Outreau en n sèmes, Outreau en cinq temps. Diachronie de la représentation sémique d'une unité lexicale

Coralie Reutenauer, Michelle Lecolle, Evelyne Jacquey, Mathieu Valette

Outreau en n sèmes, Outreau en cinq temps

Diachronie de la représentation sémique d'une unité lexicale

Coralie Reutenauer[1], Michelle Lecolle[2], Evelyne Jacquey[1] et Mathieu Valette[1]

[1] ATIFL, UMR 7118 (CNRS – Nancy Université),
{coralie.reutenauer, evelyne.jacquey, mathieu.valette}@atilf.fr
[2] CELTED, Université Paul Verlaine, Metz
lecolle@univ-metz.fr

Résumé : L'étude se situe dans un contexte de veille lexicale. Elle cherche à faire émerger de façon semi-automatique l'évolution de sens du nom propre *Outreau,* analysée manuellement dans une étude antérieure (Lecolle, 2007).

Mots-clés : Sémantique textuelle, Annotation de corpus, Evolution diachronique, Représentation sémique.

1 Introduction

Si un certain nombre de travaux se sont appuyés sur les textes pour étudier les termes implantés et stabilisés (Bourigault et Slodzian, 2000), peu d'entre eux se sont attachés aux termes émergents (lire néanmoins (Tartier, 2004)). Dans le cadre théorique de la sémantique interprétative (Rastier, 1996), on fera l'hypothèse que la naissance d'un terme en contexte s'accompagne de transformations de son environnement sémique, avec apparition et disparition progressives de traits sémantiques regroupés en classes sémantiques, jusqu'à stabilisation. Sous cette hypothèse, il semble vraisemblable que ces évolutions se manifestent à travers des phénomènes statistiques. Notre étude, située dans cette perspective, est centrée sur un mot-pôle dont l'évolution de sens diachronique a été identifiée et analysée manuellement (Lecolle, 2007). L'objectif est de mettre en évidence cette évolution et de la caractériser de façon semi-automatique en utilisant une représentation en traits sémantiques.

2 Outreau : du corpus au mot-pôle

2.1 Présentation du corpus

Le corpus porte sur l'affaire judiciaire d'Outreau. Il est constitué d'articles de presse de novembre 2001 à avril 2006, sélectionnés sur critère de présence du nom *Outreau*. Il a été initialement réalisé dans le cadre de l'étude linguistique de la polysignifiance du nom propre *Outreau* (Lecolle, 2007). Il est divisé en cinq périodes :

- 2001-2002 : "découverte" d'un réseau, arrestation de notables
- mai-juin 2004 : procès de Saint-Omer
- 1-2/07/2004 : attente du verdict de Saint-Omer
- 3-8/07/2004 : verdict du procès
- 2/12/2005 à avril 2006 : procès en appel à Paris ; suite et conséquences (commission d'enquête parlementaire)

Selon (Lecolle, 2007) le sens d'*Outreau* évolue au fil des périodes. De toponyme, il devient « l'erreur judiciaire par excellence ».

Dans l'étude actuelle, le corpus se présente sous deux versions parallèles : la version lexicale, de 400 000 occurrences de formes (issue de la version initialement réunie par (Lecolle, 2007)) ; une version « sémique » de 10 millions de ce que nous qualifions de « candidats-sèmes », par analogie aux *candidats-termes* de la terminologie, parce qu'ils sont le résultat d'un traitement automatique et n'ont pas encore été validés par le sémanticien. L'image sémique du corpus est obtenue à l'aide du logiciel Semy (Grzesitchak *et al.*, 2007) en substituant à chaque forme lexicale un sémème théorique issu des définitions lexicographiques du *TLFi* (Dendien *et al*., 2003). Sont considérés comme candidats-sèmes les lemmes des noms, verbes, adjectifs, adverbes[1] de ces définitions.

2.2 Caractérisation du mot-pôle *Outreau*

Dans le cadre de la sémantique interprétative, l'observation de l'évolution sémantique d'un mot ou syntagme est modélisée par celle de sa représentation sémique (ensemble de ses candidats-sèmes). *Outreau*, absent du TLFi en tant qu'entrée puisqu'il s'agit d'un nom propre, n'a qu' */Outreau/* comme candidat-sème affecté par le programme d'annotation. Deux autres méthodes sont alors utilisées pour générer sa représentation sémique. La première consiste à élaborer humainement une définition d'*Outreau* de type lexicographique à partir des connaissances provenant de l'étude manuelle, puis à convertir cette définition en candidats-sèmes comme l'aurait fait le programme d'annotation. D'après les connaissances tirées de son étude, Michelle Lecolle a établi la définition suivante :

Outreau :

1. Ville française du Pas-de-Calais

2. Erreur judiciaire liée à la découverte et croyance en l'existence d'un réseau pédophile puis à la réfutation publique de cette croyance.

