Approche textométrique des variations du sens

Julien Longhi, André Salem

Approche textométrique des variations du sens

Julien Longhi[1], André Salem[2]

[1]Université de Cergy-Pontoise, France – julien.longhi@u-cergy.fr

[2]Université de la Sorbonne nouvelle, France – salem@msh-paris.fr

Abstract

The use of textometric methods relies on the hypotheses, firtly, that stable units exist (forms, lemmas or their graphical approximations) and, secondly, that occurrences of these forms can be retrieved from different parts of a corpus. Once automatic counting performed, more sophisticated textometric methods can be employed to focus on textual variations (repeated segments, collocations, etc.) that occur around the same unit but in different contexts found within the corpus. This approach leads to the identification of semantic variations with relation to the context of each occurrence as highlighted through automatic segmentation. We will illustrate this by using examples of repeated segments within the corpus that contain the N-gram /enemy / taken from a widely-studied chronological text series.

Résumé

Pour pouvoir mettre en œuvre les méthodes de la textométrie, il est indispensable de postuler, dans un premier temps, l'existence d'unités stables (formes, lemmes ou leurs approximations graphiques), dont on recensera ensuite les occurrences dans les différentes parties du corpus étudié. Une fois les dépouillements automatiques réalisés, il est cependant possible d'utiliser des méthodes textométriques plus élaborées pour accéder aux variations textuelles (segments, répétés, cooccurrences, etc.) qui peuvent se réaliser autour d'une même forme dans chacun des contextes particuliers du corpus. Cette démarche permet d'accéder au repérage de variations sémantiques qui se rapportent à chacune des occurrences des formes produites par la segmentation automatique. Nous illustrons notre démarche à l'aide d'exemples prélevés dans les parties d'une série textuelle chronologique largement étudiée, des segments répétés du corpus qui contiennent le N-gram /ennemi/.

Keywords:.unité textométrique, sémantique, variation du sens

1. Introduction

Notre étude s'inscrit dans une perspective de prise en compte des dynamiques du sens à l'œuvre dans les discours, qui tiendrait compte de la variation, de l'hétérogénéité, ou encore de l'articulation entre topologie textuelle et discursive, sens et profilage. Le sens se construit dans différents champs où il est susceptible de paraître, et s'analyse « par le contexte, sous forme d'indices de position liés aux modalités de sa mise en place dans le champ » (Cadiot et Visetti, 2011), la caractérisation sémantique se faisant alors sur la base de la composition et décomposition des profils disponibles. L'automatisation du dépouillement de vastes corpus de textes, à des fins textométriques, nécessite au contraire que le repérage des unités de décompte puisse être confié à des machines. Pour pouvoir mettre en œuvre les méthodes de la

textométrie, il est indispensable de postuler, dans un premier temps, l'existence d'unités stables (lexèmes, lemmes ou leurs approximations graphiques), dont on recensera ensuite les occurrences dans différentes parties du texte. Cette manière de faire permet d'étudier la répartition de chacune des unités dans un corpus ou encore de rapprocher les différents contextes qui contiennent chaque unité textométrique. Ces simplifications, incontournables dans le premier temps de l'analyse, nous éloignent de l'étude du sens de chacune des occurrences que l'on peut élaborer dans chaque contexte particulier. Cependant, une fois les premiers dépouillements automatiques réalisés, il est possible d'utiliser des méthodes textométriques plus élaborées pour accéder aux variations textuelles qui peuvent se réaliser autour d'une même forme dans le corpus (segments répétés, cooccurrences, etc.). C'est ce croisement de perspectives et ce va-et-vient entre approche empirique et théorisation sémantique, que nous souhaitons mettre à l'épreuve dans la présente étude.

2. Application au corpus *Duchesne*

Pour illustrer notre démarche, nous appliquons ces méthodes à l'étude de la ventilation, dans les différentes parties d'une série textuelle chronologique largement étudiée, des segments répétés du corpus qui contiennent le N-gram /*ennemi*/.

2.1. Rappels sur l'analyse de la série chronologique Duchesne

La série chronologique *Père Duchesne* a déjà fait l'objet de nombreuses analyses textométriques[1]. Nous avons montré, en particulier, que les typologies réalisées à partir d'une partition de ce corpus en huit périodes, correspondant chacune à un mois de parution, mettaient en évidence un renouvellement lexical fortement lié à l'évolution dans le temps. On peut vérifier, sur la figure 1, que les parties correspondant aux périodes successives de parution sont proches sur les facteurs issus de l'analyse du tableau (8 parties x 1420 formes dont la fréquence dépasse dix occurrences)[2].

