Cooccurrences spécifiques et représentations graphiques, le nouveau " Thème " d'Hyperbase

Laurent Vanni, Adiel Mittmann

Cooccurrences spécifiques et représentations graphiques, le nouveau “Thème” d’Hyperbase

Laurent Vanni[1], Adiel Mittmann[2]

[1] BCL - UMR 7320 - CNRS - Université de Nice Sophia Antipolis (Nice, France) – lvanni@unice.fr

[2]Universidade Federal de Santa Catarina (Florianópolis, Brésil) – adiel@inf.ufsc.br

Abstract

Finding word cooccurrences and calculating the specificity scores is one of the most popular statistical methods in the analysis of textual data. Within Hyperbase, there is a “theme” feature for this purpose, which is capable of locating words that are used more commonly near a given word form, grammatical structure or lemma. The graphical representation of such an analysis is often challenging ; more than a list of the cooccurring words, it should be able to indicate the order, the score and the relations between pairs of words. Now that Hyperbase has a Web version, this article proposes a new approach for the “theme” feature : the calculation of cooccurrences has been extended to include the second level. The accompanying graphical representation is betting on new visual features in order to simplify reading the results and render the underlying calculation more explicit.

Résumé

Le calcul des cooccurrents spécifiques d’un mot est une des méthodes statistiques les plus populaires de l’ADT. Le logiciel Hyperbase a introduit cette notion avec la fonction “Thème” qui rend compte du lexique surutilisé autour d’une forme, d’une structure grammaticale ou d’un lemme. La représentation graphique d’une telle analyse est souvent loin d’être évidente. Plus que la simple liste des cooccurrents spécifiques, elle se doit de nous indiquer aussi l’ordre, l’écart ainsi que les relations entre chaque paire de mots. Avec l’arrivée de la version Web d’Hyperbase, nous proposons aujourd’hui une nouvelle approche de cette fonction. Le calcul se voit pour l’occasion approfondi, lui permettant d’identifier la cooccurrence de deuxième niveau. La représentation graphique, quant à elle, fait le pari de simplifier la lecture du résultat de cette analyse tout en explicitant plus précisément les calculs sous-jacents.

Mots-clés: polycooccurrence, visualisation, Hyperbase.

1. Introduction

L’évolution de nos outils informatiques ces 15 dernières années a contribué à façonner l’ADT. La tokenisation de textes de plusieurs millions d’occurrences est aujourd’hui digérée en quelques secondes par des processeurs extrêmement rapides. Le volume des données n’occupe que peu de place sur nos disques de plusieurs To. Les enjeux de l’ADT ont suivi ces évolutions, la taille des données est maintenant maîtrisée, la complexité des algorithmes récursifs de calculs statistiques n’est plus un frein et nous offre de nouveaux champs exploratoires. Les visualisations

graphiques ne se cantonnent pas à la simple illustration, mais au contraire remotivent une analyse plus profonde. Elles utilisent des affichages de réseaux de plusieurs milliers voire millions de liens dynamiques qui à la fois déstructurent et restructurent le texte pour en extraire différents parcours interprétatifs. Une autre forme de lecture non linéaire précise le sens de chaque mot dans son co(n)texte.

L'analyse des cooccurrences spécifiques est l'une des méthodes qui se prête le mieux au jeu des représentations graphiques du texte (Vanni et al., 2014). Quoi de plus naturel qu'une forme de réseau pour montrer les liens ou les attractions entre les mots ? Plus qu'une simple liste de cooccurrents, il convient de visualiser la topologie du corpus dans son ensemble pour apprécier les effets du co(n)texte d'un mot. Les méthodes graphiques permettent aussi de mettre en valeur plusieurs critères de cooccurrences en même temps. Si un lien semble caractériser la présence d'une relation entre deux mots, l'épaisseur du lien peut en complément indiquer l'indice de spécificité du lien.

À partir du calcul de cooccurrence spécifique déjà implémenté par Etienne Brunet dans Hyperbase et sa fonction historique "Thème", nous détaillerons comment la version Web du logiciel propose aujourd'hui de passer de la cooccurrence simple à la polycooccurrence (Martinez, 2012). Nous verrons ensuite une représentation graphique originale pour exploiter ces cooccurrences de façon dynamique et prolonger l'analyse pour approfondir l'approche coocurrentielle. Enfin, pour illustrer les résultats obtenus avec ce nouvel outil, nous utiliserons un corpus fabriqué à partir de l'ensemble des actes JADTs entre 2000 et 2014. Nous verrons avec la polycooccurrence dans Hyperbase en quoi l'évolution des moyens informatiques et l'apparition de nouvelles méthodes influent sur l'ADT et nos contributions scientifiques.

2. Spécificités et cooccurrences

2.1. Un calcul éprouvé

Le calcul des spécificités implémenté dans la plupart des logiciels d'ADT, repose sur le modèle hypergéométrique (Lafon, 1980) qui induit l'utilisation de 4 mesures fondamentales pour un mot donné dans un Corpus :

- T = taille du corpus,
- t = taille du texte,
- f = fréquence du mot dans le corpus,
- k = fréquence du mot dans le texte.

