Sémème au macroscope : genèse et variation sémiques d'une unité lexicale

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

[1] ATILF – CNRS – Nancy Université – 54000 Nancy – France

[2] CELTED – Université Paul Verlaine-Metz – 57000 Metz – France

Résumé

Cette étude se situe dans le contexte de recherches en veille lexicale. L'enjeu est d'obtenir une représentation du sémème du nom *Outreau* à partir d'un corpus annoté en traits sémantiques. Cette représentation est construite de façon semi-automatique sur critère de renforcement sémique d'informations présentes sur le plan lexical, évalué par des indicateurs statistiques. Nous étudierons l'évolution diachronique de deux types de sémèmes : d'une part, des sémèmes ponctuels du mot *Outreau*, propres à différents stades de l'affaire judiciaire éponyme et à traits sémantiques variables, d'autre part un sémème global, représentatif de l'affaire dans son ensemble mais à structure variable dans le temps par phénomènes d'actualisations. La validation s'appuiera sur la confrontation des résultats obtenus par la procédure semi-automatique à une étude linguistique antérieure de l'évolution diachronique d'*Outreau*.

Abstract

The paper is about lexical tracking. In this study, we choose the general background of textual semantics and we aim at representing the word *Outreau* as a set of semes, extracted from a corpus through a semi-automatical process. Semes are selected only if they match and statistically strengthen lexical information. We will focus on the diachronic evolution of two representations in semes of *Outreau*, a time-related one submitted to qualitative evolution in time and a global one submitted to actualization processes in time. The validation will rely on the results of a previous linguistic study.

Keywords : text-based semantics, corpus tagging, seme representation, semimetry, lexicometry, diachrony, polysignificance

1. Introduction

Dans le sillage des grandes entreprises du XXème siècle telles que le *Trésor de la Langue Française* (désormais *TLF*), de nouveaux chantiers lexicographiques s'ouvrent aujourd'hui. Parallèlement aux modèles collaboratifs du type *wiktionary*, qui semblent prometteurs en dépit des nombreuses incertitudes qu'ils soulèvent sur le plan de la qualité (Jacquemin et al., 2008), il importe aux linguistes d'élaborer des méthodologies scientifiques et des outils de *veille lexicale* destinés à détecter et identifier les phénomènes néologiques attestés ou susceptibles de mener à de nouvelles attestations. La modélisation desdits phénomènes néologiques (notamment la néologie de sens), constitue un enjeu à la fois théorique et ingénierique de taille : quels modèles linguistiques sont susceptibles de rendre compte de l'évolution du sens des mots, et comment élaborer les instruments de veille ? Cet article entend apporter quelques éléments de réponses par l'analyse semi-automatique des significations d'une unité lexicale dans un corpus diachronique.

Nous y étudierons l'évolution du sens du nom propre *Outreau* qui, de nom de ville du Pas-de-Calais en 2002, a acquis au cours du temps un sens stabilisé de « parangon des scandales judiciaires ». Caractérisé par un signifié évolutif au cours du temps mais un signifiant fixe, notre objet d'étude se rapporte à de la néologie sémantique, ou *néosémie* (Rastier et Valette, 2009), dont la détection reste un problème non résolu (Sablayrolles, 2002). Nos propositions méthodologiques s'inspirent de la sémantique interprétative de (Rastier, 1987). Elles accordent une position centrale à la textualité, à la fois pour élaborer un signifié d'*Outreau* et, à l'issue des résultats, pour en vérifier la validité. La description du signifié utilise une représentation en *sèmes*, ici qualifiés, avant validation, de *candidats-sèmes* [1].

2. Problématique

Nous chercherons d'une part à détecter des phénomènes de récurrences et groupements de sèmes dans le cotexte d'un mot, pour en extraire une représentation du contenu sémantique de ce mot, d'autre part à rendre compte de l'évolution diachronique de la représentation sémantique obtenue. Notre démarche se situe à la croisée de deux études antérieures. La première, (Lecolle, 2007) fait l'analyse diachronique du sens du nom propre *Outreau* qui, de toponyme, devient le désignateur de «l'erreur judiciaire par excellence». L'étude rend compte de l'émergence ou de la disparition de *facettes sémantiques* au cours du temps. La seconde, (Reutenauer *et al.*, in press), évalue une procédure automatique d'annotation de corpus en traits sémantiques, assimilés à des sèmes. Il s'agit d'en extraire la représentation sémique d'une unité lexicale. Elle conclut à la convergence de l'information apportée par les formes lexicales et par lesdits traits sémantiques. Elle met également à jour des *molécules sémiques* [2] dans le voisinage sémique de l'unité lexicale étudiée (le voisinage sémique s'obtient à partir des paragraphes qui contiennent l'unité lexicale étudiée ; il désigne l'ensemble des traits sémantiques affectés par annotation aux paragraphes en question). L'étude montre également que le plan sémique rend explicites des contenus sémantiques sensibles mais non patents sur le plan lexical. Il faut néanmoins rappeler que les résultats d'analyse reposent sur un filtrage manuel de listes bruitées de traits sémantiques obtenus automatiquement.