La méthode des ***segments répétés*** permet de repérer toutes les occurrences de suite de formes graphiques qui apparaissent plusieurs fois dans un corpus de textes (Lafon et Salem, 1983 ; Salem, 1986). Pour la présente étude, nous avons constitué un ensemble d'unités textuelles qui contient outre les formes graphiques *ennemi* et *ennemis*, tous les segments répétés qui contiennent l'une ou l'autre de ces formes. On a projeté sur la figure 1, en qualité d'éléments supplémentaires, cet ensemble de segments. La position sur ce graphique des différents segments montre que ces unités ne sont pas employées de manière uniforme tout au long des périodes.

---

[1] Le corpus Père Duchesne est constitué par la réunion d'un ensemble de livraisons du journal *Le Père Duchesne* de Jacques-René Hébert, parues entre 1793 et 1794. Pour une description plus avancée de ce corpus, on consultera, par exemple (Salem, 1988).

Les analyses dont nous rendons compte ci-dessous, ont été effectuées à l'aide du logiciel Lexico5. Cedric Lamalle, William Martinez, Serge Fleury ont largement contribué au développement des fonctionnalités de ce logiciel. Les auteurs tiennent à les en remercier.

[2] Ce phénomène connu sous le nom *d'effet Guttman*, a été largement décrit par Guttman (1941, 1946, 1950), Benzecri (1973) et Van Rijckevorsel (1987).