S'en suit la formule qui met en oeuvre un calcul de factorielles et retourne la probabilité pour un mot d'avoir la fréquence k dans un texte :

$$P(x=k) = \frac{f!(T-f)!t!(T-t)!}{k!(f-k)!(t-k)!(T-f-t+k)!T!}.$$

Converti en écart réduit, ce calcul rend compte de la sur-utilisation ou sous-utilisation d'un mot par rapport à la norme (usage moyen) dans un texte. C'est ici que le concept de texte a été étendu pour être adapté à la cooccurrence. Le texte n'est plus un ensemble de paragraphes qui

constituent une partie du corpus, mais devient avec la cooccurrence l’ensemble des passages qui contiennent le mot pôle recherché (Figure 1).

Mis bout à bout, les passages [1] concernés par un mot forment un texte. Pour chaque mot, nous avons ainsi les quatre mesures requises pour le calcul hypergéométrique : la taille T du corpus, la taille t du texte (le nombre de mots au côté du mot pôle dans le corpus), la fréquence f du mot dans tout le corpus et la fréquence k du mot au côté du mot pôle.

2.2. Implémentations et choix utilisateur

Par défaut ce calcul est utilisé sur les formes graphiques [2] du texte. Avec Hyperbase Web nous avons profité de l’analyse morphosyntaxique produite en amont du logiciel pour étendre le champ des possibilités. L’utilisateur peut ainsi analyser la cooccurrence sur les formes mais aussi sur les lemmes et les codes grammaticaux. Cette couche introduite par la lemmatisation et l’étiquetage du texte peut être appliquée à plusieurs niveaux. Il est possible de fixer la nature du mot pôle recherché indépendamment de celle de ses cooccurrents analysés. Nous avons laissé libre à l’utilisateur de croiser ces différentes propriétés linguisitiques. Un mot pôle peut donc devenir un code grammatical et ses cooccurrents des lemmes. Les possibilités sont nombreuses et permettent d’affiner la recherche de la cooccurrence.

Avec cette nouvelle implémentation de la fonction “Thème” dans Hyperbase nous avons introduit d’autres paramètres que l’utilisateur peut aussi faire varier. Le filtrage des mots par leur catégorie grammaticale est un exemple. Sur demande, le corpus peut être ramené à une catégorie de mot précise comme les Noms, les Adjectifs ou encore les Verbes. En associant plusieurs catégories, on peut aussi sectionner un vocabulaire plus large et demander par exemple l’ensemble des substantifs présents dans le texte. Dès lors, les valeurs T et t sont modifiées ainsi que les résultats du calcul des spécificités. Ce filtrage par catégorie grammaticale permet d’alléger un texte et de réduire le nombre de mots à analyser tout en conservant la notion de contexte et de linéarité du texte.

Les bornes de cette cooccurrence sont fixées par défaut sur une fenêtre définie par le paragraphe. Cette taille de contexte peut être modifiée à volonté pour passer d’un champ large de plusieurs dizaines de mots autour du mot pôle à un champ beaucoup plus restreint de quelques mots favorisant la cooccurrence syntaxique. L’interface d’Hyperbase Web permet donc à l’utilisateur qui le souhaite de saisir un entier correspondant au nombre de mots avant et après le mot pôle recherché.

Ces différents paramètres s’intègrent dans une fenêtre d’options associées à chacune des fonctions d’Hyperbase Web. Fixés à des valeurs par défaut, ils permettent d’appréhender la cooccurrence simplement et intuitivement. Cependant pour un utilisateur avancé ils autorisent à répondre à des hypothèses de travail plus complexes et restrictives. Fort de ces nouvelles options, nous avons aussi étendu la fonction “Thème” pour dépasser la cooccurrence simple. À

1. Le terme passage ici fait référence, par défaut, au paragraphe. Il est marqué dans un fichier texte par le “retour chariot”, en informatique le terme exacte est “caractère de fin de ligne”. Ce paragraphe peut aussi être changé au profit d’une fenêtre défini par un nombre de mots avant et après le mot pôle

2. La forme graphique est le mot tel qu’il est écrit dans le texte

l'occasion de l'implémentation de la version Web d'Hyperbase, nous nous sommes intéressés à la polycooccurrence (Martinez, 2003, 2012).

2.3. Polycoccurrence spécifique

Il convient de noter que dans la littérature, la polycooccurrence (Jean-Marc Leblanc, 2006; Martinez, 2012) a un proche parent : la cooccurrence de deuxième ordre ou d'ordre supérieur (Grefenstette, 1994). L'idée des coccurrences d'ordre supérieur est de prendre chaque cooccurrent comme point de départ (mot pôle) d'un nouveau calcul de cooccurrence sur l'ensemble du corpus. Elles ont été appliquées, par exemple, à la similarité sémantique (Lemaire et Denhière, 2006) et à l'analyse de monosémie (Bertels et Geeraerts, 2012). Cette méthode a été remplacée dans Hyperbase Web par la polycooccurrence (Martinez, 2003, 2012) qui correspond à l'idée originelle de la fonction "Thème" qui vise à explorer la sémantique d'un mot pôle au travers de ses cooccurrents. Le contexte ici est donc toujours lié au mot choisi au départ.