L'objectif de la présente étude est d'abord d'obtenir un ensemble peu bruité de candidats-sèmes d'*Outreau*. Les candidats-sèmes ciblés sont ceux qui viennent renforcer l'information présente sur le plan lexical. Ils proviennent d'une procédure semi-automatique qui annote le corpus en traits sémantiques (Reutenauer *et al.*, in press), puis combine analyse lexicométrique et analyse sémique (plus précisément *sémimétrique*) du corpus pour extraire un sémème. Il s'agit ensuite de vérifier si le comportement diachronique des candidats extraits est conforme à l'évolution diachronique des sens d'*Outreau* observée manuellement (Lecolle, 2007).

3. Focus sur *Outreau :* du corpus au mot-pôle

3.1. Présentation du corpus

Le corpus porte sur l'affaire judiciaire d'Outreau. Il est constitué d'articles de presse parus entre novembre 2001 et avril 2006, sélectionnés sur critère de présence du nom *Outreau*. Il a été initialement constitué dans le cadre de l'étude linguistique de la polysignifiance du nom propre

[1] Dénomination choisie par analogie aux candidats-termes de la terminologie.

[2] Groupement stable de sèmes, non nécessairement lexicalisé, ou dont la lexicalisation peut varier (Rastier, 1987 : 275).

Outreau (Lecolle, 2007). Ci-dessous sont résumées des conclusions tirées de (Lecolle, 2007, 2009) et réutilisées ici pour la validation des résultats d'expérience.

Selon (Lecolle, 2007, 2009), l'évolution diachronique du sens d'*Outreau* peut s'observer à travers un découpage en cinq périodes-clés, correspondant à des temps forts dans la succession des événements concernant l'affaire d'Outreau :

1. 2001-2002 : découverte d'un réseau pédophile à Outreau, arrestations
2. mai-juin 2004 : procès de Saint-Omer
3. 1-2/07/2004 : attente du verdict de Saint-Omer
4. 3-8/07/2004 : verdict du procès
5. 2/12/2005 à avril 2006 : procès en appel à Paris ; suite et conséquences (commission d'enquête parlementaire).

(Lecolle, 2007) dégage ainsi plusieurs dimensions sémantiques qui apparaissent au cours des périodes d'observation, ou sont au contraire éliminées, atténuées ou modifiées. Cinq catégories principales recouvrent les évolutions de sens.

- La *dimension locative* caractérise le sens d'*Outreau* en période 1. Elle est apparente à travers diverses facettes : emplacement géographique, structure urbaine, ou collectif propre au lieu (habitants). Même si le sens locatif d'*Outreau* ne disparaît jamais complètement, d'autres sens le supplantent progressivement.
- La *dimension policière et judiciaire* est présente aux cinq périodes, mais sous différentes formes. La période 1, moins marquée, se positionne en amont de la procédure pénale, elle recouvre des aspects policiers (arrestations) et l'enclenchement de la procédure (mise en examen). Les notions de réseau pédophile et d'inceste y sont très présentes, ainsi qu'en période 2. Le traitement judiciaire s'étale des périodes 2 à 4, avec le déroulement du procès de Saint-Omer en période 2, l'attente du verdict en période 3 et le verdict en période 4. La période 5, qui correspond au procès en appel et à la commission d'enquête parlementaire consécutive, porte à nouveau sur la procédure judiciaire mais aussi sur l'institution judiciaire comme objet d'étude en raison de ses dysfonctionnements, ce qui ajoute une dimension politique.
- *L'émotion populaire* est aussi sous-jacente dans le sens affecté à *Outreau*. Son influence dans le déroulement de l'affaire est dénoncée en période 5, on peut donc supposer qu'elle influe implicitement sur le sens d'*Outreau* aux périodes précédentes, à travers une condamnation.
- Les sens de « fiasco judiciaire » et d'« erreur judiciaire par excellence » sont caractéristiques de la période 5, même si l'erreur judiciaire émerge déjà aux périodes 2, 3 et 4.
- Le *dimension politique*, absent initialement, est surtout présent en période 5 avec l'ouverture d'une enquête parlementaire, mais apparaît avec la prise de recul sur les dysfonctionnements du système judiciaire.

3.2. Image sémique du corpus

A partir de la version lexicale du corpus, issue de la version initialement réunie par (Lecolle, 2007), est générée une deuxième version, sémique [3]. Cette version est obtenue à l'issue d'un traitement automatique décrit par (Grzesitchak et al., 2007) qui comprend l'étiquetage, la lemmatisation et l'élimination des mots grammaticaux du corpus, puis la substitution d'un *sémème* à tout lemme. Les sémèmes sont produits par extraction des substantifs, verbes, adjectifs

[3] Lire Valette, 2008 pour une discussion sur la constitution d'une ressource sémique pour l'annotation de corpus.

et adverbes sous forme lemmatisée issus des définitions lexicographiques correspondant à chaque entrée, depuis le *Trésor de la Langue Française Informatisé* (*TLFI,* Dendien et Pierrel, 2003). Chacun des lemmes extraits est considéré comme un trait sémantique. On ajoute à chaque sémème le mot vedette lui-même. Ce sémème est ainsi constitué d'un ensemble de ce que nous qualifierons, en l'absence de validation par l'expert sémanticien, de *candidats-sèmes*, par analogie aux candidats-termes de la terminologie. L'image sémique du corpus est 24 fois plus volumineuse que le corpus lui-même (un million d'occurrences de candidats-sèmes pour 400 000 formes), mais le vocabulaire reste à peu près identique (environ 24 000 unités).