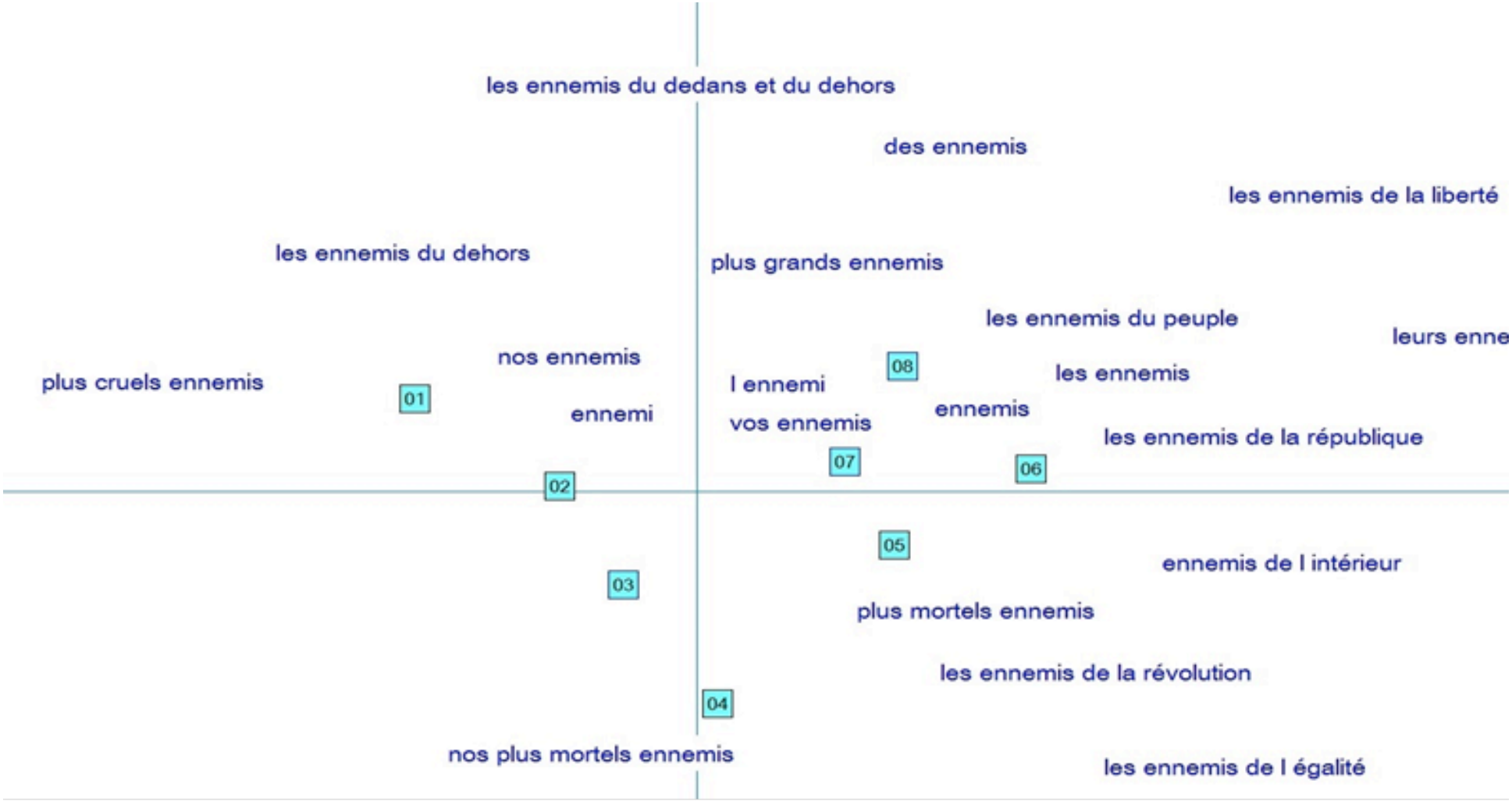

Figure 1 : Duchesne. Les segments contenant la séquence *ennemi* sur le plan des deux premiers facteurs issus de l'analyse de tableau 8 parties x 1420 formes (F>=10)

Guide de lecture pour la figure 1 : La figure fournit la représentation des huit parties du corpus ***Duchesne***, sur les deux premiers axes issus d'une Analyse des correspondances, réalisée sur l'ensemble des formes dont la fréquence dépasse 10 occurrences. Les segments répétés du corpus contenant la séquence de caractères /*ennemi* / ont été projetés sur ce même plan, en tant qu'éléments supplémentaires. La figure a été allégée des segments redondants (ex : segments contenus dans des segments plus longs). Certains des éléments superposés par l'analyse ont été très légèrement déplacés à fin de rendre la figure plus lisible.

Ainsi par exemple, le segment *plus cruels ennemis* trouve toutes ses occurrences au début du corpus alors que celles du segment *ennemis de la liberté* sont plutôt concentrées vers la fin.

L'analyse des projections des différents segments qui contiennent le n-gram /ennemi/ va nous permettre de dégager des contextes dont la distribution diffère fortement entre le début et la fin de la période temporelle couverte par le corpus.

2.2. L'évolution du contexte de la forme ennemi(s)

On peut estimer que le contenu sémantique de la forme *ennemi(s)* conserve une valeur relativement stable tout au long des périodes couvertes par le corpus que nous étudions. Le chercheur confronté à l'analyse de ces textes retrouvera sans peine, lors de l'examen de chacune des occurrences du terme, les principaux traits sémantiques décrits dans un dictionnaire de langue à propos de ce lexème (opposé, hostile, etc.). Cependant, l'analyse de ces mêmes contextes montre qu'il en va tout autrement pour ce qui concerne les référents auxquels la forme renvoie, dans chaque période particulière. Aux *plus cruels ennemis, plus mortels ennemis, ennemis du dehors* (les puissances étrangères, les expatriés), des périodes du début, succèdent bientôt *les ennemis du dedans et du dehors,* expressions qui peuvent s'analyser comme une dénonciation du fait que les *ennemis du dehors* ne constituent pas le seul danger et qui opère donc une modification manifeste du référent de départ. Par la suite la mention des *ennemis de l'intérieur* complètera la notion d'*ennemis du dedans*. Il faut noter que les *ennemis de l'intérieur* sont de plus en plus souvent précédés de l'article défini *les* qui les désigne comme une réalité dont l'existence est présupposée (elle n'est plus à démontrer).

Progressivement, *nos ennemis*, deviennent *vos ennemis*, puis *les ennemis*. Dans la dernière période les ennemis, désormais désignés, de manière préférentielle, au pluriel, ne sont plus qualifiés par leur localisation ou par leur rapport aux destinataires du message (*nos/vos ennemis*) mais par des valeurs supposée communes auxquelles ils sont censés s'opposer : *ennemis du peuple, ennemis de la république, ennemis de la révolution, ennemis de la liberté, ennemis de l'égalité.*

