valeurs possibles, « strict » et « flou ».

Par ailleurs, contrairement à Gardent & Manuélian (2005) qui modélisent les relations anaphoriques – y compris les relations associatives – par une relation entre deux unités, notre modélisation nous permet d'aller plus loin dans la description des transitions référentielles de la manière suivante : nous ajoutons à chaque schéma un champ « introduction » qui décrit la façon dont le référent est introduit : par extraction d'un autre référent, par regroupement, *in medias res*, par construction discursive, etc., et bien entendu par association. Nous intégrons ainsi avec un minimum de relations les facettes de ce thème de recherche qu'est l'anaphore associative.

4. L'annotation des éléments non référentiels

Nous avons pour l'instant considéré des expressions référentielles telles que des groupes nominaux (syntagmes nominaux de type nom propre, déterminant et nom commun, sans nom,

etc.) et des pronoms (personnel, démonstratif, adverbial, possessif, indéfini, relatif), suivant en cela les théories linguistiques classiques sur la référence (Charolles 2002). Or, à la lecture des textes, d'autres éléments nous semblent importants à prendre en compte dans les phénomènes de coréférence.

Premièrement, on peut considérer qu'un morphème peut, sans référer, rappeler au lecteur l'existence d'un référent particulier, et donc participer à la chaîne de coréférence dédiée à ce référent. C'est le cas des marques d'accord en genre et en nombre : tout verbe conjugué rappelle le nombre de son sujet, un participe passé peut également rappeler son genre, et ces formes contribuent ainsi à rappeler la nature du référent sujet. Compte tenu de l'aspect grammaticalisé et nécessaire de ces phénomènes, nous choisissons de les ignorer.

Deuxièmement, le sujet non exprimé (sujet zéro) d'un verbe à l'infinitif ou d'une participiale peut être considéré comme un élément référentiel linguistiquement non marqué. Ce n'est pas parce qu'un tel élément n'est pas exprimé qu'il n'est pas saillant à la fois du point de vue du locuteur que du lecteur, et nous choisissons de repérer ces cas en tant qu'éléments coréférentiels, ou, pour dire les choses d'une manière simple, en tant que « maillon faible » d'une chaîne de coréférence, par opposition avec une expression référentielle qui en constitue un « maillon fort » (les qualificatifs faible et fort s'appliquant à la nature de l'élément coréférentiel et non à sa contribution à la coréférence). Dans le texte de presse, « reconquérir le pouvoir » a ainsi un sujet zéro coréférent aux « héritiers » et constitue donc un maillon faible de la chaîne de coréférence des héritiers. Dans d'autres exemples, on trouve des impératifs, diverses ellipses comme dans les coordinations (« il entra et prit son chapeau »), autant d'éléments à repérer. Nous ignorons les cas où la présence d'un modal force la coréférence du sujet zéro de l'infinitif avec le sujet du modal, comme dans « ils ont voulu donner à cette entrevue […] ». Pour les autres, se pose la question de la procédure d'annotation : ce type de phénomène est toujours un obstacle aux méthodes d'annotation dans la mesure où l'on souhaite repérer et annoter du vide. C'est d'ailleurs un argument parfois avancé pour ne pas tenir compte de ce type de

phénomène (ce qui illustre un exemple d'influence des procédures de la linguistique de corpus sur la façon d'appréhender les phénomènes). Comme nous souhaitons au moins repérer ces phénomènes, quitte à les ignorer lors de certaines analyses de corpus ultérieures, nous choisissons de repérer le verbe support en tant qu'unité à annoter. Cette solution, beaucoup plus raisonnable que celle consistant à annoter une espace ou un signe de ponctuation, nous amène à repérer le verbe en tant qu'expression support d'un élément coréférentiel. Deux méthodes sont possibles : 1. ajouter un trait « support » à l'unité « expression référentielle » utilisée pour repérer tous les maillons de chaînes de coréférence ; 2. utiliser deux unités, « expression référentielle » et « élément coréférentiel » pour distinguer les maillons forts des maillons faibles.