Pour atteindre la polycooccurence, le calcul hypergéométrique (Section 2.1) est prolongé par un appel récursif de la méthode. Chacun des cooccurrents de la liste est utilisé pour définir un nouveau contexte plus restreint autour du mot pôle. Un nouveau texte est extrait du corpus, fabriqué à partir de tous les passages qui contiennent à la fois le mot pôle et le cooccurrent sélectionné (Figure 1). Dans ce nouveau contexte, un indice de cooccurrence de deuxième niveau est alors calculé pour chacun des mots présents. Ainsi, tout cooccurrent du mot pôle se voit doté lui aussi d'une liste de mots spécifiques. Ces nouveaux liens de spécificités entre les mots forment la polycooccurrence dans Hyperbase.

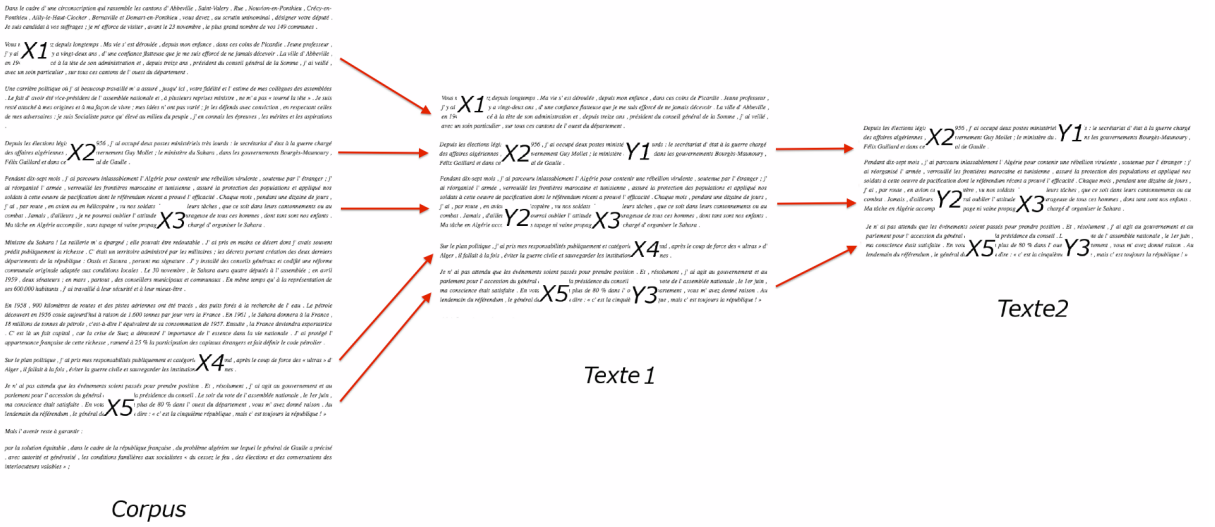

FIGURE 1 – Sélection du contexte du mot pôle X, puis du contexte de Y cooccurrent de X.

Cette couche de polycooccurrence rend le résultat assez complexe à évaluer au premier coup d'œil. Lorsque l'on invoque la fonction "Thème" d'Hyperbase Web, une liste de mots nous est présentée avec un indice de spécificité pour chaque mot. La cooccurrence de deuxième niveau ne peut pas être présentée sous cette forme, la combinatoire qui en résulte est bien trop

volumineuse pour la structurer avec une simple liste. C’est ici que les visualisations graphiques dynamiques peuvent se révéler précieuses pour apprécier le réseau de liens concurrentiels et mettre en valeur différents parcours interprétatifs d’un simple survol de la souris.

3. Représentations graphiques

La proposition de visualisation que nous faisons ici vise à montrer la cooccurrence et la polycooccurrence ensemble, comme la Figure 2. À la différence de (Martinez, 2012), cette visualisation est capable de traiter un grand nombre de mots en même temps. Afin de tirer parti de la puissance disponible dans les navigateurs modernes, la bibliothèque D3.js (Bostock et al., 2011) a été utilisée. D3 propose des outils javascripts modernes pour répondre au besoin de visualisation graphique en ligne. Cette librairie est fournie avec des exemples “clef en main” que l’on peut modifier [3], et des commandes de base qui permettent de créer nos propres visualisations. Jusqu’ici, aucune implémentation avec D3 ne permettait d’afficher un graphique répondant au aux besoins de la polycooccurrence. Nous avons donc créé une toute nouvelle visualisation avec cette libraire.