3.3. Critères de construction du sémème d'Outreau

Après l'annotation initiale, *Outreau* a pour seul candidat-sème */Outreau/,* car en tant que nom propre, il n'a pas d'entrée dans le *TLFi*. Nous cherchons donc à enrichir de façon semi-automatique ce sémème insuffisant*,* de façon à refléter les différentes facettes sémantiques repérées manuellement par (Lecolle, 2007).

Nous cherchons dans le voisinage lexical ou sémique d'*Outreau* à sélectionner les candidats-sèmes susceptibles de le définir. Par « voisinage lexical », nous entendons les cooccurrents lexicaux d'*Outreau* au sein d'un même paragraphe, avant annotation. Le voisinage sémique s'obtient à partir des paragraphes du voisinage lexical, par annotation de ceux-ci en candidats-sèmes. Le mode de sélection des candidats, que nous souhaitons automatisable et ainsi reproductible, s'appuie sur des critères de significativité statistique.

Dans cet article, nous nous intéressons aux seuls candidats-sèmes correspondant à un lemme sur le plan lexical. Par exemple, on ne prendra en compte le candidat-sème /justice/ au plan sémique que si le lemme *justice* est actualisé au plan lexical. De tels candidats sont vecteurs d'information déjà présente et sensible sur le plan lexical, mais encore plus manifeste sur le plan sémique ; on qualifiera donc ce phénomène d' « écho renforcé ». (cf 4.3). Cette restriction constitue un filtrage très sélectif de candidats-sèmes obtenus par annotation, et correspond à un angle d'observation précis, guidé par une volonté de limitation de bruit. Il ne s'agit donc pas d'exploiter tout l'apport de l'annotation ; en particulier, sont exclus de l'étude les candidats-sèmes apparemment pertinents mais vecteurs d'information totalement absente sur le plan lexical. Ceux-ci feront l'objet d'autres études.

Pour observer l'évolution diachronique d'*Outreau*, différents ensembles de candidats-sèmes en 'écho renforcé' seront générés, de façon à répondre à deux types de caractérisations :

- une caractérisation par période, avec des sémèmes propres à la période concernée, et qui pourront être différents d'une période à l'autre,
- une caractérisation globale, comportant un sémème fixe sur les cinq périodes, mais de structuration interne variable d'une période à l'autre.

4. Procédure de sélection des candidats-sèmes

4.1. Vue d'ensemble

Les deux caractérisations recherchées ont donné lieu à deux séries d'expérience, dont les grandes lignes sont schématisées dans la Fig. 1 et détaillées dans les paragraphes suivants.

On abordera en 4.2 la constitution de sous-corpus et l'évaluation de la significativité ; en 4.3 la confrontation des listes lexicale et sémique ; en 5, partie analytique, les étapes ultérieures.