Troisièmement, une construction pronominale fait apparaître un pronom supplémentaire qui rappelle le référent et contribue donc à la coréférence. Plusieurs cas doivent être distingués ici : nous choisissons d'annoter le pronom supplémentaire dans les constructions pronominales de type subjectif (réfléchies comme « il se lave », réciproques comme « ils se battent », avec coréférence dans les deux cas), et de ne pas annoter les constructions pronominales de type objectif (avec agent implicite de type « on » : « ça se dit », etc.), ni les cas où le sujet n'est pas référentiel, comme dans « il se noie plus de personnes dans la Seine que dans la Loire ».

Quatrièmement, une construction attributive, une étiquette, une parenthèse, tout syntagme qui, sans référer, vient ajouter une information à une expression référentielle est susceptible de constituer un élément coréférentiel tel qu'on l'entend. Nous choisissons donc d'annoter ces éléments.

Cinquièmement, un terme lexical peut évoquer un référent sans y référer lui-même. C'est le cas de « parricide » qui évoque le père tout en référant au fils, de « beau-parents » qui, du moins dans une acception du terme, évoque aussi le mari ou la femme. Nous choisissons d'ignorer ces référents indirects.

5. Annotation manuelle et déductions automatisées

D'une manière générale, s'il est clair que les informations morphologiques et syntaxiques peuvent être récupérées automatiquement suite au parsage du texte, il est clair également que l'annotation des expressions référentielles, l'annotation des éléments coréférentiels et la construction de chaînes de coréférence ne peuvent pas être automatisées, ou alors au prix de beaucoup d'erreurs. Avec l'objectif d'exploiter au maximum ce qui peut être automatisé, nous définissons une procédure en plusieurs étapes, certaines automatiques, d'autres manuelles, et, pour celles-ci, différents niveaux d'annotation :

- préparation du corpus : importation de texte brut ; découpage automatique en phrases (au sens graphique du terme) ; repérage du titre et des métadonnées ;
- repérage des expressions référentielles (délimitation exclusivement) : manuellement, éventuellement avec l'aide d'un outil capable de faire un pré-repérage des expressions référentielles sur des indices tels que la détermination ;
- repérage des éléments coréférentiels (délimitation) : manuellement, éventuellement avec l'aide d'un outil – qui dans le meilleur des cas devrait être un analyseur syntaxique capable de repérer entre autres les sujets zéro et les expressions attributives ;
- création d'une chaîne de coréférence (au niveau le plus simple) : regroupement manuel dans un schéma des expressions référentielles et des éléments coréférentiels qui concernent un même référent ; attribution d'un identifiant à ce référent ;
- création manuelle des relations d'appartenance entre référents, c'est-à-dire entre chaînes de coréférence.

A ce stade, l'annotateur dispose d'un corpus structuré en termes de références et de coréférences. Des analyses sont d'ores et déjà possibles sur la portée des chaînes de coréférence, sur la manière dont le texte passe d'un référent à un autre, sur la nature des groupes d'individus, etc. Aucune donnée morphologique, syntaxique ou sémantique n'a cependant été affectée. C'est l'objet des phases suivantes (facultatives) :

- annotation des propriétés de la chaîne de coréférence : détermination du type d'introduction de la chaîne ; affectation de manière manuelle ou semi-automatique de propriétés concernant le référent : type (individu, groupe strict ou groupe flou), cardinal (pour un groupe), genre (pour un individu), etc. ;
- annotation manuelle ou semi-automatique des propriétés des expressions référentielles : détermination, genre, nombre, type de syntagme, etc. (rien n'est indispensable dans ces données : plus on en crée, plus on pourra interroger le corpus sur des aspects variés et tenter d'identifier des corrélations entre les données) ;
- annotation manuelle ou semi-automatique des propriétés des éléments coréférentiels : nature du support (pour les éléments zéro), propriétés similaires à celles des expressions référentielles (même remarque qu'au point précédent) ;
- annotation de syntagmes supplémentaires qui permettent de mettre en perspective les éléments des chaînes de coréférence compte tenu de la structure et du style du texte, par exemple : repérage des dialogues, des parenthèses, des discours rapportés, des discours indirects, des suppositions, des flashbacks, etc., de manière à distinguer deux plans de narration qui permettront de considérer les maillons d'une chaîne de deux manières différentes ; repérage des cadratifs ou de toute autre partie de phrases (du discours) de manière à tester d'éventuelles corrélations avec par exemple les débuts des chaînes de coréférence.

