Utilisation de la visualisation en nuage arboré pour l'analyse littéraire

Delphine Amstutz [1], Philippe Gambette [2]

[1] CELLF – Université Paris-Sorbonne (Paris IV) – Paris – France

[2] LIRMM – Université Montpellier 2 - CNRS – Montpellier – France

Résumé

Les vertus heuristiques des outils textométriques sont de plus souvent sollicitées par l'analyse littéraire. Nous proposons un nouvel outil de visualisation qui s'intègre dans la démarche d'analyse de texte assistée par ordinateur : le *nuage arboré*. Il dispose les mots les plus fréquents d'un texte dans un arbre reflétant leur proximité sémantique calculée à partir de leur cooccurrence. La présence de couleurs permet de contraster les mots en fonction de leur saillance, leur localisation ou de leur cooccurrence avec un mot-cible. Cette nouvelle forme de visualisation, simple et parlante, a un intérêt méthodologique. Elle permet d'agencer et hiérarchiser le recours à d'autres outils déjà éprouvés. En effet, son interprétation peut générer un ensemble d'hypothèses, au delà d'une simple utilisation comme mise en forme esthétique de résultats quantitatifs. Nous présentons cette approche sur un extrait du corpus cornélien. La nouvelle version 1.3 du logiciel libre TreeCloud de construction de nuages arborés est utilisée en conjonction avec le logiciel Lexico3.

Abstract

Heuristic virtues of textometric tools are more and more acknowledged in literature analysis. We provide a new visualization tool to the process of computer assisted text analysis: the *tree cloud*. It displays the most frequent words of a text around a tree which reflects their semantic proximity computed from their cooccurrence. The use of colors helps contrasting the words depending on their salience, their location or their cooccurrence with some target word. This new visualization method, both simple and eloquent, has a methodological relevance, as it helps organizing and structuring the use of other well-known tools. Indeed, interpreting a tree cloud leads to a number of hypotheses, beyond the basic purpose of aesthetic expression of quantitative results. We introduce this approach on an extract of the Cornelian corpus. The new version 1.3 of the free software TreeCloud is used as well as Lexico3.

Keywords: textometry, visualization, literature analysis, tree cloud, cooccurrence.

1. Introduction

Le traitement des cooccurrences est une problématique majeure des logiciels d'analyse textuelle, car c'est à partir de ces présences communes au sein de portions du texte que l'outil informatique parvient à extraire des informations sémantiques. La question a été étudiée en profondeur, avec des études et variations sur le lieu des cooccurrences (phrase, paragraphe, fenêtre glissante, voir Martinez, 2003), les formules de calcul (Lafon, 1984 ; Evert, 2005) et les techniques de représentation : graphe de cooccurrences (Martinez, 2003) et lexicogrammes (Heiden, 2004), projections (Viprey, 2006) ou analyses arborées (Brunet, 2008).

C'est sur cette dernière méthode de visualisation de la cooccurrence, l'analyse arborée, et plus spécialement sa version décorée et colorée appelée *nuage arboré* (Gambette et Véronis, 2009),

que nous allons nous concentrer dans cet article, en montrant diverses utilisations possibles de l'objet simple et expressif qu'est l'arbre aux feuilles étiquetées par des mots.

Ainsi, nous n'allons pas cheminer à travers diverses fonctionnalités d'un même logiciel, comme le proposait Damon Mayaffre avec Hyperbase dans une analyse du discours électoral de Nicolas Sarkozy (Mayaffre, 2008), mais voir quelles analyses peuvent être menées en utilisant l'arbre comme origine ou complément de l'analyse, selon qu'il invite à recourir à d'autres outils informatiques ou à retourner au texte.

Illustrer de cas précis diverses utilisations de la visualisation en nuage arboré nous semble la meilleure façon de montrer l'intérêt qu'il y aurait à l'inclure à divers endroits des logiciels de textométrie existants.

Nous analyserons donc à cette fin deux pièces de Corneille : *Othon* et *Cinna* – sans toutefois ignorer les difficultés inhérentes à l'étude textométrique de textes théâtraux. En effet, dans les analyses textométriques portant sur les pièces dans leur globalité, la polyphonie énonciative est gommée. Le texte, pris d'un seul tenant, est envisagé comme l'énoncé d'un seul locuteur, l'auteur, et non comme le support d'actualisations multiples par des personnages, c'est-à-dire des acteurs, distincts les uns des autres. La textométrie ne prenant légitimement en compte que l'énoncé au détriment de l'énonciation du discours (Ducrot, 1984), des phénomènes textuels subtils relevant de la « « double énonciation » lui échappent : jeu sur l'implicite, (Kerbrat-Orrechioni, 1998), ironie tragique… Ces deux écueils sont moins des obstacles entravant l'analyse, que des limites méthodologiques bornant le champ d'un usage légitime des outils textométriques.

En fait, ces écueils deviennent des points d'appui de l'analyse textométrique si l'on complète judicieusement les incohérences apparentes de la représentation arborée par un retour au texte. Ce geste interprétatif caractérise notre approche comme une approche herméneutique (l'analyse statistique vient susciter, formaliser et/ou confirmer des intuitions subjectives). Elle se différencie ainsi d'une approche objective (« philologique ») du texte uniquement (Labbé et Labbé, 2005).

2. Le nuage arboré : simplicité, lisibilité et adaptabilité

2.1. Une hiérarchie à plusieurs niveaux

L'arbre est un système de classification naturel. Dès l'Antiquité, c'est un arbre enraciné (c'est à dire orienté d'un nœud racine vers ses feuilles) que Porphyre utilise pour classer animaux, végétaux et minéraux de façon hiérarchique. La classification arborée apparaît comme une dualité autour des concepts de regroupement et séparation : regroupement au sein d'un même sous-arbre, séparation de part et d'autre d'une arête de l'arbre. Ces deux aspects apparaissent également dans les versions non enracinées d'arbres.