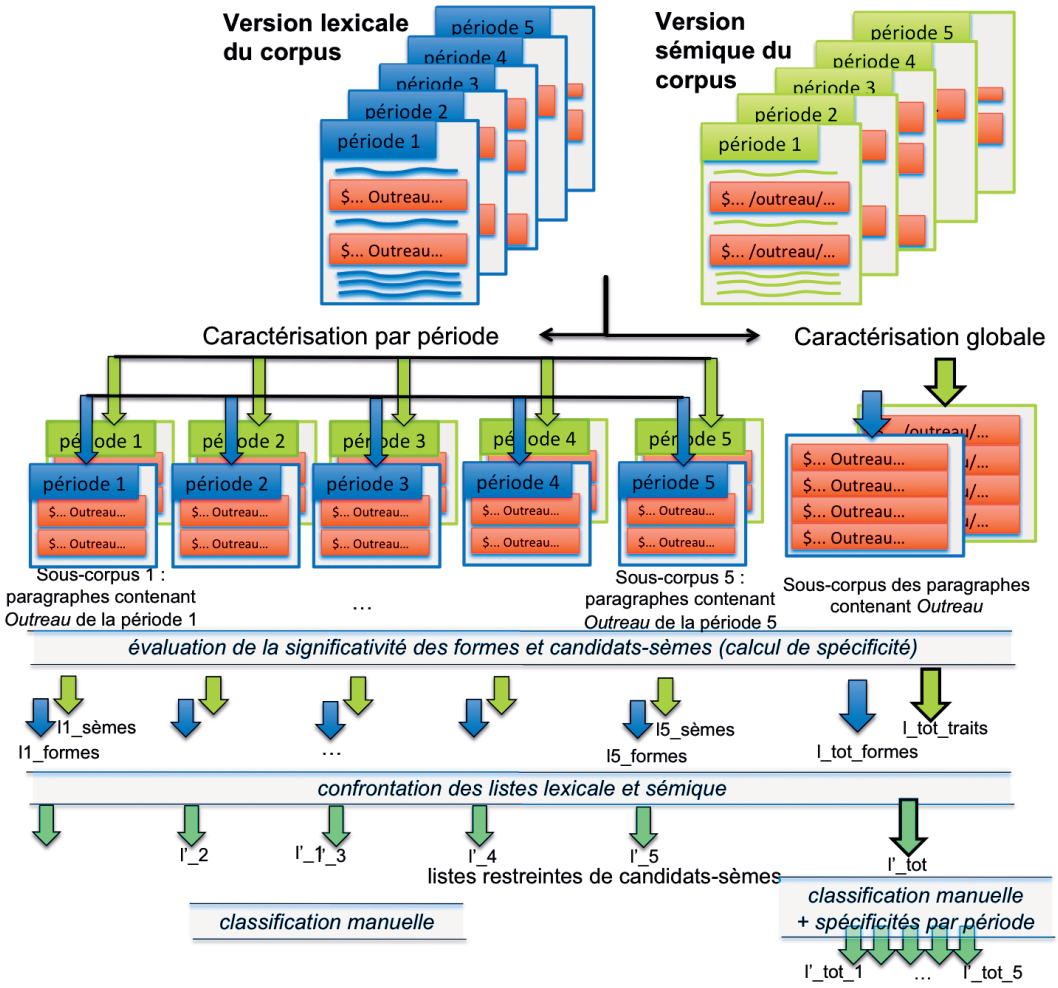


Figure 1 : Schéma d'ensemble de la procédure

4.2. Evaluation de la significativité : calcul des spécificités

La recherche et l'étude de candidats-sèmes reposent sur une évaluation statistique du degré de surreprésentation ou de sous-représentation des cooccurrents d'*Outreau*. La mesure utilisée ici est le calcul des spécificités, tel que défini par (Lafon, 1984) et implémenté par le logiciel Lexico3 (Salem et *al.*, 2003). La valeur de spécificité affectée à une unité provient de la probabilité d'observer *k* occurrences de cette unité dans un sous-corpus à préciser au préalable. Cette valeur est entière, positive si l'unité est surreprésentée dans le sous-corpus, négative sinon. Le calcul est effectué pour des seuils minimaux d'occurrences et de spécificité précisés au préalable. Dans notre cas, le seuil d'occurrences sera fixé à 10 et le seuil de spécificité sera de 3 sur le plan sémique pour la caractérisation par période et de 5 pour la caractérisation globale.

Pour la caractérisation globale, le calcul de spécificité est d'abord appliqué à deux sous-corpus parallèles : le sous-corpus de l'ensemble des paragraphes contenant *Outreau* sur le plan lexical et le sous-corpus équivalent sur le plan sémique. Ce calcul retourne une liste lexicale et une liste sémique d'unités affectées de leur spécificité, pour toute unité respectant les seuils fixés. Une fois la liste sémique filtrée par confrontation à la liste lexicale (voir paragraphe suivant), le calcul des spécificités est à nouveau appliqué aux candidats-sèmes retenus sur les sous-corpus obtenus par intersection du sous-corpus initial avec chaque période.

Pour la caractérisation par période, le calcul des spécificités est appliqué en parallèle au plan lexical et sémique à cinq sous-corpus, correspondant à l'ensemble des paragraphes contenant *Outreau* pour chacune des cinq périodes. Il en résulte une liste de spécificités sémique et lexicale pour chaque sous-corpus.

4.3. Sélection de candidats-sèmes renforcés par confrontation des listes lexicales et sémiques

Seuls sont retenus les candidats-sèmes correspondant à un « écho renforcé ». Plus précisément, pour un candidat-sème donné, nous choisissons comme référence sur le plan lexical les formes morphologiquement proches de son signifiant, c'est-à-dire les formes auxquelles il est associé de façon immédiate, mais pour lesquelles il apparaît comme le plus générique (par exemple, /magistrat/ pour *magistrats* ou *magistrature*).

Le candidat-sème sera retenu comme sème pertinent s'il respecte deux critères :

1. un critère de coactualisation qui implique :
   a. l'existence d'une forme de référence parmi les unités de spécificité positive : un des représentants lexicaux « immédiats » du candidat, c'est-à-dire proche morphologiquement, est de spécificité positive dans le sous-corpus considéré (seuil de spécificité fixé à 2)
   b. l'existence du candidat-sème lui-même comme unité spécifique positivement : le candidat-sème est lui aussi surreprésenté dans le sous-corpus considéré (seuil de spécificité fixé à 5 ou 3 selon l'approche)
2. un critère de renforcement : le candidat-sème a une spécificité strictement supérieure à celle de la forme de référence de plus grande spécificité.

Le premier critère revient à sélectionner les traits qui se « manifestent » de façon significative en tant que formes.

Le critère de renforcement exploite le nombre d'occurrences des unités lexicales à l'origine d'un candidat-sème dans la procédure d'annotation. L'accroissement de spécificité entre forme de référence et candidat est dû à une surreprésentation d'unités lexicales activant le candidat autres que la forme de référence la plus spécifique. Ces autres formes peuvent appartenir à la même famille morphologique que la forme de référence, auquel cas l'accroissement sera lié à un regroupement morphologique implicite, mais aussi, cas plus intéressant, elles peuvent provenir de formes lexicales très différentes, sans lien morphologique avec le candidat-sème.

Ces critères permettent d'obtenir des listes restreintes de candidats, de moins de cent éléments pour chacun des sous-corpus traités.