Restent deux questions qui ont fait l'objet de longues discussions dans le groupe de travail « coréférence » : la question de la saillance des référents et celle de l'analyse des types de transitions référentielles. Concernant la saillance, nous choisissons de ne pas laisser la possibilité à l'annotateur d'affecter des propriétés de saillance aux éléments annotés, mais, au contraire, de calculer automatiquement la saillance d'un référent à partir des fréquences des références et de divers paramètres linguistiques et textuels (Landragin 2004). Concernant

les transitions référentielles, nous choisissons de même de développer un outil d'analyse de corpus pour que les différents types de transition imaginables (continuation, prolongation, bifurcation, abandon, etc.) soient détectés en fonction des passages d'une référence à une autre. La technique utilisée est alors celle du motif (Mellet & Longrée 2009).

En ce sens, nous pouvons affirmer que les méthodologies de la linguistique de corpus ont influé sur notre perception des phénomènes de référence, coréférence, saillance et transition référentielle : non seulement elles ont dirigé comme nous l'avons décrit la manière de structurer nos entités et leurs propriétés (unités, relations, schémas), en mettant en avant les références et les coréférences, mais, de plus, elles nous ont conduit à ignorer dans la procédure d'annotation des phénomènes qui étaient au cœur de notre recherche mais pour lesquels nous ne pouvions pas définir de traits « attribut-valeur ». Plutôt que de forcer un annotateur à remplir pour chaque expression référentielle un trait « saillance » avec comme valeurs possibles « faible », « moyenne » et « forte », il nous a paru beaucoup plus efficace de structurer proprement les successions de références pour permettre à un outil de tester différentes méthodes de calcul de saillance à partir de cette structure.

6. Outil et schéma d'annotation pour la coréférence

Après des essais avec divers logiciels, nous avons réalisé les tests de notre procédure avec tout d'abord l'outil Glozz déjà cité, puis avec l'outil Analec (Victorri 2011). Les activités du groupe de travail « coréférence » sont d'ailleurs à l'origine d'un grand nombre d'évolutions de cet outil : gestion des unités, relations et schémas en totale compatibilité avec Glozz et de manière interactive (cf. figure 2 qui illustre une version de notre schéma d'annotation) ; visualisation conjointe ou séparée des diverses unités et des diverses annotations ; visualisation des chaînes de coréférence par sélection de propriétés et codage couleur des maillons (figure 3).

L'outil Analec est actuellement en phase de test avancé, de manière à valider l'ensemble des besoins requis par notre étude et dont les figures précédentes ne donnent qu'un rapide

aperçu : gestion, regroupement, visualisation et impression de corpus annoté ; annotation ergonomique (*via* le concept de vue, c'est-à-dire à l'aide de filtrages élaborés des informations disponibles) ; calculs de fréquences ; recherche de corrélations ; générations de tableaux de chiffres et de schémas ; etc.

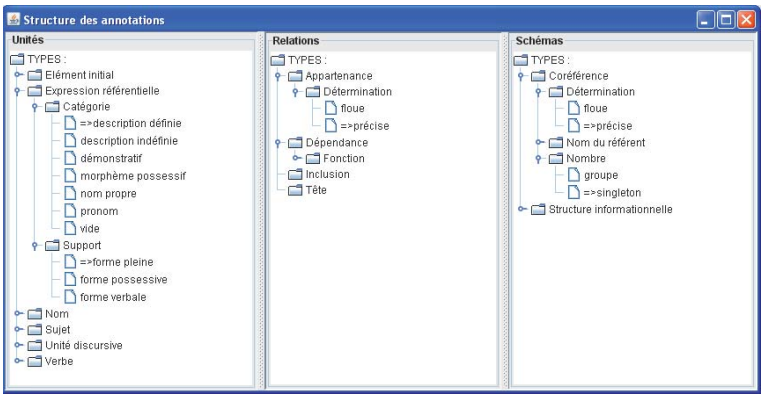

Figure 2. Gestion dans Analec des unités, relations et schémas pour différents projets d'études de corpus

