Analyse des références et des transitions référentielles : l'apport de la linguistique outillée

Frédéric Landragin, Thierry Poibeau, Bernard Victorri

Analyse des références et des transitions référentielles :
l'apport de la linguistique outillée

Frédéric Landragin, Thierry Poibeau et Bernard Victorri

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

DRAFT AUTEURS (version non finale)

1. Introduction

L'accès aisé et quasi immédiat à de très grandes masses de données textuelles, notamment sur Internet, a profondément changé les manières de travailler en linguistique. Il était encore récemment tout à fait courant de travailler à partir d'un petit nombre d'exemples fabriqués ou attestés. Ce type de pratique existe toujours, bien évidemment, mais de plus en plus les linguistes se tournent vers les corpus qui sont alors assimilés à des « réservoirs » de données et/ou de connaissances. Tout ceci n'est pas vraiment nouveau – on se rappellera l'opposition de C. Fillmore entre le « linguiste en chaise longue » et le « linguiste de terrain » (Fillmore, 1992) ou bien encore le « linguiste à l'instrument », pour reprendre le mot de B. Habert (2005) – mais la tendance actuelle semble clairement aller vers une meilleure prise en compte des corpus et de la linguistique expérimentale.

Les données permettent ainsi de valider ou d'invalider des hypothèses anciennes. L'étude de J. Bresnan *et al*. (2007) sur le datif anglais ("*He gave a book to John*" / "*He gave John a book*") a par exemple fait date en ce qu'elle a permis de « revisiter » ce phénomène, à la fois complexe et mal pris en compte dans les théories qui se fondent en majorité sur des exemples fabriqués et sur l'intuition du locuteur. Comment expliquer l'alternance entre deux formes concurrentes ? Quels sont les critères linguistiques pertinents pour expliquer ce phénomène ? L'alternance implique-telle une variation du contenu sémantique ? L'étude Bresnan a permis d'apporter des réponses à la plupart de ces questions. Son modèle (fondé sur une analyse à base de régression logistique) permet de pondérer les paramètres selon leur

importance, pour expliquer l'usage observé des constructions concurrentes. On voit qu'ici le recours aux données ne se limite pas à un simple comptage d'occurrences, mais qu'il permet d'éliminer certaines hypothèses et d'établir des modèles mêlant plusieurs paramètres que les humains ont souvent du mal à hiérarchiser tant les données sont complexes.

C'est en partie ce que nous cherchons à faire ici, en promouvant une linguistique outillée pour l'analyse de la référence, phénomène complexe mettant en jeu une multiplicité de paramètres que les études jusqu'ici ont eu du mal à classer et évaluer. Les pistes que nous proposons dans ce chapitre font écho à des préoccupations de Michel Charolles, d'une part au niveau de l'analyse outillée de la référence et des expressions référentielles dans des textes écrits en français (Charolles, 2002), d'autre part au niveau de l'analyse outillée des chaînes de références et des transitions référentielles (Charolles, à paraître). Ces préoccupations ont été discutées dans le groupe de travail « Identification des référents et transitions référentielles » du laboratoire Lattice (2009-2011), puis dans le projet PEPS MC4 « Modélisation Contrastive et Computationnelle des Chaînes de Coréférence » (2011-2012), et ce chapitre apporte un éclairage expérimental à des interrogations qui sont restées jusqu'ici théoriques, du moins dans le cadre de ces travaux effectués sur la référence dans le laboratoire Lattice.

2. Référence, suite des références et transitions référentielles

Si l'annotation, la visualisation et l'interrogation de corpus pour des phénomènes morphosyntaxiques et syntaxiques posent beaucoup de problèmes méthodologiques et requièrent des outils adaptés, il en est de même, et peut-être encore plus, pour des phénomènes sémantico-pragmatiques tels que la référence et les transitions référentielles. D'une part parce que travailler sur les phénomènes de référence implique de définir ce qu'est une expression référentielle et de construire le monde des référents, avec par exemple une modélisation des types d'accès aux référents, tâches qui ne peuvent être réalisées de manière fiable qu'à la main, d'autre part parce que travailler sur les transitions référentielles, et donc sur la suite (ou succession) de toutes les références présentes dans un texte, nécessite des outils capables de produire, pour l'utilisateur linguiste, des visualisations et des pré-analyses qui l'aident à appréhender cette suite d'annotations.