5. Analyse des listes de candidats-sèmes

Les démarches précédentes ont permis d'obtenir d'une part un ensemble de cinq listes de candidats-sèmes renforcés spécifiques du voisinage *d'Outreau* sur chacune des périodes ; d'autre part une liste globale de candidats-sèmes renforcés toutes périodes confondues. Dans les deux approches, les candidats seront observés à travers une classification dont la validation repose sur l'analyse manuelle de (Lecolle, 2007) (*cf.* la synthèse Fig. 2).

5.1. Liste de candidats-sèmes propre à chaque période

5.1.1. Classification des candidats

A partir des listes par période de candidats-sèmes, des classes sont définies manuellement. Ces classes correspondent à des regroupements sémantiques réalisés à des fins d'observation sur la base

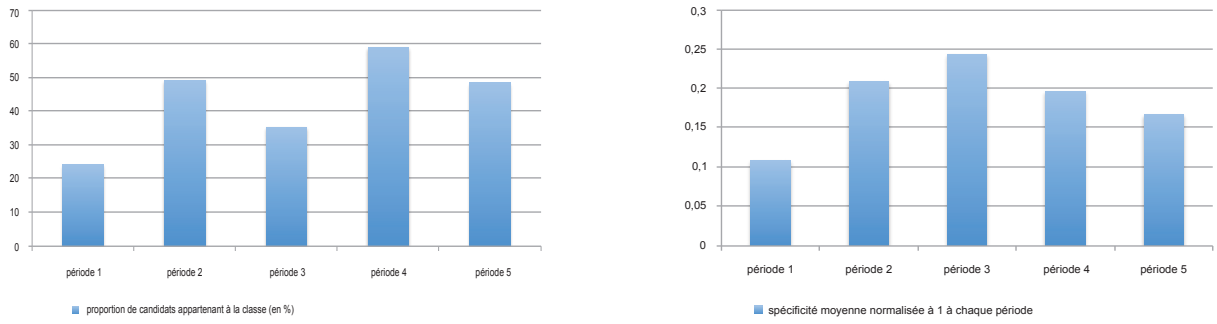
d'intuitions sémantiques. Par exemple, en période 1, émerge une classe JUDICIAIRE qui comporte notamment les candidats /police/, /procureur/, /écrouer/. Ces classes forment une partition sur les candidats retenus pour une période donnée et ne sont pas nécessairement communes à toutes les périodes. Pour chaque période, l'importance des classes est représentée par deux indicateurs : (i) l'un prend en compte la taille de la classe, par calcul de la proportion de candidats de la liste appartenant à la classe considérée ; (ii) l'autre est destiné à refléter la significativité des candidats affectés à la classe, par calcul de la moyenne des spécificités des candidats de la classe, puis, pour permettre une comparaison entre différentes périodes, homogénéisation des tailles des vecteurs-périodes des moyennes de spécificités (normalisation du vecteur à 1 en norme 1). Les conclusions avancées s'appuient sur des tendances communes aux deux indicateurs.

Période	*Dimension locative*	*Dimension judiciaire et policière (inclus réseau, pédophilie et inceste)*	*Dimension politique*	*Emotion populaire*	*Erreur judiciaire*
1	Présente [géographie, structure urbaine, habitants]	Présente [amont de la procédure pénale (arrestations et mise en examen)]	Absente	Sous-jacente [horreur]	Absente
2	Secondaire	Présente [traitement judiciaire (déroulement du procès)]	Absente	Sous-jacente [horreur]	Sous-jacente
3	Secondaire	Présente [traitement judiciaire (l'attente du verdict)]	Absente	Sous-jacente [horreur]	Sous-jacente
4	Secondaire	Présente [traitement judiciaire (le verdict)]	Emergente [dysfonctionnement du système judiciaire]	Sous-jacente [horreur]	Sous-jacente
5	Secondaire	Présente [Procès en appel et dysfonctionnements de l'institution judiciaire en tant qu'objet]	Présente [Enquête parlementaire]	Présente [Scandale, crainte vis-à-vis de l'institution judiciaire]	Présente [fiasco, erreur par excellence]

Figure 2 : Synthèse d'éléments d'analyse extraits de l'étude manuelle du corpus d'Outreau

5.1.2. Analyse des classes

Le judiciaire est présent à chaque période (cf. Fig. 3.a et 3.b). La proportion de candidats-sèmes de même que la spécificité moyenne sont plus faibles en période 1, période à laquelle le procès n'est pas encore entamé.



Figures 3.a et 3.b : évolution par période de la classe JUDICIAIRE, d'après (a) la spécificité moyenne des candidats-sèmes de la classe et (b) la proportion de candidats-sèmes de la période appartenant à la classe

Au niveau qualitatif, les candidats-sèmes de cette classe renvoient aux notions émergeant à chaque période. Par exemple, en période 1, l'ensemble de candidats-sèmes {/écrouer/, /police/, /arrestation/, /incarcération/, /incarcérer/, /prévenu/}, correspondant à près de la moitié des éléments associés à la sphère judiciaire, renvoient à l'idée d'arrestation.

Le POLITIQUE apparaît à la période 4 et se renforce à la période 5, ce qui est en accord avec la mise en place d'une commission d'enquête parlementaire et la volonté politique de se pencher sur les dysfonctionnements du système judiciaire.

