Méthode des cooccurrences : recherche sémantique sur le nom propre

Serge Heiden, Lamria Chetouani

Sémantique des noms propres
Méthode des cooccurrences

Lamria Chetouani, Serge Heiden

ENS – UMR 8503 – Grille d'honneur du Parc – 92211 Saint-Cloud – France

lamria.chetouani@bretagne.iufm.fr,
slh@ens-fcl.fr

Abstract

This research is interested in the proper noun. The objective is to determine the variety of employment and the lexical context which give meaning to the proper noun in discourse.
Our implementation of the method of co-occurrences rests on an hypertextual interface. This interface optimises the permanent comings and goings between the reading of listings of co-occurrents words and between the validation in context of their affective attraction by kwic concordances. The listings, called lexicograms, show simultaneously a particular word with the two lists of its left and right co-occurrents. Those lists are sorted by the probability that a word should meet the particular one the number of times found in the corpus and even more.

Résumé

Cette recherche s'intéresse au nom propre. L'objectif est de déterminer, la variété de ses emplois et des entourages lexicaux qui lui donnent du sens dans le discours.
Notre mise en œuvre de la méthode des cooccurrences repose sur une interface hypertextuelle, optimisant le va-et-vient permanent entre l'analyse du classement statistique des couples de formes cooccurrentes à l'intérieur des phrases, et la validation en contexte de ces attirances par la lecture des concordances de ces couples dans le texte. Les synthèses d'attirances de formes se présentent sous la forme de lexicogrammes affichant simultanément une forme pôle choisie et ses cooccurrents gauches et droits classés par la probabilité qu'ils se rencontrent le nombre de fois constaté effectivement dans le corpus ou plus.

Mots-clés : statistique textuelle, cooccurrences, lexicogrammes, concordances, hypertexte, discours, sémantique, nom propre.

1. Problématique des noms propres

Contrairement à ceux qui parlent de vacuité sémantique et de monoréférentialité ou de "désignation rigide", cette recherche soutient l'idée que les noms propres (désormais Npr) n'ont pas de sens stable et fixé une fois pour toutes par lexicalisation. Leur signification est assujettie au contexte situationnel de l'énonciation et au réglage de sens lors de leur actualisation. Ils sont assujettis aussi à l'évolution du référent et surtout à la variation des inférences qu'on peut tirer de leur évocation laquelle éveille en nous des quantités d'"idées". En effet, les Npr ont un grand pouvoir d'évocation. Ils influencent notre imagination en bien ou en mal en créant des associations d'esprit et des images.
Ils créent surtout des associations lexicales au sein du discours où ils sont employés car ils fonctionnent avec une prédilection pour la compagnie d'un ou de plusieurs mots qui les suivent (ou les précèdent) et qui sont révélateurs de leur registre symbolique. Ces termes accompagnateurs peuvent être immédiatement collés à eux ou séparés d'eux d'*une* à *n* unités intermédiaires. L'objectif de cette étude qui s'appuie sur l'analyse des cooccurrences, est de défendre l'idée que la signifiance du Npr dépasse la problématique saussurienne du signe en rendant compte de la variabilité des sens actualisés en discours. Nous essaierons d'examiner les choix dominants, opérés par le narrateur, parmi le paradigme des couples aptes à désigner un même objet en mobilisant diverses valeurs langagières (historiques, socio-politiques et culturelles). L'ensemble des informations peu à peu accumulées peut orienter le lecteur vers telle ou telle représentation du référent car, selon nous, les alliances ne se construisent pas de façon aléatoire mais suivant une logique interne au discours.