3. La sémantique de *ennemi(s)*

Les variations constatées montrent que la forme *ennemi(s)* prend différents sens selon les contextes dans lesquels elle s'inscrit, en ce qu'ils sont associés à des référents distincts. Plutôt que de représenter le sens comme la somme des cooccurrences constatées, nous souhaitons analyser ces valeurs comme un sous-ensemble prélevé sur un ensemble de valeurs acquises. Les espaces sémantiques déterminés et caractérisés par l'analyse statistique jouent un rôle fondamental qui, au-delà des synonymies, ou des polysémies, se renouvellent « en étant confronté aux textes – ce qui impliquerait de prêter attention à d'autres corrélations » (Visetti 2004 : 11). La description sémantique que nous proposons s'inscrit dans le champ de la sémantique lexicale[3], du côté des approches qui envisagent la construction des référents comme extrinsèque. Cependant, alors que ces approches mobilisent en général des analyses phrastiques, et travaillent sur des exemples forgés, nous introduisons une perspective statistique qui précède la représentation du sens. La description de l'objet *ennemi(s)* n'est pas séparée des rapports que l'on entretient avec lui, et sa description suppose une prise en

[3] Cadiot et Némo (1997 : 127-128)

compte différenciée de ses propriétés extrinsèques (relatives à ces rapports), et de ses propriétés intrinsèques (supposées stables et indépendantes).

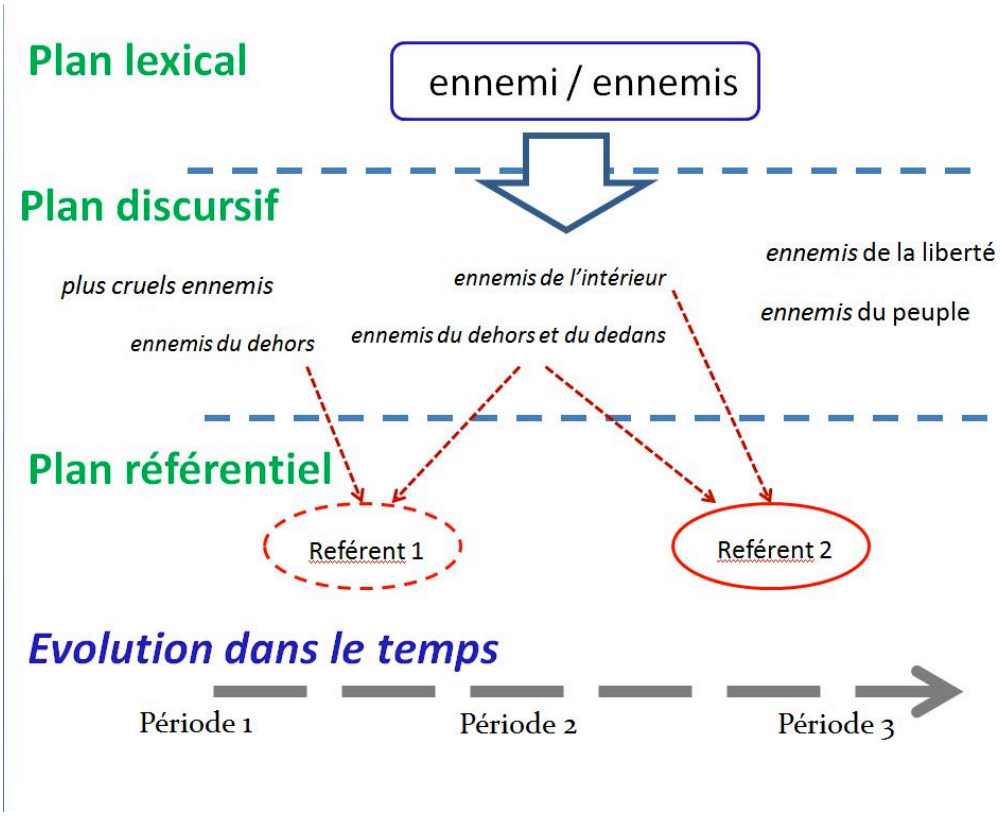

Figure 2 : Niveaux et unités d'analyse

L'intérêt de cette démonstration textométrique est pour nous de fournir des résultats concrets et matériels pour l'analyse des sens d'une unité lexicale. Ceci a plusieurs conséquences pour la mise en œuvre d'une sémantique soucieuse de l'exploitation des constats empiriques :

1) la représentation des variations du sens en contexte nous a permis d'identifier la manière donc les propriétés sont introduites et attribuées dans le corpus. Le référent change au fil du temps, puisque les *ennemis*, initialement définis comme *du dehors*, et introduits par *nos*, deviennent *vos ennemis*, et se présentent finalement sous la forme *ennemi(s) de* + N. Le besoin d'être déterminé par un complément du nom, ou son équivalent, qui indique avec quoi le terme « relatif » se trouve mis en relation », cette complémentation explicitant « ainsi la référence identitaire » (Steuckardt, 2008).