Inversement, la classe du crime (évoquée par des candidats tels que /pédophilie/, /meurtre/, /viol/, /délinquant/), incluse dans la dimension judiciaire et policière, est présente en périodes 1 à 3, puis cette présence chute en période 4 et enfin disparaît en période 5 (Fig. 4.a et 4.b). Cette évolution s'explique doublement : d'une part, par l'ampleur accordée à l'émotionnel populaire dans un premier temps, puis le recul de son influence avec la prise de conscience de l'erreur judiciaire ; d'autre part, par le glissement du scandale de l'affaire de pédophilie vers l'erreur judiciaire.

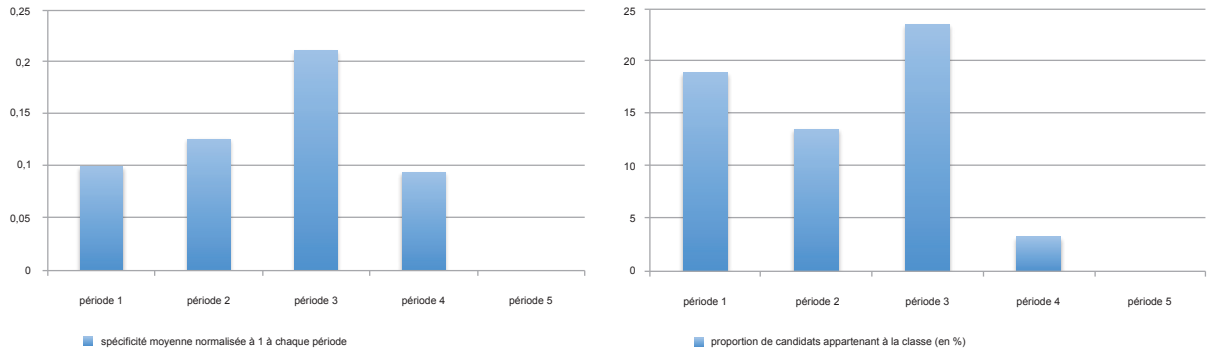


Figure 4.a et 4.b : évolution par période de la classe liée au crime en (a) spécificité moyenne et (b) proportion de candidats-sèmes affectés à la classe

Les classes rattachées au LIEU (LIEU D'HABITATION, LIEU GÉOGRAPHIQUE) ne sont représentées significativement qu'à la période 1, ce qui rejoint l'évolution du sens locatif d'*Outreau*, initialement dominant voire exclusif, puis supplanté par d'autres sens. Inversement, la classe du FIASCO (constituée de /naufrage/, /drame/, /faillite/, /faute/) s'impose comme représentative de la période 5.

5.2. Liste de candidats-sèmes globaux

Alors que l'approche précédente portait sur des candidats-sèmes supposés être caractéristiques du sens d'*Outreau* pour la période considérée, nous abordons dans ce paragraphe des candidats globalement caractéristiques d'*Outreau*, mais dont la pertinence, à une période donnée, n'est pas nécessairement avérée. Nous nous situons ici dans une perspective d'*activation* ou d'*inhibition* de candidats par période.

La confrontation de listes globales lexicale et sémique de spécificités fournit une liste de candidats-sèmes non pondérés, représentatifs de l'ensemble des paragraphes contenant *Outreau*. Afin de mesurer l'évolution diachronique de cette image sémique non pondérée, 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 retenus sur le sous-corpus sémique des paragraphes contenant *Outreau* de la période concernée. On obtient ainsi une représentation numérique du sémème sous forme d'un tableau de spécificité des candidats-sèmes par période.

Deux méthodes d'analyse sont utilisées.

La **première** étudie l'activation des candidats-sèmes considérés séparément. Pour évaluer cette activation, des listes de données qualifiées sont constituées manuellement, dans lesquelles les candidats sont classés selon les valeurs «activé», «non-activé» ou «indécidable» pour chaque période. Sont indécidables en particulier les candidats-sèmes qui ne peuvent être traités isolément, comme /commettre/ (*commettre une erreur ou commettre un crime*?) [4].

Afin de mettre en parallèle les listes numériques avec les listes manuelles, on considère que les valeurs de spécificités négatives ou faibles (strictement inférieures à 2), correspondent à une non-activation du candidat-sème, et les spécificités supérieures à 2, à son activation. Les deux types de résultats sont alors en adéquation dans 67% des cas (hors indécidables), avec convergence nette aux périodes 1 et 2, mais peu satisfaisante aux périodes suivantes. Cependant, en assimilant les candidats de faible spécificité (entre -2 et 2) à des indécidables, donc en ne conservant que les cas tranchés d'activation ou non activation, le taux de convergence atteint 91% au total, et est supérieur à 85% pour chaque période (Fig. 5).

L'information saillante pour un regard humain l'est donc également au niveau des coefficients.

La **seconde méthode d'analyse** s'appuie sur la constitution de classes à partir des connaissances du corpus et sans indication sur les résultats numériques. L'évolution des classes d'après les données numériques est ensuite confrontée à l'analyse manuelle.