Concernant les phénomènes de référence, une approche outillée consiste ainsi à repérer dans le texte les matérialisations de références. Suivant (Charolles, 2002, p. 39), « pour qu'il y ait référence, il faut que quelqu'un accomplisse par le biais de la parole un comportement précis visant à mentionner une certaine chose ». Sans entrer ici dans les détails de ce qu'est la parole ou de la distinction entre dénotation et référence, nous noterons que

cette action de mentionner se fait par la production d'une expression référentielle. Avec un outil d'annotation, notre tâche consiste donc à repérer ces expressions référentielles. Suivant (Charolles, 2002, p. 1), nous nous intéressons en priorité aux types d'expressions référentielles suivants : « les noms propres, les groupes nominaux définis, démonstratifs, indéfinis, et les pronoms ». Rien qu'en dressant cette liste de types, nous introduisons une donnée qui n'est pas exprimée par le texte. Notre tâche consiste alors à annoter les expressions référentielles repérées, de manière à enrichir notre corpus par des données qui sont autant d'analyses. En plus du type d'expression référentielle, nous attachons à chaque expression repérée plusieurs traits (couples « attribut / valeur ») : fonction grammaticale, position dans la phrase, rôle thématique, etc. Toutes ces données sont enregistrées dans le corpus, et permettent d'explorer celui-ci en choisissant un ensemble de critères, par exemple toutes les expressions référentielles qui apparaissent en tête de phrase avec la fonction grammaticale sujet et le rôle thématique d'agent. C'est une première étape vers l'analyse outillée des expressions référentielles. Par ailleurs, nous enregistrons également dans le corpus des données sur le monde des référents : chaque expression référentielle est liée à un référent, qui est lui aussi repéré par un ensemble de traits (identifiant, type de référent – individu, objet concret, etc. – genre, nombre) et relié éventuellement à d'autres référents. Ces traits permettent d'étudier les références, et l'analyse combinée des traits des expressions référentielles et de ceux des référents permet d'entrer pleinement dans l'analyse outillée de la référence, avec – comme exemple rendu possible par l'outil – le test d'une corrélation entre type d'expression référentielle et type de référent, ou encore entre fonction grammaticale et identifiant du référent (à quel personnage s'applique le plus souvent la fonction sujet ?).

Concernant les phénomènes de transition référentielle, Michel Charolles a longtemps imaginé l'étude des références d'un texte par fenêtres de trois ou quatre références successives, une telle « fenêtre » étant relativement facile à appréhender cognitivement et permettant d'explorer des hypothèses sur les types de transitions référentielles. Pour être menée de manière rationnelle, l'analyse d'un texte avec ce principe de fenêtre requiert cependant des outils, et ce n'est que récemment, avec par exemple les avancées du logiciel ANALEC (Analyse de l'Ecrit), que l'on peut commencer à explorer une telle voie (Victorri, 2012 ; Landragin, 2011 ; Mellet & Longrée, 2009). Notre objectif dans cet article est d'illustrer les apports récents de la linguistique outillée pour des études linguistiques sur ces aspects. Nous considérons ainsi la suite des références d'un texte, c'est-à-dire l'enchaînement successif de toutes les expressions référentielles repérées lors de la phase décrite ci-dessus. Cette suite de références, par définition, est l'ensemble ordonné des expressions référentielles

du texte, accompagnées de leurs annotations et des annotations des référents associés. L'outil permet de visualiser une telle suite, et de l'explorer en ayant recours ici aussi à des paramètres correspondant aux données enregistrées dans le corpus. L'outil permet également d'opérer des calculs sur cette suite, et notamment des calculs visant à déterminer les « fenêtres » successives. La suite des références devient ainsi un objet d'étude moins hermétique : à travers la liste calculée des types de fenêtre les plus fréquents (nous y reviendrons), le linguiste peut désormais repérer où sont les suites de références consécutives à un même individu, où sont les suites alternées de références à deux référents (qui sont par là même en compétition), ou encore par combien de transitions d'un référent à un autre le texte se caractérise. Ce sont ces aspects que nous allons maintenant détailler, avec une section centrée sur la visualisation et deux autres sur les calculs réalisables par l'outil.

3. Visualisation de la suite des références

Une fonctionnalité indispensable d'un outil conçu pour des analyses linguistiques est la génération de représentations graphiques : la linguistique de corpus ne se satisfait plus des traitements de texte, et des outils permettant des visualisations colorées et ergonomiques s'avèrent nécessaires. L'objectif est d'accéder de manière conviviale et efficace aux données d'un corpus, en exploitant les capacités de la perception visuelle humaine : utilisation de codes graphiques (couleur, taille, distance) ; mise en saillance des données les plus importantes ou les plus pertinentes compte tenu de la requête de l'utilisateur.

Les outils disponibles actuellement (Cunningham *et al.*, 2002 ; Morton & LaCivita, 2003 ; Müller & Strube, 2006 ; Venant, 2008 ; Widlöcher & Mathet, 2009) regroupent généralement l'annotation et la visualisation (sans l'oublier l'interrogation). Certains sont optimisés pour rendre la phase d'annotation la plus rapide possible, d'autres sont optimisés pour la visualisation et le confort visuel. Chaque outil est également optimisé en fonction d'un objet d'étude linguistique. MMAX puis MMAX 2 (Müller & Strube, 2006) sont ainsi conçus au départ pour l'annotation et l'analyse des anaphores et des chaînes de coréférence. S'ils sont capables de bien gérer des expressions référentielles et des relations entre ces expressions, ils présentent des lacunes pour la gestion d'ensembles d'annotations, comme celui de la suite des références d'un texte. GLOZZ (Mathet & Widlöcher, 2012), conçu initialement pour l'annotation et la visualisation des structures discursives, s'avère en revanche très bien adapté à l'étude de la suite des références d'un texte. Il permet plusieurs types de visualisation pour un ensemble d'annotations, par exemple sous forme de boîte englobante, ou, mieux, sous forme de liste chaînée (cf. figure 1). Il ajoute à l'état de l'art la possibilité de visualiser

l'intégralité du texte (bandeau vertical à gauche de l'interface représentée en figure 1), de manière à repérer visuellement les zones les plus concernées par des références à tel personnage, par exemple.