C'est donc cet objet d'organisation classique que nous choisissons d'étiqueter, aux feuilles, par des mots de taille et de couleurs différentes. Restreindre l'étiquetage aux feuilles permet de gagner en lisibilité, tout comme les variations de taille inspirées du *nuage de mots*. Ce système de représentation est étrangement absent de la plupart des outils actuels de textométrie. Certes, Hyperbase ou Astartex proposent des visualisations de mots dont la taille de police est variable, mais ces variations restent discrètes et n'atteignent pas les rapports habituellement présents dans les nuages de mots. Ces derniers ont connu le succès avec Internet et le web 2.0, où leur simplicité de création automatisée et leur capacité d'attirer l'œil par des mots-clés percutants

sans “mots vides” leur donnent une fonctionnalité de résumé rapide autant que d’aide à la recherche de contenu. C’est d’ailleurs dans le contexte d’une amélioration des nuages de mots pour le web (Hassan-Montero et Herrero-Solana, 2006 ; Kaser et Lemire, 2007 ; Fujimura et al., 2008 ; Viégas et Wattenberg, 2008 ; van Ham et al., 2009) que la première visualisation en nuage arboré est apparue sur le blog de Jean Véronis en décembre 2007 [1].

Dans les nuages de mots, un autre type de hiérarchie entre donc en jeu : celle induite par les différences de taille de mots, les plus gros apparaissant comme le plus lisibles, et donc les plus importants. Ces différences proviendront simplement des fréquences des mots dans le texte, ou de critères plus subtils comme la spécificité des mots, dans le texte par rapport à un autre, ou par rapport à un corpus de référence.

En combinant les deux concepts comme montré en Fig. 2, nous superposons donc deux hiérarchies : l’une basée sur l’importance, et l’autre sur l’emboîtement des sens, calculées respectivement à partir de l’occurrence et de la cooccurrence. C’est le principe du nuage arboré implémenté dans le logiciel libre TreeCloud [2] (Gambette et Véronis, 2009). Cet outil propose deux possibilités pour choisir la taille des mots : un calcul direct à partir des fréquences, ou l’importation d’une liste de tailles personnalisées.

2.2. La sémantique des couleurs pour guider la lecture

Les variations de couleur s’ajoutent aux variations de taille de caractères pour rendre plus clair ou plus expressif le nuage arboré. Elles jouent alors le même rôle que dans un nuage de mots, et peuvent donc être utilisées de plusieurs manières. Le plus simple est la coloration hiérarchique, qui répète l’information déjà portée par la taille des mots, en colorant par exemple en rouge les plus fréquents du texte et en bleu pâle les mots plus rares, en passant par des nuances expressives, comme en Fig. 7. Un petit nombre de couleurs doit être utilisé pour éviter de perdre en lisibilité (les règles classiques d’ergonomie suggèrent que 7 est un maximum).

D’autres alternatives de coloration reflétant une certaine place du mot par rapport à l’ensemble du texte peuvent être choisies : grammaticale (noms en rouge, verbes en bleu, …) nécessitant un étiquetage morpho-syntaxique, chronologique (noms apparaissant au début du texte en rouge et à la fin en bleu, avec des nuances entre les deux)... Les colorations hiérarchiques et chronologiques sont disponibles dans TreeCloud, ainsi qu’une coloration de dispersion liée à l’écart-type de l’ensemble des positions du mot.

La couleur peut également concerner seulement une partie du texte, dans le cas d’une coloration autour d’un mot cible. Cette idée, apparue dans le logiciel Astartex (Viprey, 2006), consiste à colorer les mots en fonction de leur cooccurrence avec un mot en particulier. Si la coloration (variant du bleu au rouge) est calculée par une fonction affine de la cooccurrence, il est possible que le nuage devienne presque entièrement bleu ou presque entièrement rouge. Pour éviter cet écueil, ce n’est pas directement par rapport à la valeur de cooccurrence, mais par rapport au rang de chaque mot dans la liste classée par cooccurrences, que la couleur est calculée par TreeCloud.

Enfin, les informations de couleurs peuvent provenir de sources extérieures au texte : on peut penser à des colorations thématiques, ou une mise en relief d’éléments particuliers, comme certains noms propres : noms de personnages, lieux, etc.

[1] http://aixtal.blogspot.com/2007/12/actu-une-ferrari-dans-un-arbre.html.
[2] http://fr.treecloud.org.

2.3. Un outil de visualisation polyvalent

Le nuage arboré peut représenter l'ensemble d'un texte, en l'utilisant en intégralité pour calculer les cooccurrences sur les mots les plus fréquents, après élimination des "mots vides", mais d'autres données peuvent être utilisées en entrée pour répondre à des problématiques très diverses. Une possibilité est par exemple de ne pas fournir l'ensemble du texte en entrée de la chaîne de traitement, mais uniquement les contextes ou la concordance d'un mot, pour obtenir un nuage ciblé sur un mot. Dans le contexte d'analyse des pièces de théâtre, comme nous le verrons dans la section suivante, il peut être intéressant de se focaliser sur le discours d'un personnage pour en effectuer le nuage arboré. Nous avons donc préparé, pour ce contexte, un document tableur qui permet, à l'aide de formules diverses, d'extraire rapidement l'ensemble des répliques d'un personnage d'une pièce de théâtre au bon format (en l'occurrence, avec les balises de Lexico3). Le même document tableur fournit également, automatiquement, des histogrammes pour indiquer la longueur de chaque acte, chaque scène et chaque rôle [3].

Il est aussi possible de considérer le nuage arboré comme une visualisation améliorée d'un ensemble quelconque de mots associés chacun à une valeur quantitative. Par exemple les listes de mots spécifiques d'un texte calculées par le logiciel Lexico3, et fournies dans ce logiciel sous forme d'une liste triée par spécificité décroissante, peuvent être visualisées sous forme d'un histogramme, d'un nuage de mots (où la taille reflète la spécificité), ou d'un nuage arboré qui permet une meilleure vision d'ensemble du matériel lexical en insérant l'information sémantique extraite du texte (voir Fig. 7).

De même, les distances entre mots utilisées pour construire l'arbre peuvent refléter soit directement leur cooccurrence absolue, soit la significativité de leur cooccurrence par rapport à une cooccurrence attendue, en fonction de la formule mathématique choisie. Cette cooccurrence attendue peut être celle estimée d'après un modèle d'apparition des mots du texte, par exemple le modèle hypergéométrique (Lafon, 1984), ou d'après un corpus de référence. Dans ce dernier cas, de la mê me façon que des formules statistiques permettent d'estimer la significativité du nombre d'occurrences d'un mot dans un texte par rapport à un autre, des formules pour estimer la significativité de la cooccurrence de deux mots dans un texte par rapport à un autre pourraient être développées.