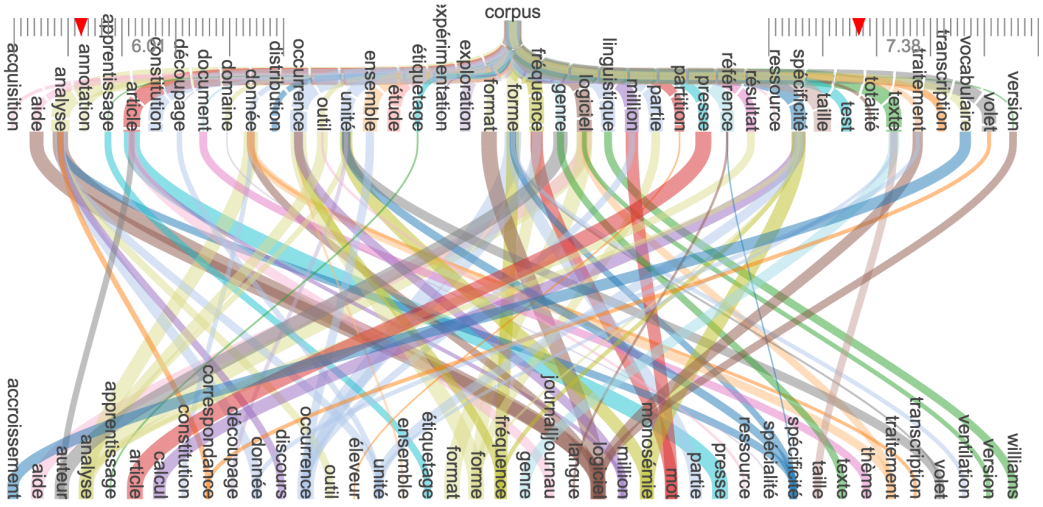

FIGURE 2 – Polycooccurrence : visualisation en couches dynamiques.

Nous avons choisi une visualisation en couches successives qui affiche le mot pôle en tête du graphique, la cooccurrence simple dans la première couche et la polycooccurrence dans la seconde. Les mots de premier ou second niveau sont filtrés en fonction de seuils de spécificité indépendants. Lorsque les seuils sont ajustés [4], les changements ont lieu au moyen d’une animation fluide qui commence immédiatement.

Dans chaque couche, les mots sont affichés de gauche à droite, dans l’ordre alphabétique. S’il n’y a pas assez de place pour les afficher horizontalement, ils sont tournés de 90 degrés vers

3. Cette méthode a déjà été utilisée dans Hyperbase pour afficher les graphes des spécificités et la distribution des mots dans le corpus

4. Les seuils de spécificité peuvent être ajustés graphiquement au moyen d’un curseur que l’on déplace sur une règle.

la droite ; si cela ne suffit toujours pas, la taille de la police est réduite suffisamment pour que chacun des mots soit présenté sans superposition.

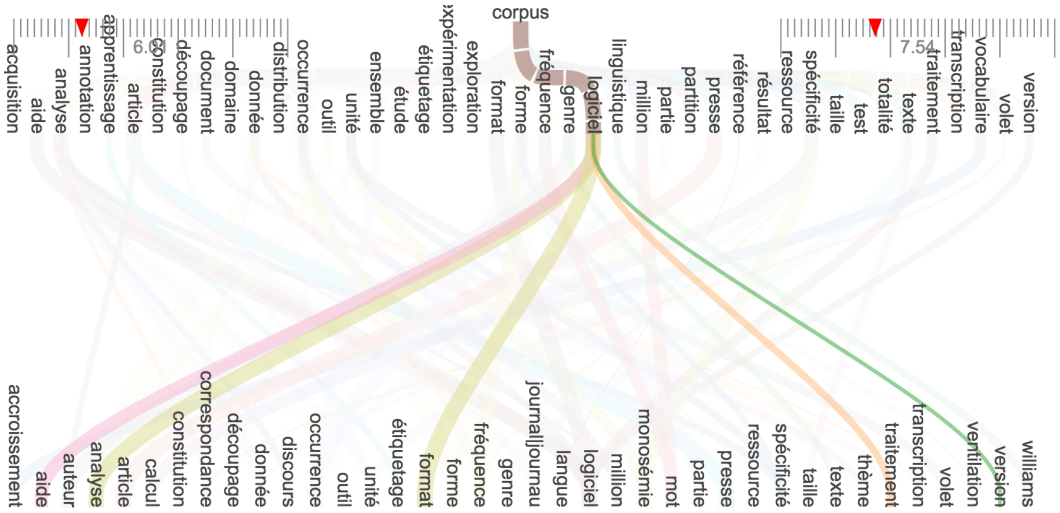

FIGURE 3 – Filtrage du graphique au survol de la souris.

Les lignes qui relient chacun des mots représentent la spécificité du mot de la couche inférieure par rapport au mot de la couche supérieure. Ces lignes sont donc orientées, mais aucune flèche n'a été ajoutée par souci de lisibilité. Le graphique se lit naturellement de haut en bas.