– Figure 1 à insérer à peu près ici –

ANALEC, dans sa toute dernière version (Victorri, 2012), propose une fonctionnalité supplémentaire consistant à représenter la suite des références 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 par conséquent le code couleur. Il peut ainsi confronter diverses représentations, les exporter, les superposer, etc. La figure 2 présente l'interface de visualisation de cette suite des références, avec un code couleur lié au type d'expression référentielle et un affichage mettant en perspective le découpage en paragraphes du texte, de manière à visualiser directement des liens éventuels entre par exemple utilisation du nom propre et amorçage d'un nouveau paragraphe. Il est à noter qu'en plus des données d'annotation, d'autres données ou calculs peuvent être exploités en tant que code couleur. C'est notamment le cas du nombre de références à un même référent. Cette possibilité a fait apparaître lors de l'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, et c'est là tout l'intérêt d'une visualisation graphique appropriée.

– Figure 2 à insérer à peu près ici –

Plusieurs représentations visuelles peuvent être générées, de manière à multiplier les possibilités de détection rapide de phénomènes intéressants. La suite complète des références, même pour un texte ayant la taille modeste d'une nouvelle (cf. figure 3), peut cependant devenir trop riche et par là même moins informative que ne l'est la visualisation d'un sous-ensemble de cette suite.

– Figure 3 à insérer à peu près ici –

4. Calculs sur la suite des références, pour un référent donné

Ces visualisations ne suffisent pas à identifier aisément toutes les caractéristiques de la suite des références d'un texte, et une deuxième étape d'analyse consiste alors à exploiter des outils de repérage automatique de phénomènes remarquables tels que des motifs dans les transitions référentielles. Nous proposons dans ce chapitre quelques pistes dans ce sens, avec notamment

l'exploitation de bigrammes ou de trigrammes, pistes que nous illustrons avec le déroulement et les résultats de tests effectués sur des textes courts. Cette section aborde la question des calculs réalisables sur des suites de données telles qu'un sous-ensemble de la suite des références du texte, les exemples utilisés ici ayant trait à la nouvelle « Les bijoux » de Maupassant et étant ciblés sur l'étude d'un référent donné. Avec la visualisation de la suite des références, l'utilisateur obtient comme nous l'avons vu dans les figures 2 et 3 (voir la suite de chiffres qui apparaît au-dessus des points colorés) un message codé avec un alphabet particulier, message qui peut être analysé à l'aide de techniques classiques telles que le comptage des N-grammes.

Les calculs les plus simples concernent des fréquences : nombre d'éléments de la suite de références en cours d'analyse, répartition des expressions référentielles par chapitre ou par paragraphe, taille moyenne des intervalles (en nombre de mots, en nombre de phrases) entre deux expressions référentielles, etc. Il s'agit de comptages sur la structure même de la suite des références, sans tenir compte des annotations. Si l'on intègre celles-ci, par exemple les types d'expressions référentielles, on peut alors aborder des calculs sur la répartition de ces types dans la suite ou dans un sous-ensemble de cette suite correspondant par exemple à l'un des personnages : noms propres plutôt en début, pronoms de 3^{e} personne plutôt à la fin, etc. Sans aller jusque-là, la figure 4 montre avec des calculs de fréquences la constitution de quelques-uns des sous-ensembles de la suite des références, chaque sous-ensemble étant lié à l'un des référents principaux.

– Figure 4 à insérer à peu près ici –

Nous évoquions plus haut la mise en œuvre d'un alphabet et des calculs de fréquences des N-grammes sur des messages codés selon cet alphabet. La figure 5 montre un exemple d'une telle application, avec un alphabet à trois lettres « 0, 1, 2 » mettant en avant le référent donné (pour lequel on utilise le code « 1 »), de manière à tester sa fréquence d'apparition et les transitions référentielles avec d'autres référents (pour lesquels on utilise le code « 2 »), tout cela en tenant compte des changements de paragraphe (code « 0 ») et en laissant de côté les référents non humains de manière à focaliser l'analyse sur les seuls personnages. Le premier trigramme obtenu, celui correspondant à la suite des références à Mr. Lantin, montre que les trigrammes les plus fréquents sont avant tout des continuations (« 1 1 1 » et « 2 2 2 »), et que Mr. Lantin est clairement le personnage principal de « Les bijoux », évoqué de manière continue (sans compétiteur) à de nombreux endroits. On a ensuite des successions de références à Mr. Lantin outrepassant les limites des paragraphes (« 1 0 1 » et « 0 1 1 »), puis les choses se mélangent sans que rien d'intéressant ne ressorte. Le deuxième trigramme

obtenu, celui correspondant à Mme. Lantin en tant que référent donné, ne montre pas grand-chose à part le fait que celle-ci n'est pas le personnage principal.