1 Lire (Grzesitchak *et al.*, 2007) pour une présentation du programme d'annotation sémique et (Valette, 2008) pour une discussion sur la constitution d'une ressource sémique à partir d'un dictionnaire.

Cette définition donne l'ensemble de candidats-sèmes { /ville/, /français/, /pas-de-calais/, /erreur/, /judiciaire/, /découverte/, /croyance/, /existence/, /réseau/, /pédophile/, /réfutation/, /publique/[2] }.

La deuxième image sémique, dite 'de résonance', est générée par une procédure semi-automatisée qui sélectionne des candidats-sèmes faisant écho à des formes lexicales spécifiques du voisinage d'*Outreau* (voir 2.a).

3 Expériences

3.1 Outil mathématique : calcul des spécificités

Les expériences reposent sur le calcul des spécificités, implémenté ici par Lexico3 (Salem *et al.* 2003). Cette mesure, construite sur le modèle hypergéométrique, utilise des comparaisons entre partie et tout (le tout étant généralement l'ensemble du corpus, la partie, l'ensemble des contextes contenant le mot-pôle). Elle estime, à partir de calculs de probabilité, le degré de surreprésentation ou sous-représentation d'une forme donnée dans un sous-corpus selon sa fréquence, donc détermine dans quelle mesure cette forme caractérise le sous-corpus. La forme sera spécifique si sa fréquence est supérieure à celle attendue théoriquement d'après sa distribution dans l'ensemble du corpus. Pour plus de précisions, lire (Lafon, 1984).

3.2 Génération de la représentation sémique de résonance

Pour appliquer un calcul de spécificité, il est nécessaire de différencier une ou plusieurs partie(s) du corpus. L'expérience présente ayant pour but d'observer le comportement dans le temps du mot-pôle *Outreau*, la partie à laquelle est appliqué le calcul est l'ensemble des paragraphes contenant *Outreau* sur le plan lexical (respectivement, ces mêmes paragraphes sur le plan sémique). Le résultat du calcul de spécificité se présente sous forme de listes : liste de formes sur le plan lexical, liste de candidats-sèmes sur le plan sémique.

La représentation sémique de résonance est obtenue en confrontant ces deux listes, restreintes aux items de spécificité positive supérieure à 2. Cette confrontation a pour objectif de filtrer les candidats-sèmes : seuls sont conservés les candidats dont au moins une forme morphologiquement proche est dans la sous-liste lexicale. Par exemple, si la forme lexicale *débattre* et le candidat-sème /débat/ ont une spécificité supérieure à 2, /débat/ sera conservé. L'image sémique sera donc constituée de l'ensemble des sèmes ainsi sélectionnés.

3.3 Quantification de l'évolution par période des candidats-sèmes

[2] Les candidats /puis/, ininterprétable, et /lier/, provenant du métalangage lexicographique, sont exclus.

Afin de mesurer l'évolution diachronique de l'image sémique d'*Outreau*, nous cherchons à quantifier le degré de surreprésentation ou de sous-représentation de chaque candidat-sème à une période donnée. Ainsi, pour chaque période, le calcul des spécificités est appliqué aux candidats-sèmes sur le sous-corpus sémique des paragraphes de la période concernée contenant *Outreau*. Chaque candidat-sème se voit ainsi affecter un coefficient par période.

4 Analyse et validation des résultats

Les résultats mettent en évidence l'évolution des candidats-sèmes d'une période à l'autre ou encore leur positionnement respectif au sein d'une même période, avec émergence de candidats ou groupes de candidats statistiquement caractéristiques d'une période. Par exemple, le candidat-sème /ville/ voit sa spécificité décroître au fil du temps, tandis que /judiciaire/ ou /procès/, absents en période 1, s'imposent aux périodes suivantes. Pour analyser les résultats statistiques, nous avons mis en place une évaluation manuelle indépendante de la connaissance du processus de traitement automatisé. Nous aborderons d'abord les résultats de l'image sémique issue de la définition théorique, puis ceux de l'image sémique en résonance.