Si le mot pôle choisi par l'usager est très fréquent et que les seuils sélectionnés sont bas, la quantité de lignes dans la visualisation peut devenir visuellement écrasante. Pour que l'usager puisse explorer les informations disponibles confortablement, nous avons mis en place un système de filtrage dynamique au survol de la souris. Ce système met en évidence les associations d'un mot au milieu de centaines d'autres en effaçant les liens inutiles. Un exemple est donné sur la Figure 3. Les mots sont positionnés de manière séquentielle pour permettre de balayer les associations en succession rapide et d'identifier des résultats pertinents dans la masse de données affichées par le graphique.

En sélectionnant, avec la souris, un mot de la première couche et un autre de la seconde, la visualisation adopte un mode secondaire d'affichage (Figure 4). Ici le mot choisi dans la première couche et le mot pôle se placent côte à côte en haut du graphique. La première couche est alors destinée à faire apparaître tous les cooccurrents spécifiquement liés aux mots sélectionnés. La deuxième couche est réservée au second mot sélectionné. Le but ici est de mettre en évidence tous les liens spécifiques qui existent entre le mot pôle et deux de ses cooccurrents en occupant tout l'espace disponible. Le graphique propose alors plusieurs chemins interprétatifs à l'utilisateur par l'intermédiaire d'autres mots spécifiques de l'environnement des mots choisis. Afin d'éviter toute confusion, les liens de cooccurrence directe (issue directement du mot pôle) sont affichés dans un motif en pointillés, tandis que ceux issus de la polycooccurrence sont solides.

Dans cette visualisation, l'écart réduit [5] est représenté par l'épaisseur des traits. Cette épaisseur

5. Pour rappel, l'indice de spécificité présenté en 2.1 est converti dans Hyperbase en écart réduit.

p, en pixels est calculée en fonction de chaque écart e par la formule :

$$p = m\ln(1+e-\theta_i)+k,$$

où m représente un facteur d’échelle et θ_i le seuil de spécificité de la couche supérieure. Le paramètre k représente l’épaisseur minimale qui se produit lorsque $e = \theta_1$ et $p = k$. Cette formule garantit que l’épaisseur reste informative et visuellement agréable sur un large éventail d’écarts réduits et de seuils.

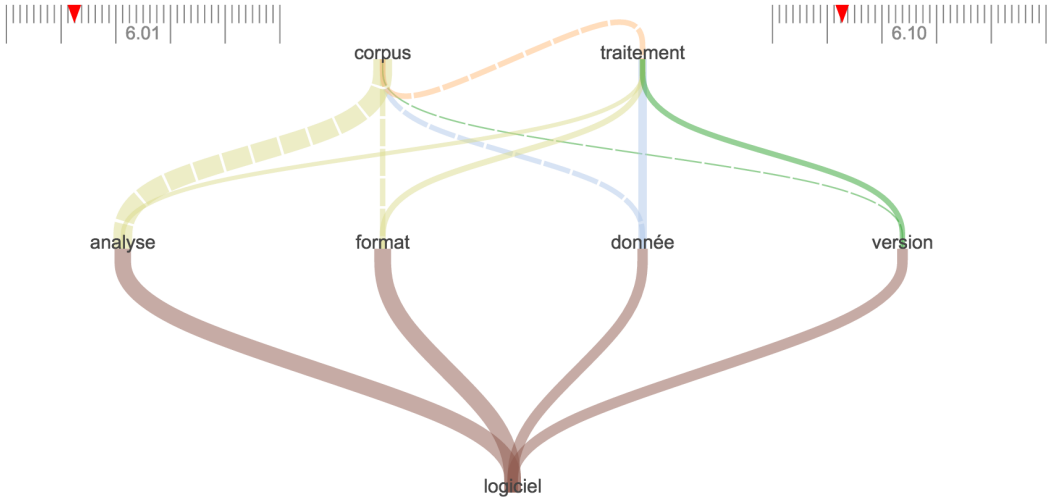

FIGURE 4 – Zoom sur l’environnement lexical des mots “traitement” et “logiciel” autour du mot pôle “corpus”.

En fixant des seuils différents pour la première et la seconde couche, l’usager peut produire une visualisation significative et informative. Toutefois, la formule qui prend en compte le seuil de spécificité implique des mesures différentes pour les épaisseurs des deux ensembles de lignes : étant données une ligne du mot pôle à la première couche et une autre ligne de la première couche à la seconde, une épaisseur identique ne signifie pas que les mots cibles ont le même écart réduit. Cette épaisseur est à mettre en relation seulement avec les liens de la même couche. La valeur exacte de cet écart est affichée dynamiquement au survol de la souris sur le lien de cooccurrence. Les couleurs des liens sont arbitraires et sélectionnées parmi une liste de couleurs prédéfinies. Une fonction de hachage [6] est appliquée au mot cible pour déterminer la couleur du lien choisie. Cette méthode garantit une bonne répartition des couleurs pour optimiser le contraste et la lisibilité du graphique.