– Figure 5 à insérer à peu près ici –

Avec la suite des références à Mr. Lantin, l'analyse en bigrammes glissants donne des résultats complémentaires : « 1 1 » (111 fois) ; « 2 2 » (73 fois) ; « 1 2 » (55 fois) ; « 0 1 » (54 fois) ; « 1 0 » (51 fois) ; « 2 1 » (51 fois) ; « 2 0 » (31 fois) ; « 0 2 » (27 fois). Ce classement montre encore une fois que Mr. Lantin est bien le personnage principal – on s'en doutait – mais montre également que la lettre « 0 » apparaît avant tout avec « 1 » et beaucoup moins fréquemment avec « 2 ». On en déduit que les paragraphes sont beaucoup plus calqués sur Mr. Lantin que sur tout autre référent, ce qui est déjà plus intéressant en soi.

Nous aurions pu illustrer ici d'autres alphabets avec d'autres types de résultats. Il est par exemple intéressant d'ignorer complètement les expressions concernant d'autres référents que celui en cours d'analyse, et, plutôt que de réduire l'alphabet à deux lettres, remplacer le « 1 » par un caractère codant le type d'expression référentielle ou la fonction grammaticale, ou encore le rôle thématique. De même, il est intéressant de calculer les écarts entre références, et d'ajouter cette donnée aux annotations existantes. Si on choisit comme unité pour cet écart le nombre d'expressions intercalées, on obtient un ensemble de chiffres qui peut lui aussi être considéré comme un alphabet.

Un point important concernant ces calculs est leur matérialisation dans un nouveau trait d'annotation. Avec cette donnée dont le coût de production est négligeable puisqu'elle est automatisée, la linguistique outillée rend possible la recherche de corrélations, par exemple entre le type d'expression – on pense au nom propre – et l'écart avec l'expression précédente. Ces calculs, une fois effectués sur plusieurs dizaines de textes de tailles et de genres variés, permettrait d'apporter des arguments quantitatifs pour ou contre l'importance du nom propre en tant qu'indice de redémarrage d'une chaîne de référence (Schnedecker, 1997). Un autre point important a trait à la taille de la « fenêtre » : nos exemples ont impliqué des fenêtres de deux (bigrammes) et de trois références (trigrammes), mais nous n'avons pas réalisé ces seuls tests. Nous avons notamment constaté qu'une fenêtre de plus de trois références entrainait trop de configurations différentes, et par conséquent des fréquences faibles, peu susceptibles d'observations intéressantes. Bigrammes et trigrammes semblent être les plus pertinents, et nous allons les tester maintenant sur la suite complète des références.

5. Calculs sur la suite complète des références

D'autres calculs consistent à utiliser tout simplement l'identifiant de chacun des référents comme une lettre de l'alphabet sur lequel effectuer des calculs de fréquences. Toutes les références sont alors prises en compte, ce qui permet de se focaliser sur la typologie du texte, avec l'ensemble des transitions référentielles. La prise en compte d'une donnée supplémentaire comme le type d'expression référentielle est alors problématique, dans la mesure où combiner référent et type conduit à appliquer un produit (au sens mathématique) pour obtenir un alphabet avec un nombre élevé de lettres. Du fait de ce nombre élevé, quelques tests nous ont montré que même le trigramme le plus fréquent était en fait très peu fréquent (avec 3 ou 4 occurrences dans tout le texte) et donc peu susceptible de fournir des résultats pertinents. A part ignorer tout simplement les annotations, une solution consiste à n'analyser plus que les bigrammes glissants, ce qui s'avère réducteur. Une autre solution consiste à donner à l'utilisateur le contrôle de filtres et de fonctions de simplification, dont le but est de réduire l'alphabet, ce qui passe nécessairement par la réduction du nombre de valeurs de chaque donnée prise en compte. Réduire le nombre des valeurs prises par un champ tel que « type d'expression référentielle » ou « fonction grammaticale » pose des problèmes théoriques, qui reviennent à dire quelles sont les valeurs importantes et quelles sont les autres. Dans le but d'un calcul, c'est potentiellement une solution acceptable, surtout si elle fait apparaître une observation qui n'aurait pas été possible autrement. Réduire le nombre de référents, outre la simplification extrême consistant à se focaliser sur un seul référent comme nous l'avons fait dans la section précédente, peut se faire en suivant les relations entre référents éventuellement spécifiées dans le monde des référents (appartenance d'un individu à un groupe, notamment). Grosso modo, il s'agit alors, dans le graphe reliant les référents entre eux, d'identifier les sous-graphes connexes et de considérer ceux-ci comme les éléments principaux de l'alphabet.