Autre réglage pouvant avoir des conséquences importantes sur le résultat, et donc le contexte d'utilisation de l'outil : le choix du lieu de la cooccurrence. En effet, on peut forcer TreeCloud à considérer que deux mots apparaissent ensemble s'ils font partie d'une même fenêtre de *n* mots de largeur. L'algorithme consiste alors à faire se déplacer une fenêtre glissante (le pas de glissement étant également paramétrable) le long du texte. Il est possible de choisir la valeur de *n* en prenant en compte divers critères objectifs : la quantité d'information contenue dans la fenêtre (Martinez, 2003) ou la stabilité de l'arbre construit après altération du texte (Gambette et Véronis, 2009), qui convergent généralement pour privilégier une petite fenêtre de cooccurrence (d'une dizaine de mots de largeur). Elle peut aussi se régler en prenant en compte les caractéristiques attendues du nuage arboré : une fenêtre de cooccurrence très petite aura tendance à bien rapprocher les collocations, à défaut de cooccurrences plus éloignées. L'autre alternative que nous avons ajoutée dans la version 1.3 de TreeCloud est de considérer que deux mots sont cooccurrents s'ils apparaissent dans un même bloc délimité par deux caractères (ou mots) séparateurs. On peut ainsi définir la phrase ou le paragraphe comme lieu de la cooccurrence, au moyen d'un prétraitement basique du texte.

[3] disponible en matériel supplémentaire de cet article sur http://theatre.treecloud.org.

Ces divers réglages sont donc principalement guidés par l'utilisation prévue du nuage arboré. Tout réglage conduit à un résultat qui reflète un certain signal présent dans le texte, c'est le cas en particulier des paramètres proposés par défaut. C'est lors de l'analyse du résultat que l'utilisateur de cette visualisation doit prendre du recul, en essayant quelques changements de paramètres (formule de cooccurrence, nombre de mots du nuage, principalement) pour vérifier si les tendances qu'il identifie dans l'arbre sont conservées, ce qui est par exemple le cas pour les regroupements de mots que nous analysons dans les nuages arborés de la section 3. C'est enfin le retour au texte qui permet la confirmation d'hypothèses nées de ces diverses « projections » du texte sur un arbre.

2.4. Une méthode modulaire et pratique

2.4.1. Une méthode modulaire

L'algorithme de construction d'un nuage arboré est modulaire, c'est à dire qu'il peut se découper en petites briques de traitement indépendantes, comme montré en Fig. 1. Ceci est particulièrement important quand il s'agit d'utiliser plusieurs sources de données pour répondre à des problématiques diverses comme mentionné dans la section précédente, et illustré dans la suivante.

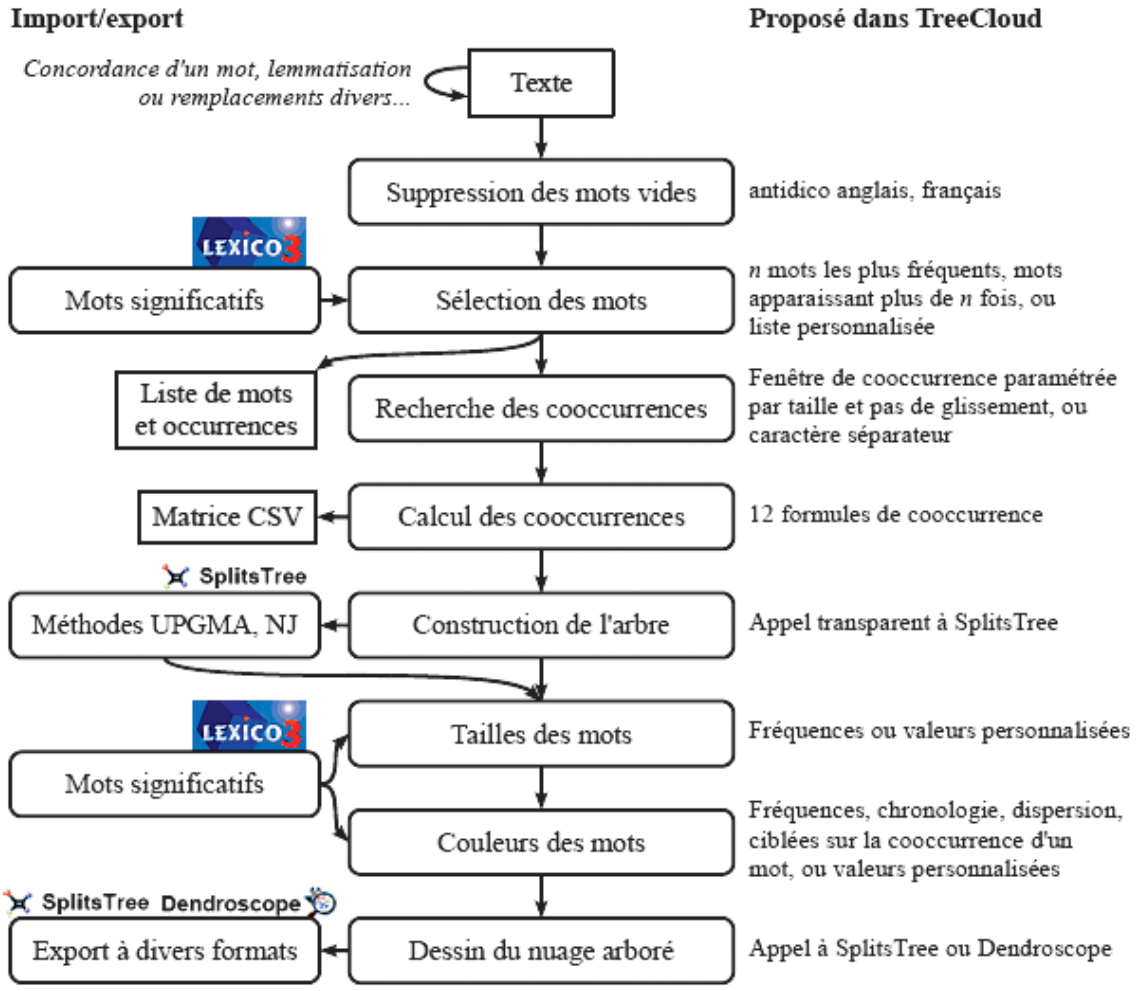

Figure 1 : La chaîne de traitements pour la construction d'un nuage arboré à partir d'un texte, et son implémentation dans le logiciel TreeCloud