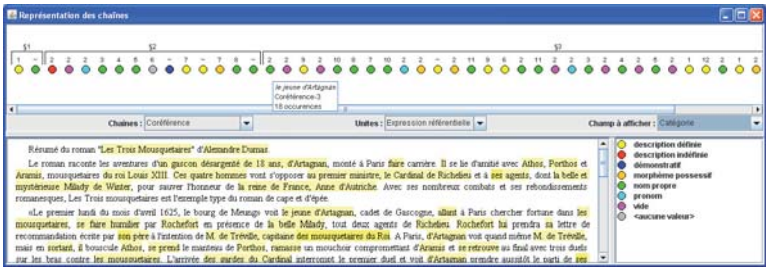

Figure 3. Visualisation dans Analec de chaînes de coréférence, avec ici un code couleur affecté à la structure des expressions référentielles (type de morphème ou de pronom, détermination)

7. Conclusion

Nous avons décrit dans cet article une procédure d'annotation adaptée à l'analyse des références et des chaînes de coréférence dans des textes écrits. Cette procédure est issue d'un dialogue permanent entre les méthodologies actuelles de la linguistique

de corpus et notre besoin d'une linguistique outillée pour l'étude de phénomènes sémantiques complexes et parfois relativement éloignés du matériau linguistique. Elle se fonde sur des choix concernant les données à annoter et les données qui seront ultérieurement déduites d'annotations, sur une structuration en unités, relations et schémas, et sur un outil ergonomique qui permet d'optimiser l'ensemble des opérations nécessairement manuelles.

L'avantage essentiel de cette méthodologie est de prendre en compte à la fois la réalité conceptuelle des référents et la réalité linguistique, sans éluder – dans la mesure du possible – les problèmes d'ambiguïtés, de subjectivité des interprétations, de multiplicité des niveaux d'analyse, de multiplicité des plans discursifs. Son principal inconvénient est par conséquent sa complexité et le nombre des données à annoter, du moins dans le cas de figure où l'on cherche à maximiser les possibilités d'interrogation et de recherche de corrélations.

Nous espérons avec ce travail avoir commencé à répondre aux trois problèmes qui se sont posés il y a maintenant presque deux ans : problème de temps (trop d'éléments à annoter et surtout trop de manipulations à faire avec l'interface de l'outil d'annotation) ; problème de décision (primauté de l'interprétation de l'annotateur dès qu'on traite de phénomènes sémantiques et pragmatiques) ; problèmes techniques : logiciels existants peu stables, peu ergonomiques, et surtout très lents dès qu'on annote un texte de plusieurs pages, ce qui s'avère indispensable pour l'étude de la coréférence…

Références bibliographiques

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

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

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

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

Gardent C. & Manuélian H. (2005). « Création d'un corpus annoté pour le traitement des descriptions définies », *Traitement Automatique des Langues*, vol. 46,1 : 115-139.

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

Hobbs J. (1979). « Coherence and coreference », *Cognitive Science* 3 : 67-90.

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

Landragin F. (2004). « Saillance physique et saillance cognitive », *Cognition, Représentation, Langage (CORELA)* 2(2) : http://edel.univ-poitiers.fr/corela.

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

Longo L. & Todirascu A. (2010). « RefGen : un module d'identification des chaînes de référence dépendant du genre textuel », *17e Conférence sur le Traitement Automatique des Langues Naturelles*. Montréal.

Mellet S. & Longrée D. (2009). « Syntactical 'Motifs' and Textual Structures », *Belgian Journal of Linguistics* 23 : 161-173.

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

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

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

Victorri B. (2011). « Analec : logiciel d'annotation et d'analyse de corpus écrits », logiciel téléchargeable sur : http://www.lattice.cnrs.fr/-Analec-.

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