Cette forme de visualisation de la cooccurrence et de la polycooccurence nous permet d’explorer nos corpus sous un angle nouveau. Là où la simple cooccurrence s’arrête sur le lien direct entre deux mots, le graphique présenté ici nous propose d’aller plus loin et d’identifier des différences de “Thème” qui n’apparaissent qu’à la lumière de la polycooccurrence. Ainsi pour

6. Une fonction de hachage en informatique calcule une empreinte unique de la donnée passée en paramètre. Elle garantit une répartition homogène des valeurs retournées dans un intervalle défini.

deux parties distinctes d'un même corpus, même si nous constatons les mêmes cooccurrents d'un mot pôle, nous pouvons différencier leur valeur sémantique en analysant les polycooccurrents associés. L'exemple qui suit nous présente ce cas de figure dans le corpus des JADTs.

4. Cas d'utilisation

Nous avons vu en 3 que les représentations graphiques sont des outils indispensables pour révéler toutes les ramifications d'un résultat statistique tel que celui de la cooccurrence spécifique. L'ADT en général a-t-elle pour autant pris ce virage graphique ces dernières années ? Fait-elle le pari de la visualisation graphique lorsque les données deviennent plus complexes à représenter ? Pour le savoir nous avons pris l'ensemble des articles publiés aux JADTs entre 2000 et 2014, soit 14 ans et une taille de corpus de 2,5 millions de mots pour vérifier si les pratiques de l'ADT ont changé au cours de ces dernières années.

Pour cette analyse diachronique, nous avons séparé le corpus en deux périodes. Une première entre 2000 et 2006 qui représente une montée en puissance des outils informatiques et l'apparition de nouveaux moyens de communication. Et une deuxième période qui couvre les années de 2008 à 2014, où l'accès aux ressources textuelles a été dopé par l'utilisation d'internet. Le web a ainsi ouvert l'accès à un volume d'informations encore jamais atteint jusqu'à présent.

Au début des années 2000, le calcul des cooccurrences spécifiques du lemme "donnée" (présenté sur la Figure 5) nous apprend que ces "données" sont fortement associées au lemme "traitement" (+14). Le traitement est déjà automatique ou semi-automatique et assisté par des "logiciels" (+9,85). Il s'appuie aussi sur les "méthodes" (+8,92) historiques de l'ADT. Nous sommes en présence d'une approche traditionnelle où la qualité du corpus prime sur la quantité. Les outils statistiques servent avant tout à éprouver une hypothèse de travail. On est encore loin d'une approche heuristique. La "fouille" de données est une pratique débutante et encore peu répondue. Ce terme fait une apparition timide (+4,74) dans la liste des cooccurrences qui montre un intérêt naissant pour ce type d'analyse qui nécessite un volume de données importantes.

Dans les années 2008 à 2014, la taille des corpus traités s'est décuplée, le cap de la "fouille" de données a été franchi (+14,61). Elle détrône même le simple "traitement" (+8,31) avec un score qui place cette approche parmi les trois premiers cooccurrents les plus spécifiques du lemme "donnée". Avec la démocratisation d'internet et le libre-accès à la ressource textuelle, les corpus se sont vus enrichir de plusieurs millions de mots. Ils intègrent maintenant de nouvelles formes de données issues de diverses sources tels que les journaux numériques, les blogs, les tweets et autres réseaux sociaux.

Comment la communauté ADT se propose-t-elle de représenter ces nouvelles ressources ? Nous observons que le lemme "visualisation" (+8,64) est maintenant l'un des 10 cooccurrents les plus spécifiques de la liste. Ce changement méthodologique a même fait reculer l'usage du mot "méthode" dans les cooccurrents directs du mot "donnée". Pour comprendre cette évolution, il est nécessaire de poursuivre notre analyse avec d'autres concepts très présents dans les publications des JADTs, à savoir le "corpus" et le "texte". Nous allons maintenant regarder la polycooccurrence autour de ces mots pour en extraire quelques éléments de réponse.

Que ce soit du point de vue du "texte" ou du "corpus", une question récurrente qui se pose en