– Figure 6 à insérer à peu près ici –

En attendant, la figure 6 montre les résultats de la prise en compte de la suite complète des références, sans autre donnée que les identifiants des référents. Le classement en haut à gauche correspond aux trigrammes glissant avec comme donnée de départ l'ensemble des expressions référant aux neuf référents du texte, et la lettre « 0 » pour les changements de paragraphe. Nous y retrouvons la continuité fréquente sur le personnage « 1 » (Mr. Lantin), même au-delà des frontières de paragraphes (« 1 0 1 » et « 0 1 1 »), puis les apparitions fréquentes du personnage « 2 » (Mme. Lantin), et, plus loin, du personnage « 7 » (le deuxième bijoutier, qui devient de fait le troisième personnage principal, largement devant tous les

autres qui n'apparaissent même pas ici). L'analyse en bigrammes glissants donne le classement suivant : « 1 1 » (111 fois) ; « 0 1 » (54 fois) ; « 1 0 » (51) ; « 2 2 » (48) ; « 1 2 » (26) ; « 2 1 » (22) ; « 7 1 » (18) ; « 1 7 » (16) ; « 2 0 » (14) ; « 0 7 » (13) ; « 7 0 » (11) ; « 0 2 » (10). Sans surprise, les trois bigrammes les plus fréquents ne concernent que Mr. Lantin. Les trois suivants reflètent les alternances fréquentes entre des références à Mr. Lantin et des références à Mme. Lantin, sans qu'un seul changement de paragraphe n'apparaisse (autrement dit, ceux-ci restent liés à Mr. Lantin seul). Les deux suivants indiquent que les alternances entre Mr. Lantin et le deuxième bijoutier sont également fréquentes, toujours à l'intérieur des paragraphes et sans intervention d'autre référent. Enfin, les quatre derniers bigrammes de la liste partielle ci-dessus font apparaître des changements de paragraphes, soit avec Mme. Lantin, soit avec le deuxième bijoutier, ce qui rapproche encore un peu les façons qu'a l'auteur de référer à ces deux personnages, façons apparemment bien différentes de celles relatives au personnage principal.

Conclusion

Nous avons montré quelques apports de la linguistique outillée pour l'analyse des références et des transitions référentielles dans un texte de la taille d'une nouvelle : annotation sous forme de données structurées des expressions référentielles ; annotation structurée du monde des référents ; visualisations diverses de la suite des références d'un texte ; observations calculées sur la typologie de cette suite et de sous-ensembles correspondant à quelques référents particulièrement intéressants. Ces premiers résultats montrent que nous sommes désormais bien loin de la linguistique de corpus effectuée dans les strictes limites d'un traitement de texte : bien au contraire, les outils ergonomiques récents montrent tout leur intérêt, surtout quand ils jouent avec des codes graphiques qui permettent au linguiste de repérer rapidement des régularités ou des phénomènes remarquables.

Nous avons montré également qu'à partir de calculs simples tels que des comptages de N-grammes, il était possible d'observer des phénomènes difficiles à percevoir autrement, et qu'il était même possible d'apporter des arguments à des hypothèses linguistiques, comme celle portant sur le fait qu'un paragraphe commence souvent avec le même référent que le dernier référent mentionné dans le paragraphe précédent (Schnedecker, 1997). C'est le sens de la fréquence importante du trigramme « 1 0 1 » dans le texte étudié. Par ailleurs, pour répondre à l'une des interrogations de Michel Charolles, nous avons montré ici que dans le cadre des tests réalisés, la taille de la fameuse « fenêtre » de références pouvait se réduire à

deux ou trois éléments : quatre éléments ou plus ne permettent plus aucune observation valable.

Enfin, nous voulons insister sur un dernier point, à savoir que l'utilisation d'outils ne remplace en aucune façon les analyses linguistiques réalisées manuellement : les deux sont complémentaires, et les observations issues de calculs ne prennent sens qu'en les rapprochant d'analyses faites par le linguiste. L'outil ne fait qu'apporter des facilités et des points de vue plus complexes et plus variés que ne le permet l'analyse fondée sur la lecture.

Références bibliographiques

Bresnan J., Cueni A., Nikitina T., Baayen H., « Predicting the Dative Alternation ». In: Boume G., Kraemer I., Zwarts J. (Eds.) *Cognitive Foundations of Interpretation*, Royal Netherlands Academy of Science, Amsterdam, 2007, pp. 69–94.

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

Charolles M., *L'anaphore et les chaînes de référence en français*, Paris, Ophrys, à paraître.

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

Fillmore C. J., « 'Corpus linguistics' vs. 'Computer-aided armchair linguistics' ». In: *Directions in Corpus Linguistics (Proceedings from a 1992 Nobel Symposium on Corpus Linguistics, Stockholm)*, Mouton de Gruyter, La Hague, 1992, p. 35–60.

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

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

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

Mathet Y., Widlöcher A., « Glozz Annotation Platform », http://www.glozz.org/, 2012.

Mellet S., Longrée D., « Syntactical 'Motifs' and Textual Structures », *Belgian Journal of Linguistics*, vol. 23, 2009, p. 161–173.

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

Müller C., Strube M., « 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, 2006.

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

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

Victorri B., « ANALEC : logiciel d'annotation et d'analyse de corpus écrits », logiciel téléchargeable sur : http://www.lattice.cnrs.fr/ANALEC, 2012.