4.1 Image sémique issue de la définition théorique

Nous avons retenu trois axes d'observation : l'étude de la pertinence des candidats une fois la définition déstructurée ; l'estimation de l'activation par période de chaque candidat sans connaissance préalable des résultats numériques puis confrontation des listes établies manuellement et automatiquement ; la validation de l'allure, sous forme d'histogrammes, de l'évolution observée sur les cinq périodes à candidat fixé.

Concernant la pertinence des candidats-sèmes, l'analyse se heurte au caractère prédicatif de certains candidats, à savoir /découverte/, /existence/, /croyance/ et /réfutation/ : si ces candidats sont traités de façon isolée, l'analyse est ambiguë et délicate, voire impossible.

Pour évaluer l'activation des candidats-sèmes, des listes de données qualitatives sont constituées manuellement, où les candidats sont classés avec les valeurs "activé", "non-activé" ou "indécidable" pour chaque période, puis confrontées aux listes correspondantes de spécificités calculées automatiquement. Afin de mettre en parallèle les résultats, on considère que les valeurs de spécificités négatives ou faibles (inférieures à 2), correspondent à une non-activation du candidat-sème, et les spécificités supérieures à 2, à son activation. Hors cas ambigus mentionnés précédemment, on constate une convergence parfaite en période 1 et sur l'essentiel de la période 2. En revanche, le taux de convergence est médiocre aux périodes 3 à 5, mais, dans les cas tranchés, c'est-à-dire sur les spécificités les plus fortes en valeur absolue, les listes manuelle et automatique s'accordent.

Evaluation humaine

	période				
candidat-sème	1	2	3	4	5
ville	oui	non	non	non	non
français (adj)	oui	non	non	non	non
pas-de-calais	oui	oui	oui	oui	non
erreur	non	non	-	-	oui
judiciaire	non	oui	oui	oui	oui
découverte	oui	-	-	-	oui
croyance	-	-	-	-	oui
réfutation	non	-	-	-	oui
existence	oui	oui	oui	oui	oui
réseau	oui	oui	oui	oui	oui
pédophile	oui	oui	oui	oui	oui

Légende :
"oui" : activation du candidats-sème ;
"non" : non-activation du candidat ;
"-" : candidat inclassable

Résultats numériques (spécificités)

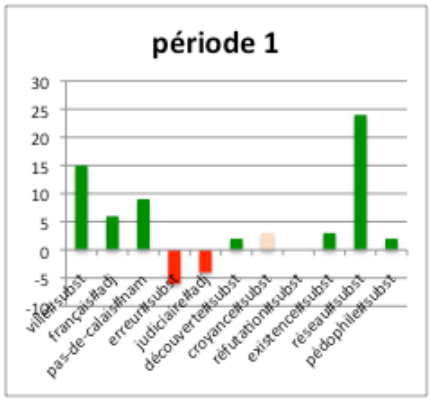

Fig. 1 – Confrontation des résultats manuels et numériques et période 1

Enfin, les histogrammes d'évolution par période des candidats ont été globalement jugés cohérents avec l'analyse manuelle, à l'exception de /pédophile/, dont l'évolution, non couplée à celle de /réseau/, est en désaccord avec la connaissance du corpus, et hors cas ambigus.

4.2 Image sémique de résonance

Deux méthodes d'analyse ont été utilisées pour l'image sémique de résonance.

La première confronte, comme précédemment, des listes manuelles et numériques de spécificités sur l'activation des candidats. Si une spécificité supérieure à 2 est considérée comme une activation et inférieure à -2 comme une non-activation, les deux types de résultats sont en adéquation dans 67% des cas (hors indécidables), avec convergence nette aux périodes 1 et 2, mais peu satisfaisante aux périodes suivantes. Cependant, en écartant les candidats de faible spécificité (entre -2 et 2), donc en ne conservant que les cas où l'activation ou non-activation est nette, le taux de convergence atteint 89% au total, et est supérieur à 80% pour chaque période.

taux de convergence	période					
	1	2	3	4	5	**total**
cas 1 : spécificités faibles assimilées à une non activation	83%	86%	56%	54%	56%	**67%**
cas 2 : spécificités faibles exclues	87%	93%	89%	94%	83%	**89%**

Fig. 2 – Proportion de candidats-sèmes pour lesquels données numériques et évaluations humaines s'accordent

L'information saillante humainement l'est donc également au niveau des coefficients.