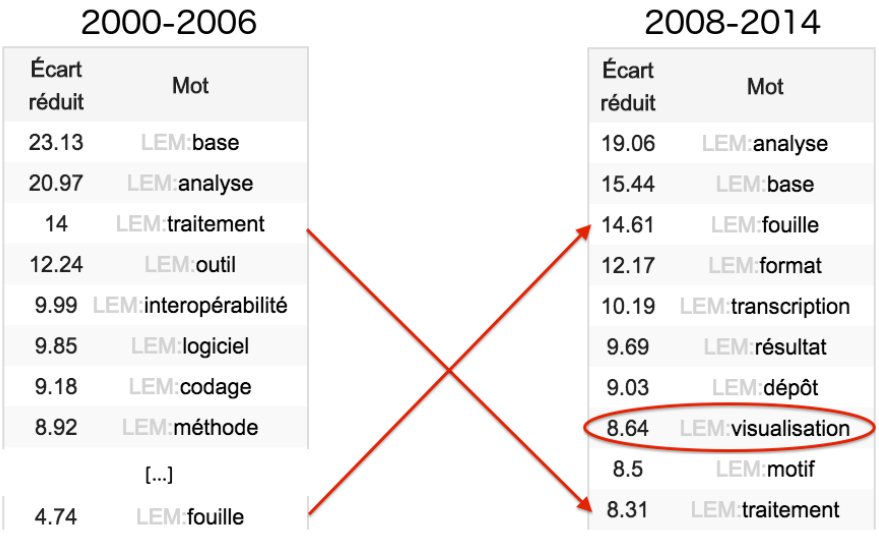

FIGURE 5 – Extrait des spécificités du lemme “donnée” dans les 2 parties du corpus.

ADT est la “taille”. En d’autres termes le nombre de mots que l’on souhaite analyser. Avec un score de +11,38 en 2000-2006 et +8,63 en 2008-2014 autour du mot pôle “corpus”, nous avons voulu observer la polycooccurence pour distinguer les différences entre ces 2 périodes. Les Figures 6 et 7 nous montrent, respectivement en première et deuxième période, la polycooccurrence du mot “taille”. Ce qui apparait alors semble vérifier la différence de méthodologie qui était pressentie par la liste des cooccurrents de la Figure 5. La “taille” dans les années 2000-2006 (Figure 6) faisait écho à l’“hétérogénéité” du corpus. Le passage à une large échelle avec un volume important de mots était sujet à discussion autour de la composition du corpus. Le problème de l’hétérogénéité qui introduit des biais interprétatifs était donc étroitement lié à la question de la taille. Avec la Figure 7 nous constatons que les nouvelles ressources numériques issues du web de 2008 à 2014 ont redistribué les problématiques et les questions liées à l’ADT.

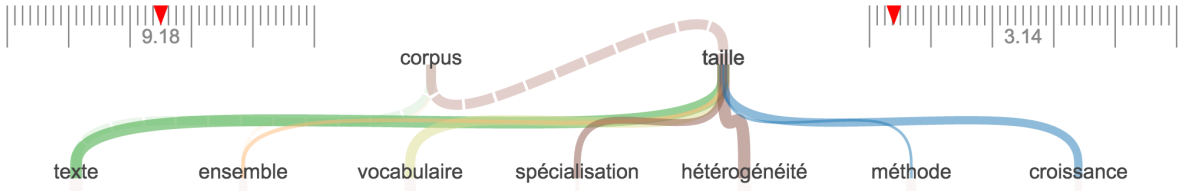

FIGURE 6 – Polycooccurrence du mot “taille” dans l’environnement du mot pôle “corpus” - période 2000 à 2006.

À partir de 2008 et jusqu’en 2014, le problème de l’hétérogénéité a laissé la place aux “millions” de “mots” (Figure 7). À l’aire du tout numérique, les ressources textuelles en ligne permettent d’atteindre des tailles de corpus démesurées. Déjà en 2012, Étienne Brunet (Brunet, 2012) proposait d’étudier 44 milliards de mots (offerts par google books) avec son logiciel Hyperbase.

Puis en 2014 nous proposions de revenir sur ce corpus (Brunet et Vanni, 2014) pharaonique gonflé à 86 milliards de mots. Certes nous avions critiqué l'hétérogénéité nécessairement associée à ce volume et les erreurs d'étiquetage morphosyntaxiques produites par une annotation rapide. Mais nous étions forcés de constater qu'il était désormais possible d'analyser l'évolution de la graphie des mots du français sur les 200 dernières années avec nos outils statistiques habituels.

Ce changement de priorité s'observe aussi dans le graphique par la présence du lemme "résultat" qui a pris place dans le contexte de "corpus" et de "taille". Si l'on considère que ces résultats s'enrichissent de graphiques comme nous laisse entendre le lemme "visualisations" (cooccurrent de "donnée"), nous avons là un moyen d'explorer ces nouveaux corpus. L'hétérogénéité devient ici une question qu'il est possible de mesurer graphiquement. À titre d'exemple nous pouvons citer un autre cooccurrent qui apparait, le lemme "classification". Les classifications de textes sont issues de graphes tels que les arbres qui utilisent des méthodes statistiques pour nous donner de précieuses informations quant à la constitution du corpus.