Widlöcher A., Mathet Y., « 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, 2009.

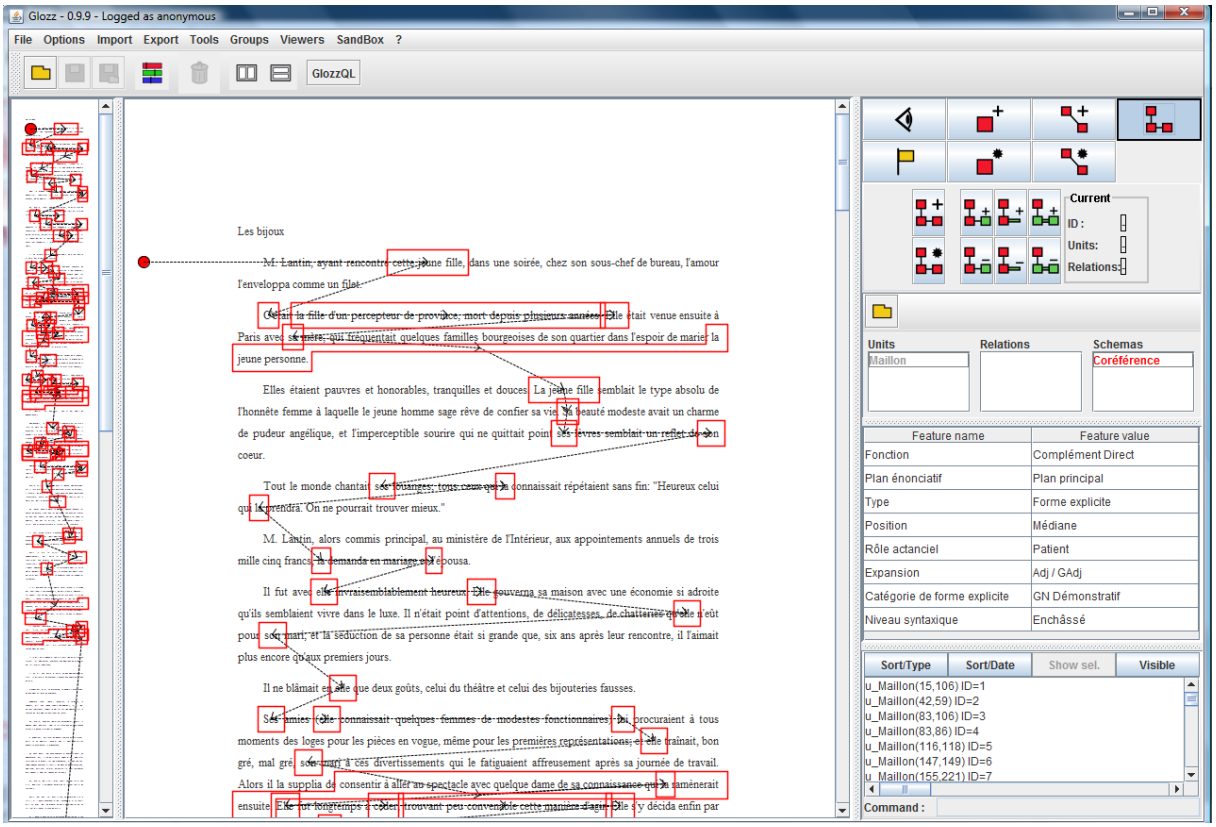

Figure 1. Interface principale de GLOZZ. La fenêtre centrale montre le texte, ici la nouvelle « Les bijoux » de Maupassant, avec en surimpression la suite des références à la jeune fille, l'un des personnages principaux du texte. A droite se trouve l'interface d'annotation avec les traits que nous avons évoqués dans la section 2 : colonne « attribut » à gauche, colonne « valeur » à droite.