2) L'évolution dans le corpus au fil du temps permet de rendre compte de la dynamique sémantique à l'œuvre, laquelle rend compte diachroniquement des évolutions de sens. La textométrie permet ainsi de saisir les processus, et donc de donner du sens à la dimension potentiellement « hétéroclite » des propriétés des référents.

Ainsi, au plan linguistique, le passage du référent 1 ou référent 2 se fait par l'intermédiaire d'une transformation des propriétés de *ennemi(s)* : défini de manière situationnelle (*du dehors*) et relative (*nos*, *nos plus cruels*), il acquiert des propriétés plus polémiques (*vos*, *du dedans et du dehors*), pour s'intégrer ensuite dans un processus discursif qui construit le référent (ennemi de + N : *ennemi de la liberté* ; *ennemi du peuple*), par l'introduction de termes à fort charge axiologique. Le référent introduit alors un point de vue, qui n'est pas

strictement géographique ou institutionnel, mais aussi politique et idéologique. L'approche statistique dévoile, en outre, que c'est le pluriel qui est prioritairement mobilisé.

3. Conclusion

De manière désormais classique, les méthodes de la textométrie permettent de mettre en évidence les variations du vocabulaire qui surviennent au cours des périodes successives d'une même série textuelle chronologique. Dans la présente étude, nous avons appliqué les méthodes d'analyse statistique multidimensionnelle (AFC) à l'étude d'un ensemble particulier, celui des segments répétés réunis sur la base du fait qu'ils contenaient tous une même unité graphique (en l'occurrence, le n-gram /*ennemi/*).

La confrontation des segments ainsi sélectionnés nous permet d'observer des variations autour des formes graphiques *ennemi* et *ennemis*. L'analyse de ces variations dans le temps nous conduit à distinguer des référents qui varient en fonction des périodes réunies dans le corpus.

Au-delà des séries textuelles chronologiques, la méthode que nous avons présentée est susceptible de recevoir des applications dans l'étude de nombreux types de corpus. L'extraction semi-automatique des unités dont les contextes varient fortement en fonction des parties d'un corpus textuelle peut également être envisagée.

References

Benzécri J-P. and coll. (1981). *Pratique de l'analyse des données, Linguistique et lexicologie*. Dunod.

Cadiot P. and Nemo F. (1997). Propriétés extrinsèques en sémantique lexicale. *Journal of French Language Studies*, 7(2): 127-146.

Cadiot P. and Visetti Y.-M. (2001). *Pour une théorie des formes sémantiques*.PUF.

Guttman L. (1941). The quantification of a class of attributes: a theory and method of a scale construction. In P. Horst, *The prediction of personal adjustment*, SSCR New York.

Lafon P. and Salem A. (1983). L'Inventaire des segments répétés d'un texte. *Mots. Les langages du politique,* 6 : 161-177.

Lamalle C, Martinez W, Fleury S, and Salem A. (2002). *Les dix premiers pas avec Lexico3. Outils lexicométrique*s. http://www.cavi.univ-paris3.fr/Ilpga/ilpga/tal/lexicoWWW

Lebart L. and Salem A. (1994). *Statistique textuelle*. Dunod.

Longhi J. (2008). *Objets discursifs et doxa. Essai de sémantique discursive*. L'Harmattan, coll. « Sémantiques ».

Rastier F. (2011). *La mesure et le grain. Sémantique de corpus*. Honoré Champion, coll. « Lettres numériques ».

Salem A. (1987). *Pratique des segments répétés*. Klincksieck.

Salem A. (1988). Approches du temps lexical. *Mots. Les langages du politique*, 17 : 105-143.

Steuckardt A. (2008). Les ennemis selon *L'Ami du peuple*, ou la catégorisation identitaire par contraste. *Mots. Les langages du politique* [En ligne], 69 | 2002. http://journals.openedition.org/mots/10023

Van Rijckevorsel J. (1987). *The application of fuzzy coding and horseshoes in multiple correspondances analysis.* DSWO Press.

Visetti Y.-M. (2004). Le Continu en sémantique : une question de formes. *Texto ! juin 2004*. http://www.revue-texto.net/Inedits/Visetti/Visetti_Continu.html