L'histogramme de la Fig. 6 présente les moyennes de spécificité sur les candidats-sèmes des classes constituées par période. Nous n'aborderons que 5 des 7 classes constituées, pour faciliter la lecture et permettre une mise en parallèle aisée avec le tableau récapitulatif de la figure 2.

taux de convergence	*période*					
	1	*2*	*3*	*4*	*5*	*total*
cas 1 : spécificités faibles assimilées à une non activation	79%	89%	97%	54%	57%	67%
cas 2 : spécificités faibles exclues	86%	96%	88%	100%	85%	91%

Figure 5 : Proportion de candidats-sèmes pour lesquels données numériques et évaluations humaines s'accordent

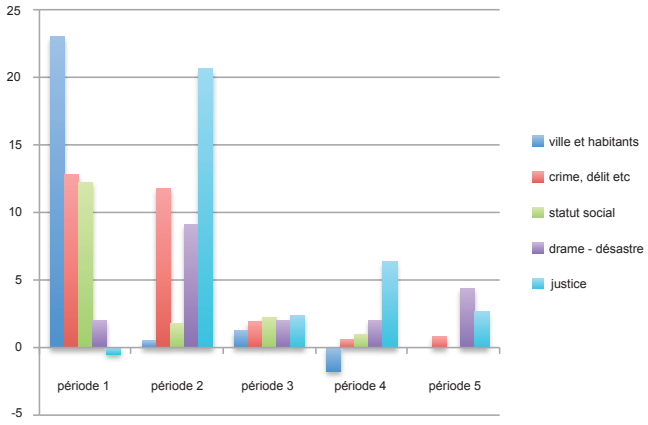


Figure 6 : Evolution par période de la moyenne de spécificité des candidats-sèmes de cinq classes (regroupements sémantiques constitutés manuellement à partir des listes de candidats)

[4] De fait, ce que nous avions considéré, dans une première approche, comme des candidats-sèmes s'avère difficile, voire impossible à traiter comme tel.

Cette représentation permet d'observer l'agencement interne des classes au sein d'une période. On constate en particulier une prépondérance du LIEU en période 1 alors que le JUDICIAIRE est inhibé ; inversement, en période 5, la classe LIEU est inexistante alors que les classes DÉSASTRE et JUDICIAIRE dominent. Un autre axe d'approche consiste à observer l'évolution d'une classe donnée d'une période à l'autre, malgré un biais introduit sur les spécificités par la taille des sous-corpus. Par exemple, le déclin rapide du LIEU dès la deuxième période apparaît nettement, alors que le JUDICIAIRE s'impose en période 2 et reste très représentée par la suite. Pour une approche plus fine des résultats, on peut observer le comportement des candidats-sèmes constitutifs de la classe ciblée au cours des périodes. Ainsi, pour la classe JUDICIAIRE, le comportement du sous-ensemble de candidats-sèmes {/écrouer/ ; /emprisonner/}, évolue à l'opposé des autres candidats de la classe, avec une spécificité fortement positive en période 1 et faible aux autres périodes. En effet, ce sous-ensemble fait écho à des aspects de la phase amont, policière, de la procédure pénale, et renvoie aux arrestations au début de l'affaire d'Outreau.

Les résultats observés convergent donc avec l'analyse manuelle. L'interprétation nécessite cependant d'autres clés, comme en témoigne l'évolution des ensembles de candidats qui évoquent le drame ou le désastre. Les spécificités les font apparaître comme saillants en périodes 2 et 5, mais ces indicateurs ne suffisent pas pour déterminer si les aspects dramatiques sont liés à l'affaire de pédophilie ou au désastre judiciaire.

5.3. Confrontation des deux caractérisations

Dans les deux approches, des classes ont été générées manuellement à partir des listes de candidats-sèmes obtenues par procédure semi-automatique. Ces classes, réalisées indépendamment par deux linguistes différents, présentent un certain nombre de similarités :

- les classes JUDICIAIRE et CRIME sont communes ; le LIEU également, bien que le nombre d'éléments de cette classe soit restreint dans l'approche globale (4 candidats-sèmes) ;
- les deux approches font l'une et l'autre apparaître le caractère désastreux de l'affaire d'Outreau, mais pas de façon identique. Dans l'approche par période, la distinction entre le désastre lié au fiasco judiciaire et le drame de l'affaire pédophile émerge nettement, alors que dans l'approche globale, cette distinction n'est pas présente ;
- la classe POLITIQUE n'apparaît que dans la caractérisation par période ; elle n'émerge que tardivement à l'échelle du corpus, et son poids dans l'ensemble du corpus n'est probablement pas suffisant pour la faire ressortir dans la caractérisation globale.

Au niveau de l'évolution des classes similaires, les tendances évolutives sont globalement les mêmes et en conformité avec les résultats tirés de l'analyse manuelle : CRIME en déclin sur les dernière périodes ; JUDICIAIRE présent en périodes 2 à 5 et dans une moindre mesure en période 1 ; LIEU caractéristique de la période 1 ; émergence d'une classe DRAME / DÉSASTRE en périodes 2 et 5, mais correspondant à des phénomènes différents (Fig. 7).

Enfin, on peut se demander dans quelle mesure les éléments constitutifs des classes similaires sont communs aux deux approches. Pour cela, nous avons étudié la proportion d'éléments communs aux classes similaires.