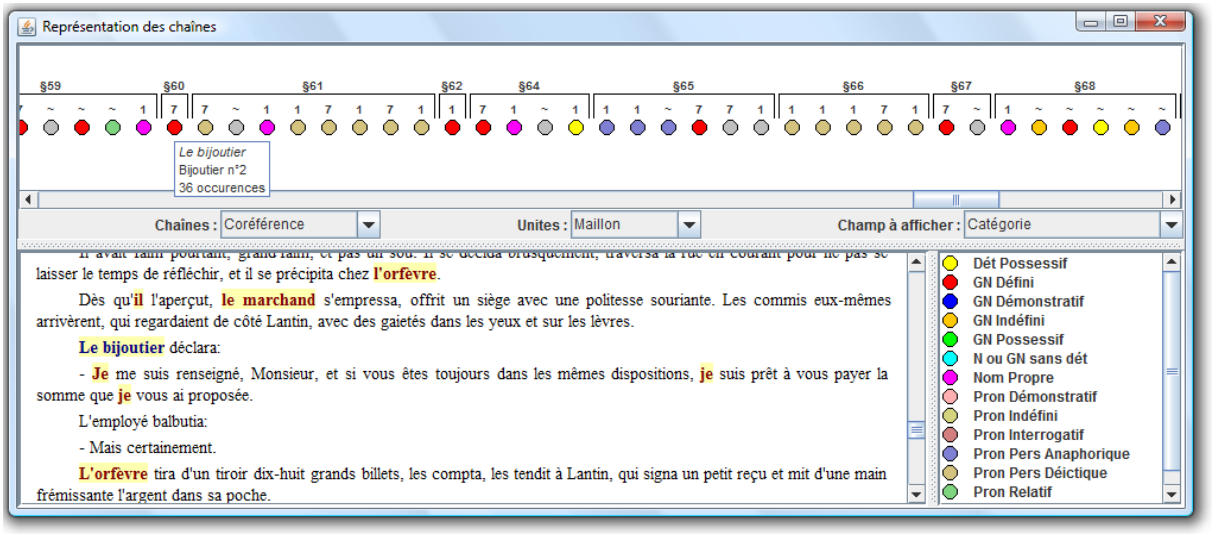

Figure 2. Interface d'ANALEC dédiée à la visualisation graphique de la suite des références d'un texte, en exploitant plusieurs codes graphiques.

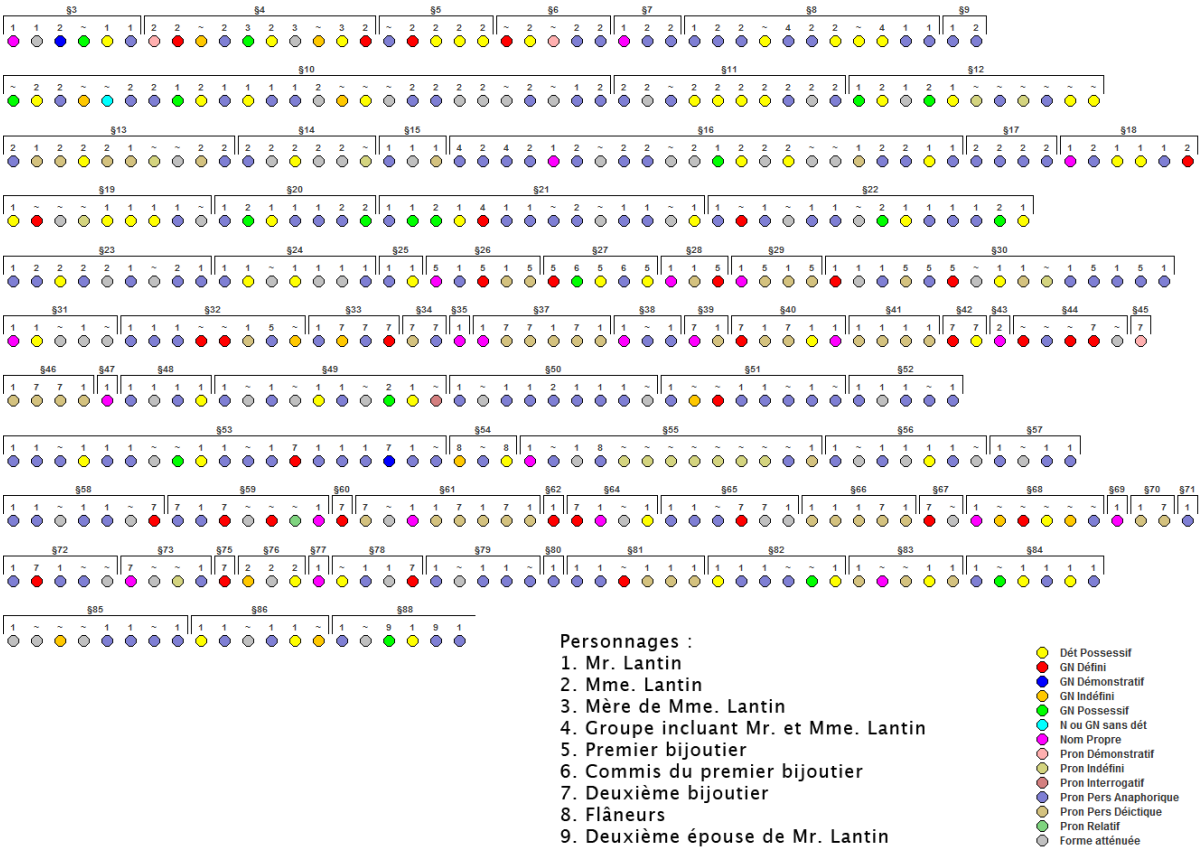

Figure 3. Suite complète des références de la nouvelle « Les bijoux » de Maupassant, avec un code couleur lié au type d'expression référentielle.