Ceci est aussi primordial pour tester les différentes alternatives possibles, à certaines étapes. Par exemple, une douzaine de formules de cooccurrence sont implémentées, et bien séparer cette partie du code du logiciel permet de tester facilement ces diverses formules pour voir lesquelles assurent la meilleure stabilité de l'arbre construit à l'étape suivante (Gambette et Véronis, 2009).

Dans l'implémentation TreeCloud, plusieurs étapes de traitement créent un fichier de résultats qui peut être utilisé indépendamment. L'étape de sélection de mots renvoie un document tableur associant à chaque mot du texte son nombre d'occurrences, ce qui permet de créer un histogramme directement à partir du tableur, ou de fournir le fichier à un logiciel de création de nuages de mots. L'étape de calcul des cooccurrences fournit aussi un document tableur contenant l'ensemble des valeurs de cooccurrences, qui peuvent, après un tri des colonnes, livrer la liste des cooccurrents fréquents de chaque mot sélectionné. Enfin, quand le logiciel SplitsTree (Huson et Bryant, 2006) est appelé automatiquement par TreeCloud, on peut récupérer le fichier Nexus qu'il renvoie, et qui contient en particulier l'arbre au format parenthésé classique Newick.

Cette modularité de la technique a permis de la coder dans plusieurs langages de programmation, en réutilisant des briques logicielles déjà existantes par ailleurs (un logiciel complet, SplitsTree, ou des extraits de code de projets logiciels antérieurs). La version en Python de TreeCloud, très paramétrable et multi-plateformes, est disponible en tant que logiciel libre avec une interface graphique en Delphi pour Windows. Il fait appel au logiciel gratuit SplitsTree (multiplateforme, en Java) pour la construction et le dessin de l'arbre. Deux versions minimalistes, l'une en Delphi et l'autre en C et assembleur (avec interface web [4]), ont aussi été développées pour traiter l'intégralité du processus de création de nuages arborés, mais sont moins paramétrables et ne sont plus tenues à jour.

2.4.2. Une méthode pratique

De la même manière que de nombreux codes sources existent pour construire des nuages de mots (ou de tags, sur des sites web), on peut envisager que la visualisation en nuage arboré connaisse une intégration dans de nombreuses plateformes. Outre la modularité, la rapidité est un avantage des algorithmes mis en jeu. La recherche des cooccurrences demande un temps proportionnel à la longueur du texte multiplié par le carré du nombre de mots sélectionnés. Ainsi, pour des nuages d'une centaine de mots (la limite pour rester lisible sans outil de zoom nécessitant une exploration par zones), une vingtaine de secondes suffisent à la création du nuage sur un texte de 300 000 mots dans TreeCloud, et moins de deux dans le programme optimisé en C et assembleur.

On peut aussi noter que contrairement aux méthodes de visualisation des cooccurrences faisant intervenir des graphes, le problème du dessin esthétique d'un arbre non enraciné se résout très bien par l'algorithme EqualAngle (Felsenstein, 2004), et seul le placement automatique des étiquettes n'a pas de solution élégante naturelle (même si l'algorithme glouton proposé dans SplitsTree pour placer les étiquettes sans recouvrement fonctionne plutôt bien en pratique, surtout quand il est possible de modifier facilement le résultat de façon manuelle). Avec un graphe en revanche, de nombreux algorithmes, plus ou moins rapides, existent pour répondre au problème de placement des noeuds et de dessin des arêtes, mais aucun ne s'est encore distingué par la qualité ou l'uniformité de ses résultats, même si des outils récents proposent des solutions esthétiques pour de petits graphes (van Ham et al., 2009).

[4] Implémentation par Jean-Charles Bontemps disponible sur http://poulphunter1.free.fr/Cooccurrences.

3. Illustration : approche contrastée de *Cinna* et *Othon*

Afin de montrer comment l'analyse textométrique peut enrichir un type de travail universitaire qui lui reste habituellement hermétique, une thèse d'histoire littéraire et d'histoire des idées, nous procéderons à une étude de cas ciblée. Le choix d'un corpus restreint d'application n'est donc cohérent que si l'on le rapporte au projet d'ensemble dans lequel il s'intègre – à savoir l'étude de la figure du favori royal sous Louis XIII.

L'auteur emblématique de la période concernée est bien évidemment le dramaturge Pierre Corneille, dont les œuvres sont toutes numérisées. Le personnage du favori royal apparaît principalement dans deux de ses œuvres dramatiques, *Cinna* (1643) et *Othon* (1664). Dans ces deux pièces seules en effet, le rôle dramatique prépondérant du héros lui est conféré par son statut exceptionnel de favori du souverain. Dans ces deux pièces seules, le statut politique de favori est l'objet d'une mise en débat, au cœur même de l'intrigue et de la dynamique dramatique – ce qui n'est le cas ni dans *Polyeucte* (Sévère est favori de l'Empereur Decius, mais il s'agit là d'une attribution accessoire indiquée dans la *dramatis personae*), ni dans *Le Cid* (Rodrigue, au terme de la tragi-comédie, peut être considéré comme le favori de Don Fernand, mais il s'agit là d'une qualification déduite *a posteriori*) par exemple.

Cinna et *Othon* se prêtent tout particulièrement à une étude contrastive. Ecrite chacune à une période charnière de la carrière du dramaturge, considérée, chacune dans son ordre et de l'avis de l'auteur même, comme un sommet de la production cornélienne, *Cinna* et *Othon* mettent en œuvre deux visions antithétiques de la politique: glorieuse et sublime malgré la trahison dans *Cinna*, dysphorique et manœuvrière malgré la victoire dans *Othon*.