On constate que l'enrichissement qualitatif propre à chaque période (c'est-à-dire la proportion de candidats de la période absents dans la liste globale) est net, et plus important que dans la liste globale. Cependant, la proportion de candidats-sèmes exclusivement présents dans la liste globale est loin d'être négligeable. Chaque approche a donc son propre apport qualitatif, et, malgré les différences de nature des unités, la convergence vers des classes similaires s'opère.

	proportion de candidats de la liste globale présents dans une autre période	*proportion de candidats de la période présents dans la liste globale*				
		1	2	3	4	5
JUDICIAIRE	89%	7%	36%	33%	22%	12,50%
CRIME	40%	18%	43%	50%	0	–
DRAME - DESASTRE	43%	–	75%	–	–	25%
LIEU	75%	33%	–	–	–	–

Figure 7 : Proportion de candidats-sèmes communs aux classes similaires de l'approche globale et par période

6. Discussion conclusive

Deux approches ont permis de proposer des sémèmes *d'Outreau* et les valider. Ces deux approches intègrent par écho de l'information apportée sur le plan lexical (sélection des candidats-sèmes coactualisés avec une de leurs formes lexicales de référence) ; de plus, cette information est amplifiée sur le plan sémique par renforcement de spécificité. Les approches sont toutes deux en accord avec les analyses manuelles : elles parviennent à faire émerger les mêmes tendances saillantes et mettent en évidence une évolution diachronique conforme à l'étude manuelle. Le sémème généré par période est qualitativement plus riche, puisqu'il permet de nuancer les classes saillantes et accroît la diversité des représentants au niveau des classes communes aux deux approches. Toutefois, l'approche par période s'appuie principalement sur l'écho, donc sur des informations aussi bien caractéristiques des formes que des sèmes, tandis que l'approche globale adopte une position plus centrée sur le plan sémique et exploite davantage l'apport propre des sèmes, puisque, certes, l'écho intervient dans la sélection des candidats-sèmes, mais les indicateurs sont propres aux sèmes. Par ailleurs, la perspective des deux approches diffère : la première cherche à obtenir une image qualitative ponctuelle d'*Outreau*, propre à une période donnée, alors que l'autre cible une représentation sur l'ensemble du corpus qui, ensuite, se structure dans le temps par activation ou inhibition de candidats.

La convergence entre analyses manuelle et automatique montre la validité de l'approche automatique. Néanmoins, la confrontation repose sur l'identification manuelle de classes, extraites à la fois de traitements de résultats issus de la procédure semi-automatique et d'un regard orienté par l'étude linguistique. Cette émergence de classes, comme par exemple celle correspondant à la dimension judiciaire et policière, est le résultat d'un processus d'interprétation complexe et que l'on ne sait, à ce jour, pas encore décrire précisément. Cependant, des outils théoriques tels que les fonds sémantiques (constitués d'isotopies, c'est-à-dire de récurrences de sèmes) et formes sémantiques (groupements de traits sémantiques) semblent adaptés pour préciser les processus interprétatifs à l'oeuvre. En particulier, les récents concepts de *diffusion* et de *sommation*, empruntés à (Rastier, 2006), rendent compte des échanges sémiques entre fonds et formes. Propres à modéliser les interactions entre le signifié d'une unité lexicale et un faisceau d'isotopies locales, ils ouvrent des perspectives pour modéliser une construction dynamique de sens.

Références

Dendien J. and 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) : 11-37.

Grzesitchak M., Jacquey E. and Valette M. (2007). Systèmes complexes et analyse textuelle : Traits sémantiques et recherche d'isotopies. *ARCo'07*, 227-235.

Jacquemin P., Lauf A., Poudat C., Hurault-Plantet M. and Auray N. (2008). La fiabilité des informations sur le web : le cas Wikipedia. *CORIA2008*. pp. 227-235.

Lafon P. (1984). *Dépouillements et statistiques en lexicométrie.* Slatkine- Champion.

Lecolle M. (2007). Polysignifiance du toponyme, historicité du sens et interprétation en corpus. Le cas de *Outreau*. Corpus, 6 : 101-125.

Lecolle M. (2009). Changement de sens du toponyme en discours : de Outreau « ville » à Outreau « fiasco judiciaire ». In LecolleM., Paveau A.-M. and Reboul-Toure S., editors, *Le nom propre en discours, Les Carnets du Cediscor,* 11 : 91-106.

Mayaffre Damon (2002). « Les corpus réflexifs : entre architextualité et hypertextualité », *Corpus*, 1 : 51-69.

Rastier F. (1987). *Sémantique interprétative*. PUF.

Rastier F. (2006). Formes sémantiques et textualité. *Langages*, 163 : 99-114.

Rastier F. and Valette M. (2009). De la polysémie à la néosémie, *Le français moderne*, in S. Mejri, éditor, *La problématique du mot*, 77 : 97-116.

Reutenauer C., Valette M. and Jacquey E. (in press). De l'annotation sémique globale à l'interprétation locale : environnement et image sémiques d' « *économie réelle* » dans un corpus sur la crise financière ». *ARCo'09*.

Sablayrolles F. (2002). Fondements théoriques des difficultés pratiques du traitement des néologismes. *Revue française de linguistique appliquée*, VII(1) : 97-111.

Salem A., Lamalle C., Martinez W., Fleury S., Fracchiolla B., Kuncova A. and Maisondieu A. (2003). *Lexico3 – Outils de statistique textuelle. Manuel d'utilisation*. Syled-CLA2T, Université de la Sorbonne nouvelle – Paris 3 [logiciel disponible sur http://www.cavi.univ-paris3.fr/Ilpga/ilpga/tal/lexicoWWW].

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