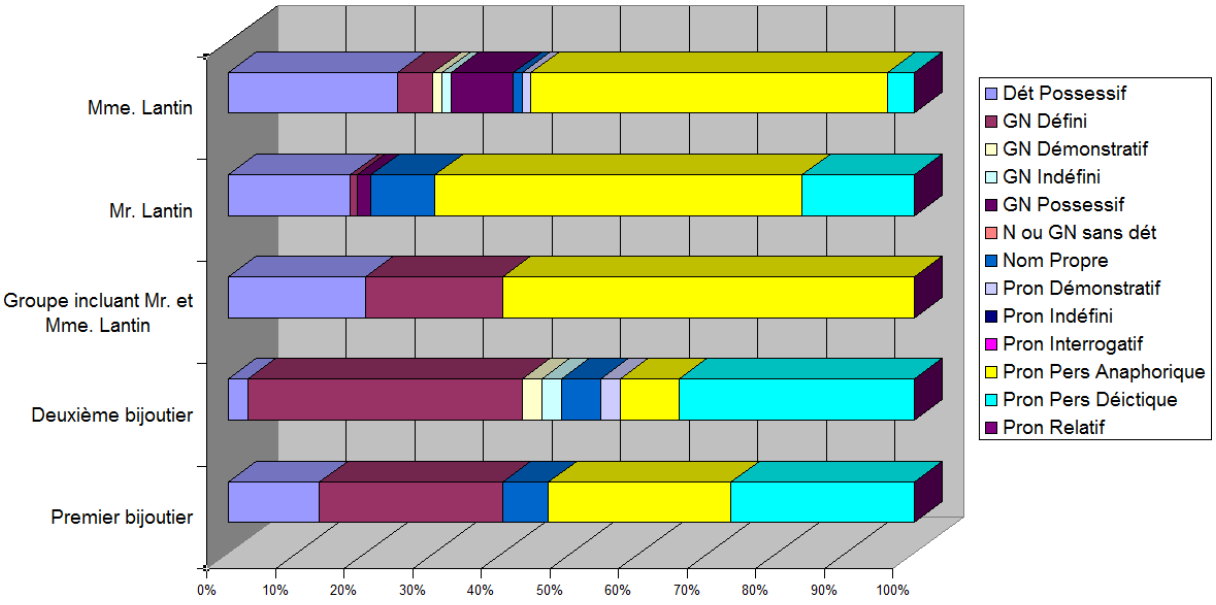

Figure 4. Représentation graphique, produite par MS Excel, de la répartition des types d'expressions référentielles pour les personnages les plus cités dans « Les bijoux », à partir d'un tableau de corrélation généré dans ANALEC et exporté en CSV.

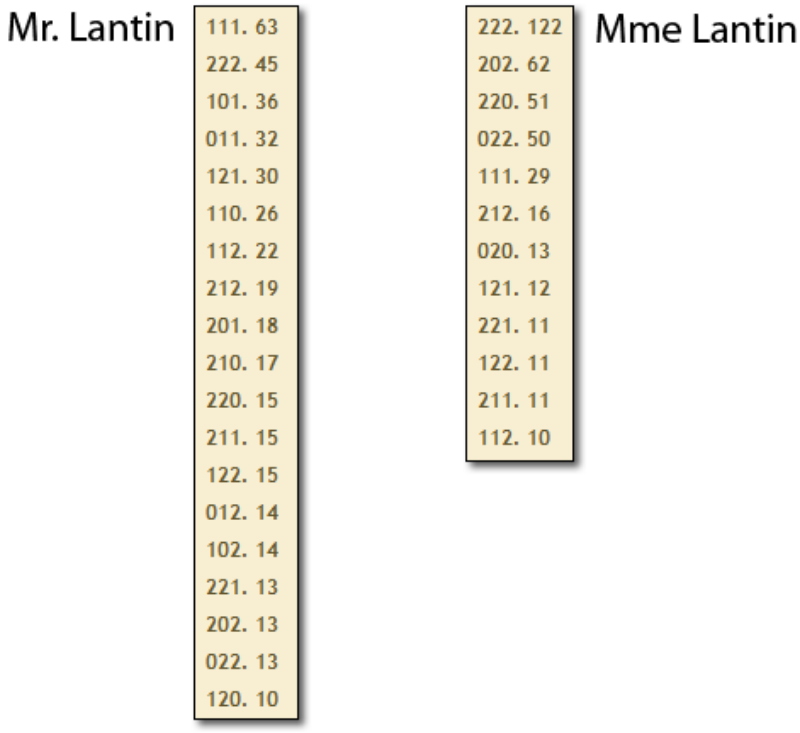

Figure 5. Analyse en trigrammes glissants de la suite des références à Mr. Lantin et de celle correspondant à Mme. Lantin dans la nouvelle « Les bijoux ». Dans les deux cas, la première colonne indique les trigrammes, avec un code dans lequel « 1 » vaut pour le personnage concerné, « 0 » pour un changement de paragraphe, et « 2 » pour une expression référant à un autre personnage, quel qu'il soit. La deuxième colonne indique les fréquences de chacun des trigrammes détectés (pour ceux qui apparaissent au moins à 10 reprises).

Figure 6. Analyses en trigrammes glissants puis en bigrammes glissants de l'ensemble des références de la nouvelle « Les bijoux », avec prise en compte des changements de paragraphe dans le premier cas et pas dans le second cas. Ici comme dans la figure précédente, les copies d'écran sont issues de l'outil en ligne DCODE (http://www.dcode.fr/analyse-frequences).