Le recours à la représentation arborée pour mener à bien cette comparaison devrait permettre de faire émerger des hypothèses de travail, d'étayer des intuitions générales, d'affiner l'analyse en suscitant des parallèles, des biais, que la simple lecture linéaire ou cursive des textes n'aurait pas suffi à déceler et articuler.

De prime abord (voir Fig. 2), le nuage arboré fait apparaître, pour chacune des deux pièces concernées, une caractérisation simple et pertinente des personnages, selon les rapports de force qui les opposent ou les relations qui les unissent. Ainsi dans *Cinna*, les personnages de Maxime et Cinna sont immédiatement perçus comme des « amis », Auguste comme un « tyran ». Dans *Othon*, les deux protagonistes sont définis en fonction de leur relation au pouvoir: Othon aspire à « régner » tandis que Galba, « empereur », détenteur du « pouvoir », tente de réguler les intrigues de la « cour ».

La représentation arborée qui rend compte de la totalité du texte de chaque pièce laisse également apparaître les linéaments des intrigues principales et secondaires. Le nuage arboré d'*Othon*, pièce « implexe », est de fait moins lisible que celui de *Cinna*, pièce simple. Dans *Cinna*, pièce de la fondation politique, c'est au terme d'une lutte acharnée pour la « liberté » de « Rome », qu'Auguste, d'abord essentiellement « tyran », devient, par un acte extraordinaire de clémence, digne du pouvoir et de la « grand[eur] » d'un César, aux yeux de Cinna et Emilie. Cette conclusion est confirmée par l'histogramme de la Fig. 3. L'arbre permet également de délimiter une sous-intrigue, celle de la vendetta familiale (Emilie se doit de venger la mort de son père, ancien tuteur d'Octave, assassiné sur l'autel d'Auguste).

Dans *Othon* en revanche, l'adjectif « grand », s'il est toujours central et étroitement associé à la sphère politique (il est au voisinage de « état » et de « maître »), côtoie cependant le mot « choix ».

Othon est en effet une pièce de la succession dynastique. Le pouvoir de Galba, l'« empereur » en place, se délite à mesure que les manœuvres de cabinet – les intrigues amoureuses au premier chef – pour lui trouver un successeur pullulent.

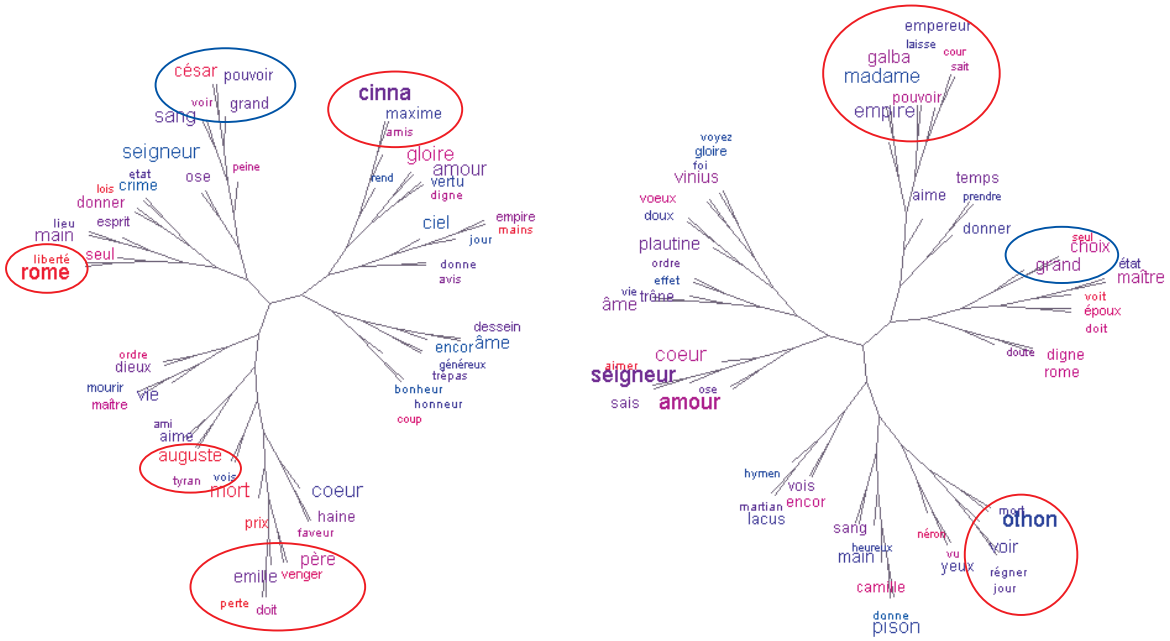

Figure 2 : Nuages arborés globaux des 60 mots les plus fréquents dans Cinna et Othon (distance Liddell, fenêtre de largeur 20), colorés chronologiquement (rouge au début, bleu à la fin)

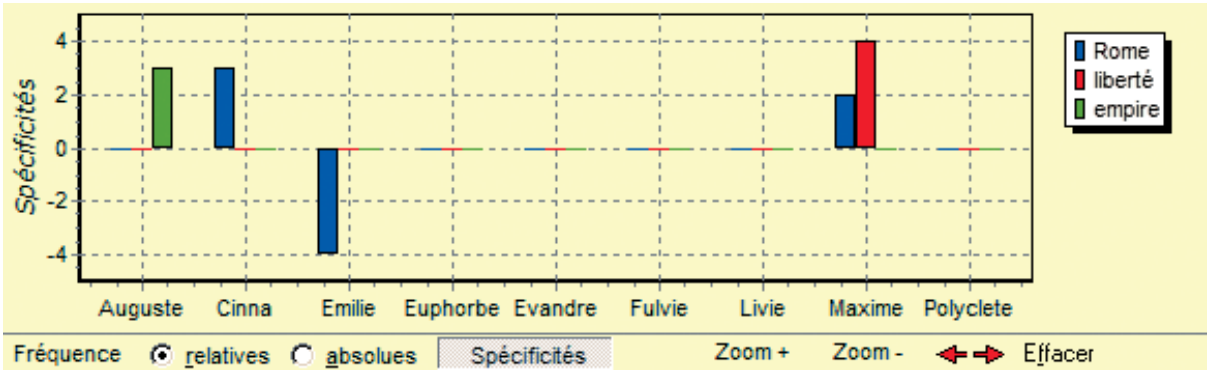

Figure 3 : Spécificités d'emploi de « Rome », « liberté » et « empire » chez les différents personnages de Cinna dans Lexico3

La représentation arborée des paroles d'Auguste en Fig. 4 permet d'isoler un pôle « euphorique » nettement isolé des autres ensembles lexicaux figurés, plutôt dysphorique. L'analyse plus précise de cet ensemble centré sur les relations amicales (« amis ») permet de dégager quelques conclusions intéressant l'idéologie politique de la pièce.