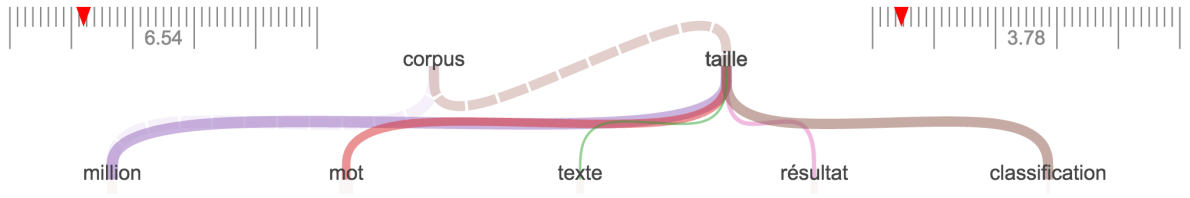

FIGURE 7 – Polycooccurrence du mot "taille" dans l'environnement du mot pôle "corpus" - période 2008 à 2014.

Cette analyse contrastive illustre l'évolution des pratiques de l'ADT. Les méthodes sont sensibles aux moyens technologiques dont elles disposent. La numérisation des données était encore il y a 16 ans un challenge en soi. Aujourd'hui, les "humanités numériques" nous laissent espérer des corpus sans limites de taille. Les outils associés se doivent de prendre en compte ce nouveau paradigme. Les visualisations graphiques apparues ces dernières années nous donnent maintenant les moyens d'explorer en détail ces corpus. Les choix ergonomiques qui s'offrent à nous sont nombreux et il ne serait pas étonnant de voir apparaître de nouvelles problématiques dans les prochaines années liées à ces choix.

5. Conclusion

Entre mise à jour du calcul et refonte de la visualisation concurrentielle, Hyperbase Web nous propose aujourd'hui une nouvelle fonction "Thème". Cette fonction du logiciel permet aux utilisateurs d'analyser leur corpus en ligne du point de vue de la polycooccurrence. L'analyse du co(n)texte d'un mot est soutenue par des visualisations graphiques dynamiques directement exécutées dans le navigateur de l'utilisateur. Hyperbase suit donc l'évolution des moyens et des méthodes technologiques et propose un autre point de vue sur la cooccurrence.

Le jeu de paramètres fourni avec l'outil permet de moduler à la demande les données en fonction des besoins de l'utilisateur. Le filtrage par codes morphosyntaxiques ainsi que la lemmatisation

autorisent des hypothèses très fines sur la cooccurrence, sans pour autant s’éloigner du calcul originel qui a fait le succès de cette fonction. Le réseau cooccurrentiel obtenu par la polycooccurrence a fait l’objet d’une étude approfondie de la représentation graphique. L’utilisateur dispose maintenant d’un jeu d’outils qui lui permet de parcourir différents chemins interprétatifs en quelques clics.

Hyperbase Web s’inscrit donc dans les nouvelles technologies dont dispose aujourd’hui l’ADT. La technique et les moyens ont évolué depuis le début des années 2000. Les corpus ont changé dans la nature et dans la taille, à l’image des Actes des JADTs dont on retrouve facilement la trace numérique à partir de l’édition 2000 sur le web. Cette base est d’ailleurs gratuite et en libre-accès, [7] ; une autre des caractéristiques observables de nos outils d’aujourd’hui.

Références

Bertels, A. et Geeraerts, D. (2012). L’importance du recoupement des cooccurrents de deuxième ordre pour étudier la corrélation entre la spécificité et la monosémie. In *Actes de JADT 2012*, pages 135–147.

Bostock, M., Ogievetsky, V., et Heer, J. (2011). D3 : Data-driven documents. *Visualization and Computer Graphics, IEEE Transactions on*, 17(12) :2301–2309.

Brunet, E. (2012). Au fond du goofre, un gisement de 44 milliards de mots. *JADT*.

Brunet, E. et Vanni, L. (2014). Goofre version 2. *JADT*.

Grefenstette, G. (1994). Corpus-derived first, second and third-order word affinities. In *Proceedings of EURALEX 1994*, pages 279–290.

Jean-Marc Leblanc, W. M. (2006). L’analyse contrastive des réseaux de cooccurrence : Le monde dans les discours des présidents de la cinquième république. In *Actes de JADT 2006*, pages 603–615.

Lafon, P. (1980). Sur la variabilité de la fréquence des formes dans un corpus. *Mots*, 1(1) :127–165.

Lemaire, B. et Denhière, G. (2006). Effects of High-Order Co-occurrences on Word Semantic Similarities. *Current Psychology Letters - Behaviour, Brain and Cognition*, 18(1) :1.

Martinez, W. (2003). *Contribution à une méthodologie de l’analyse des cooccurrences lexicales multiples dans les corpus textuels, Thèse de Doctorat en Sciences du Langage*. Thèse de doctorat, Université de la Sorbonne nouvelle - Paris 3,.

Martinez, W. (2012). Au-delà de la cooccurrence binaire... poly-cooccurrences et trames de cooccurrence. *Corpus*, (11).

Vanni, L., Luong, X., et Mayaffre, D. (2014). Arbre et co-occurrences. nouvel outil logométrique sur le net. application au discours de françois hollande. *JADT*.

7. http ://hyperbase.unice.fr/jadt