La seconde approche s'appuie sur la génération de classes sémantiques à partir des connaissances du corpus et sans indication sur les résultats numériques. Elle consiste à confronter l'émergence par période des classes d'après les données numériques et une analyse manuelle indépendante. A titre d'exemple, la classe //ville_et_habitants//, considérée comme très saillante en période 1 et non saillante aux autres périodes, présente un profil de spécificités conforme aux attentes (figure 3).

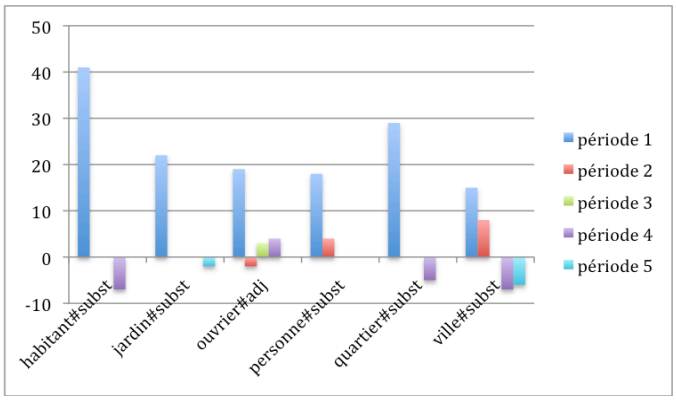

Fig. 3 – Profil de spécificités des candidats de la classe //ville et habitants//

5 Conclusion

La méthode présentée met en place une représentation sémique quantifiée du mot-pôle *Outreau*. Celle-ci permet d'observer une évolution diachronique de candidats-sèmes de façon isolée ou en groupements sémiques et leur émergence au sein d'une période. Les comportements sensibles à travers les données ont été validés par une analyse manuelle disjointe de la production des résultats, soit en amont avec prédiction des comportements, soit en aval au vu de résultats extraits selon l'axe d'observation. Si les résultats sont dans l'ensemble positifs, le traitement automatisé révèle toutefois ses limites au niveau de l'analyse des données, notamment pour les candidats-sèmes ininterprétables s'ils sont isolés. Ce constat invite à réfléchir à des méthodes d'extraction d'information exploitable ou d'articulation de candidats-sèmes ambigus.

Pour conclure, si on ne peut naturellement allouer le statut de terme à *Outreau*, le processus d'émergence de nouvelles facettes sémantiques supplantant le sens initial n'est pas sans évoquer des processus en jeu dans l'émergence ou évolution de termes. Le phénomène statistiquement sensible de disparition ou apparition de classes

sémantiques, assimilables aux taxèmes de la sémantique textuelle, ouvre des perspectives en termes d'automatisation, et donc de veille lexicale.

Références

BOURIGAULT D. & SLODZIAN M. (2000). Pour une terminologie textuelle. *Terminologies Nouvelles*. 19, p. 29-32.

CONDAMINES A., REYBEROLLE J. & SOUBEILLE A. (2004). Variation de la terminologie dans le Temps : une Méthode Linguistique pour Mesurer l'Evolution de la Connaissance en Corpus. *Actes Euralex International Congress*. p. 547-557. Université de Lorient.

DENDIEN J. & PIERREL J.-M. (2003). Le trésor de la langue française informatisé. Un exemple d'informatisation d'un dictionnaire de langue de référence. *TAL*. 44-2, p. 11-37.

LECOLLE M. (2007). Polysignifiance du toponyme, historicité du sens et interprétation en corpus. *Corpus*. 6, p. 101-125.

GRZESITCHAK M., JACQUEY E. & VALETTE M. (2007). Systèmes complexes et analyse textuelle : Traits sémantiques et recherche d'isotopies. *ARCo'07*. p. 227-235.

RASTIER F. (1996). Sémantique interprétative. PUF. Paris. Première édition : 1987.

SALEM A., LAMALLE C., MARTINEZ W., FRACCHIOLLA B., KUNCOVA A. & MAISONDIEU A. (2003). Lexico3 – Outil de statistique textuelle. Manuel d'utilisation. *Syled-CLA2T, Université de la Sorbonne Nouvelle*. p. 227-235. http://www.cavi.univ-paris3.fr/Ilpga/ilpga/tal/lexicoWWW.

TARTIER A. (2004). Analyse automatique de l'évolution terminologique : variations et distances. *Thèse de doctorat en informatique, Université de Nantes.*

VALETTE M. (2008). A quoi servent les lexiques sémantiques ? Discussion et proposition. *Cahiers du CENTAL. 5*, p. 43-58. P.U. de Louvain.