On relève en effet, à l'aide de la carte des sections Lexico3 de la Fig. 5, cinq occurrences du mot « ami » dont la répartition dans la pièce n'a rien d'aléatoire. Les occurrences se font de plus en plus fréquentes à mesure que l'intrigue se déroule. Le retour au texte permet de comprendre les différents emplois du mot « ami » dans le discours d'Auguste. Dans les trois premiers actes, le mot « ami » est avant tout employé comme un terme hypocoristique d'interpellation. Son usage est descriptif (occurrence 3) ou expressif (occurrences 1 et 2). Dans les deux derniers actes en revanche, le terme « ami » n'est plus employé par Auguste comme allant de soi. Le terme ne relève plus de l'apostrophe familière, il est commenté, posé et imposé par Auguste. La

quatrième occurrence, tout particulièrement, est une formule performative. Qualifier quelqu'un d'« ami » n'est plus une ressource disponible à tout un chacun: c'est la prérogative exclusive de l'empereur, l'espace politique se définissant, comme le dit Carl Jung, par la distinction fondamentale entre ami et ennemi. Décider de qui est ami appartient au pouvoir impérial, seul. C'est une création voire un acte politique et non plus une donnée naturelle (relevant de l'affection, de la passion). Les amis « naturels » deviennent, après l'épreuve de la trahison, des amis d'institution, des « favoris » politiques. L'étude du mot « faveur » dans l'ensemble du texte, suggérée par le nuage arboré de *Cinna* en Fig. 2, corrobore d'ailleurs cette analyse. Dans les premiers actes, le mot « faveur » apparaît dans des expressions lexicalement figées (« en ma faveur »). Il renvoie uniquement à un sens amoureux, personnel, alors que dans les trois dernières occurrences, il prend une coloration politique. A la faveur amoureuse et privée qui déstabilise l'univers politique (comme ferment de toutes les conspirations) succède la faveur politique, dispensée par le seul empereur, qui instaure un nouvel ordre politique stable et hiérarchisé.

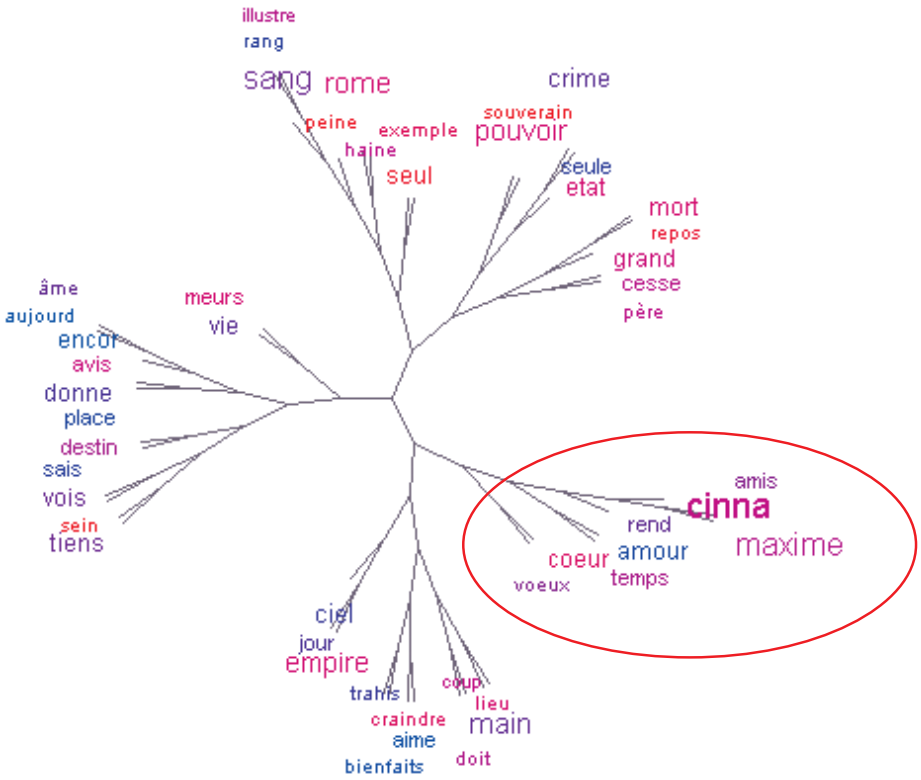

Figure 4 : Nuage arboré des 50 mots les plus fréquents des paroles d'Auguste dans Cinna

50

1. Voilà, mes chers amis, ce qui me met en peine.
2. Quoi ! mes plus chers amis ! quoi ! Cinna ! quoi ! Maxime !
3. Reprenez le pouvoir que vous m'avez commis, Si donnant des sujets il ôte les amis
4. Soyons amis, Cinna, c'est moi qui t'en convie
5. Il nous a trahis tous ; mais ce qu'il a commis Vous conserve innocents, et me rend mes amis.

Figure 5 : Carte des sections Lexico3 et contextes de « amis » dans les paroles d'Auguste dans Cinna

L'étude des deux nuages arborés de la Fig. 6 met en évidence deux polarités. Si « Galba » est associé, naturellement, au titre d'« empereur », c'est à « Othon » que revient la qualification de « maître ». En effet, Othon s'affirme comme le véritable « maître » de la Cour. Son pouvoir d'influence, authentique quoique officieux, bat en brèche l'autorité de Galba, fantoche vieillissant. Dans *Othon*, l'exercice du pouvoir ne repose plus sur la gloire, l'honneur, l'acte héroïque, mais sur la manipulation, la maîtrise de l'autre. L'espace politique est un échiquier sur lequel les pions se déplacent selon des alliances, matrimoniales ou politiques, selon des intérêts particuliers – sans considération du bien public: « Ce sont intrigues de cabinet qui se détruisent les unes les autres » (« Préface » d'*Othon*).

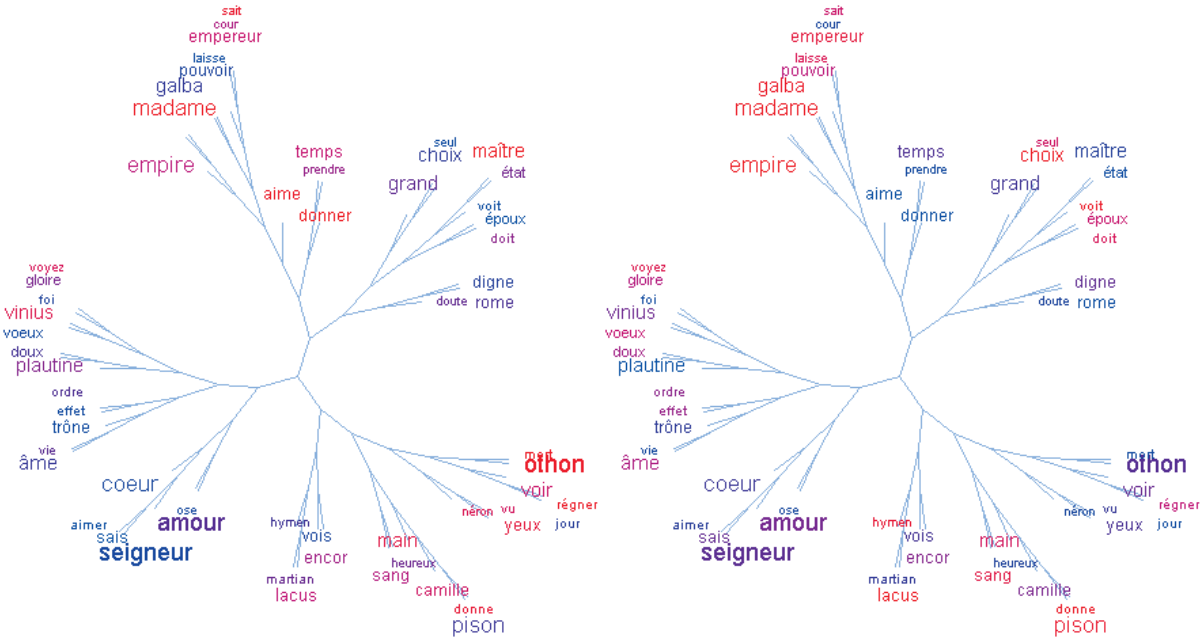

Figure 6 : Nuage arboré des 60 mots les plus fréquents de la pièce Othon, coloré à gauche par rapport aux cooccurrences avec « Othon », à droite par rapport à celles avec « Galba »

La caractérisation contrastive de chacune des deux pièces en Fig. 7, reprise dans le Tab. 1, rend manifeste la conception singulière de l'amour qui distingue *Othon* de *Cinna*.

L'amour héroïque est un motif mineur de *Cinna*, alors que trois ensembles régulièrement dispersés – « amour-cour-souffrir », « hymen-empire » et « trois-époux » – ressortent lorsque l'on figure le vocabulaire spécifique d'*Othon*. Dans la pièce de 1665, l'amour se dépouille de ses naïfs oripeaux sentimentaux, il est directement lié aux intérêts de la cour, où les alliances politiques s'expriment en termes d'engagement matrimonial. De *Cinna* à *Othon*, l'amour politique et galant succède à l'amour héroïque et romanesque. Il est un voile pudiquement jeté sur des conflits d'influence. « A la grandeur de Rome succèdent les grandeurs de Rome ; aux âmes fières et généreuses, les âmes souples, intéressées ; au portrait des héros, celui des politiques ; à leurs actions exemplaires, les ruses, les complots, les intrigues de cour. Une sorte de byzantinisme politique, mélange de finesses, d'audaces, de prudences, de sourdes machinations, s'y double d'un byzantinisme intime tout aussi délicat et retors. (…) Il faut remonter jusqu'à Pompée (l'épisode romanesque Cinna-Emilie-Maxime est d'une autre nature) pour trouver l'ébauche de ces amours politiques. (…) Othon est peut-être le portrait le plus achevé de toute la galerie. » (Nadal, 1948).

Le nuage arboré est une plateforme d'analyse, une forme de représentation qui suscite des propositions d'interprétation, distribue et hiérarchise le recours à des logiciels textométriques

complémentaires. Ses éventuelles défaillances sont palliées par un retour au texte constant et facilité. Ainsi l'étude comparative de *Cinna* (1643) et *Othon* (1665) a-t-elle montré les ressources nouvelles qu'est susceptible d'apporter à l'analyse littéraire cet outil, qui offre une approche originale et aisée d'objets divers (caractérisation des personnages, structure dramatiques ou actancielle, enjeux idéologiques, etc.).

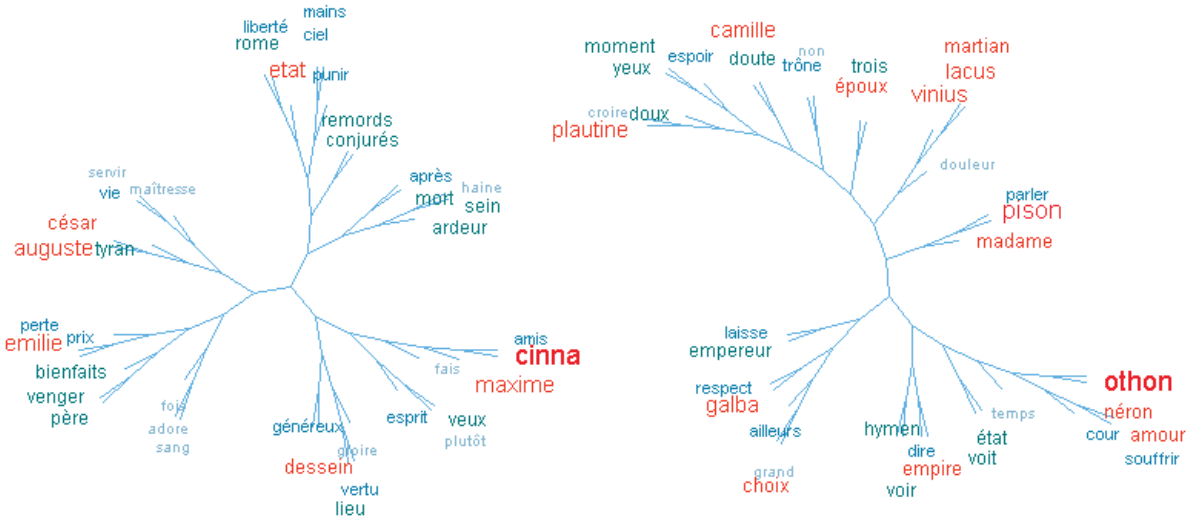

Figure 7 : Nuage arboré des mots spécifiques de Cinna et Othon, dimensionnés et colorés d'après leur spécificité calculée dans Lexico3

	Cinna	*Othon*
Lieu du pouvoir et objet de la confrontation entre les personnages	Rome («liberté»)	Empire («trône»)
Souverain en place	tyran	Empereur
Membres du corps politique	amis	maîtres / seigneurs
Moyens au service de la cause politique	gloire	amour matrimonial («amour», «hymen», «choix»)
Caractérisation de la pièce	Pièce de FONDATION	Pièce de SUCCESSION DYNASTIQUE

Tableau 1 : Tableau comparatif des pièces Cinna et Othon

4. Conclusion

Nous avons vu que la visualisation en nuage arboré constitue un outil intéressant à plusieurs niveaux pour l'analyse textuelle et littéraire. Pour une utilisation la plus pratique possible, il reste à améliorer l'interactivité des implémentations actuelles : soit en ajoutant des fonctionnalités de retour au texte au logiciel TreeCloud, soit en intégrant l'algorithme de construction d'un nuage arboré dans les outils de textométrie classiques.

Ceci constituera le premier pas vers l'identification d'autres problématiques où le nuage arboré pourrait apporter une réponse automatique, ou une ébauche de réponse incitant à approfondir l'exploration dans une démarche semi-automatique. Il serait également intéressant de voir comment le nuage arboré peut s'intégrer dans la méthodologie d'utilisation du logiciel Coocs, consacré à une analyse fine de la cooccurrence (Martinez, 2003).

Remerciements

Nous remercions les organisateurs et les formateurs de l'école thématique CNRS-SHS MISAT, qui est à l'origine de ce projet d'intégration de la visualisation en nuage arboré dans des analyses textométriques. Merci à Daniel Huson pour ses indications techniques ayant permis l'intégration de SplitsTree et Dendroscope dans la chaîne de traitement de TreeCloud, et à Virginie Lethier et Jean Véronis pour d'intéressantes discussions préliminaires à cet article. Ce travail a été soutenu par le projet ANR 08-EMER-011-01 (PhylARIANE).

Références

Brunet E. (2008). Les séquences (suite). In Heiden S. and Pincemin B., editors, *Proc. of JADT'08*.

Ducrot O. (1984). *Le Dire et le dit*. Editions de Minuit.

Felsenstein J. (2004). *Inferring Phylogenies.* Sinauer Associates.

Fujimura K., Fujimura S., Matsubayashi T., Yamada T. and Okuda H. (2008). Topigraphy: visualization for large-scale tag clouds. In *Proc. of WWW'08*.

Gambette P. and Véronis J. (2009). Visualising a text with a tree cloud. In Locarek-Junge H. and Weihs C., éditors, *Classification as a Tool of Research, Proc. of IFCS'09 (11th Conference of the International Federation of Classification Societies)*.

van Ham F., Wattenberg M. and Viégas F.B. (2009). Mapping text with phrase nets. In *Proc. of IEEE InfoVis'09.*

Hassan-Montero Y. and Herrero-Solana V. (2006). Improving tag-clouds as visual information retrieval interfaces. In *Proc. of InSciT2006.*

Heiden S. (2004). Interface textuelle à un espace de cooccurrences : implémentation dans Weblex. In Prunelle G., editor, *JADT 2004, Le poids des mots.* Presses universitaires de Louvain, pp. 577-588.

Huson D.H. and Bryant D. (2006). Application of Phylogenetic Networks in Evolutionary Studies. *Molecular Biology and Evolution*, 23(2): 254-267 [logiciel disponible sur www.splitstree.org].

Kaser O. and Lemire D. (2007). TagCloud drawing: algorithms for cloud visualization. In *Proc. of WWW'07*.

Kerbrat-Orecchioni C. (1998). *L'Implicite*. Armand Colin.

Labbé C. and Labbé D. (2005). How to measure the meanings of words? Amour in Corneille's work. *Language Resources and Evaluation*, 39: 335-351.

Lafon P. (1984). *Dépouillements et Statistiques en Lexicométrie*. Slatkine-Champion.

Martinez W. (2003). *Contributions à une méthodologie de l'analyse des cooccurrences lexicales multiples dans les corpus textuels.* Thèse de Doctorat en Sciences du Langage, Université Sorbonne Nouvelle Paris 3.

Mayaffre D. (2008). Quand "travail", "famille", "patrie" co-occurrent dans le discours de Nicolas Sarkozy. Etude de cas et réflexion théorique sur la co-occurrence. In Heiden S., Pincemin B., editors, *Proc. of JADT'08*.

Nadal O. (1948). *Le sentiment de l'amour dans l'œuvre de Pierre Corneille*. Gallimard.

Viégas F.B. and Wattenberg M. (2008). Tags clouds and the case for vernacular visualization. TagCloud drawing: algorithms for cloud visualization. *ACM Interactions*, 15(4): 49-52.

Viprey J.-M. (2006). Ergonomiser la visualisation AFC dans un environnement d'exploration textuelle: une projection géodésique. In Viprey J.-M., editor, *Proc. of JADT'06*, Besançon, PUFC.