2. Fonctionnement lexico-discursif des toponymes chez Jury

Nous avons choisi d'analyser spécifiquement les toponymes les plus fréquents (*"France"* et *"Algérie"*) dans un corpus homogène de deux romans de Maurice Jury : Le *Péché d'omission* et *Tala*. Cette option est motivée, d'une part, par le fait que les ouvrages traitent d'une période cruciale de l'identité des deux nations (entre 1958 et 1961). La simple évocation de ces noms de pays éveille une mémoire historique qui n'est pas sans produire des valeurs symboliques. D'autre part, l'intérêt de cette étude tient à ce que ces mots sont analysés "en situation", dans leur fonctionnement réel en discours et non pas dans leur définition dictionnairique. Ce n'est donc pas une approche lexicographique mais une analyse à la fois mathématique et lexico-discursive que S. Heiden et P. Lafon (Heiden et Lafon, 1998), situent dans une *"position médiane (...) à peu près équidistante du vocabulaire et du discours"*. Ils la définissent *"comme une tentative, presque désespérée, de saisir les mots au moment où ils ne sont plus tout à fait des mots mais déjà du discours."* Il s'agit là de l'appréhension du jeu de va-et-vient entre les mots et le discours car l'actualisation du Npr dans le discours, met en œuvre le fonctionnement pragmatique qui ne peut être compris sans la caractérisation de la situation de communication. En l'occurrence, le locuteur, sous l'emprise de la détermination culturelle, historique, sociale et de son expérience personnelle du monde, donne son point de vue en nommant les pays. Il façonne ainsi des représentations symboliques, au moyen du langage. Chaque emploi toponymique produit du "sens". La totalité des "sens" qui se manifestent dans le corpus ne constitue qu'un aspect de la représentation du référent "exprimable" par le locuteur. Elle ne correspond qu'à une facette particulière, mais jamais à l'essence (ou à la propriété ontologique) de l'objet nommé. Le désignateur ne nomme pas le réel "en soi" mais tel qu'il le perçoit, et selon le rapport qu'il entretient avec lui. C'est ce fonctionnement dynamique de la nomination qui permet de définir la variabilité sémantique du mot.

3. Analyse contextuelle et méthode des cooccurrences

La méthode informatique des cooccurrences permet de déterminer la variété des emplois et des entourages lexicaux qui donnent du sens au nom propre dans le discours. En se focalisant sur le jeu des attirances des formes lexicales au sein des énoncés, elle concilie l'étude du vocabulaire et du discours. Notre analyse qui ne s'intéresse qu'à deux formes particulières (*France* et *Algérie)* n'étudie que les couples formés par ces deux pôles et les formes qui viennent fréquemment avec eux, à leur droite ou à leur gauche, dans les phrases du discours. Le système mis en œuvre étant binaire, le programme relève tous les couples de cooccurrents du corpus et les classe de façon exhaustive. La probabilisation des rencontres de formes à l'intérieur des phrases a été formulée par Pierre Lafon (Lafon 1984). Étant donnés le nombre de phrases et la fréquence de chaque forme, il s'agit d'établir le rapport entre toutes les combinaisons possibles de faire se rencontrer ces deux formes à l'intérieur de certaines phrases et toutes les manières possibles de répartir ces formes parmi les phrases du texte.
Notre mise en œuvre de la méthode des cooccurrences repose sur une interface hypertextuelle par Intra/Internet optimisant le va-et-vient permanent entre l'analyse du classement statistique des couples de cooccurrents et la validation de ces attirances en contexte par la lecture des concordances des couples effectifs du texte. Les synthèses exhaustives d'attirances se présentent sous la forme de lexicogrammes présentant simultanément une forme pôle choisie et ses cooccurrents gauches et droits classés par la probabilité (voir figure 2). Chaque cooccurrent est renseigné par sa fréquence (f), le nombre de ses rencontres avec le pôle (cf), la probabilité de cooccurrence (p) et sa distance moyenne au pôle (d_m). Chaque liste de cooccurrents au pôle est exclusivement élaguée en seuillant la valeur de ces renseignements. Le calcul de la concordance d'un couple particulier est déclenchable directement par un simple clic sur le nombre cf de rencontres effectives des deux formes dans le texte (voir figure 3). Diverses possibilités de tris multi-critères des lignes de la concordance permettent ensuite de systématiser l'analyse des contextes d'apparition de chaque couple. Un lien hypertextuel relie alors directement chaque ligne de concordance à la page de l'édition en ligne du texte contenant le couple pour élargir rapidement la lecture de son contexte d'apparition.

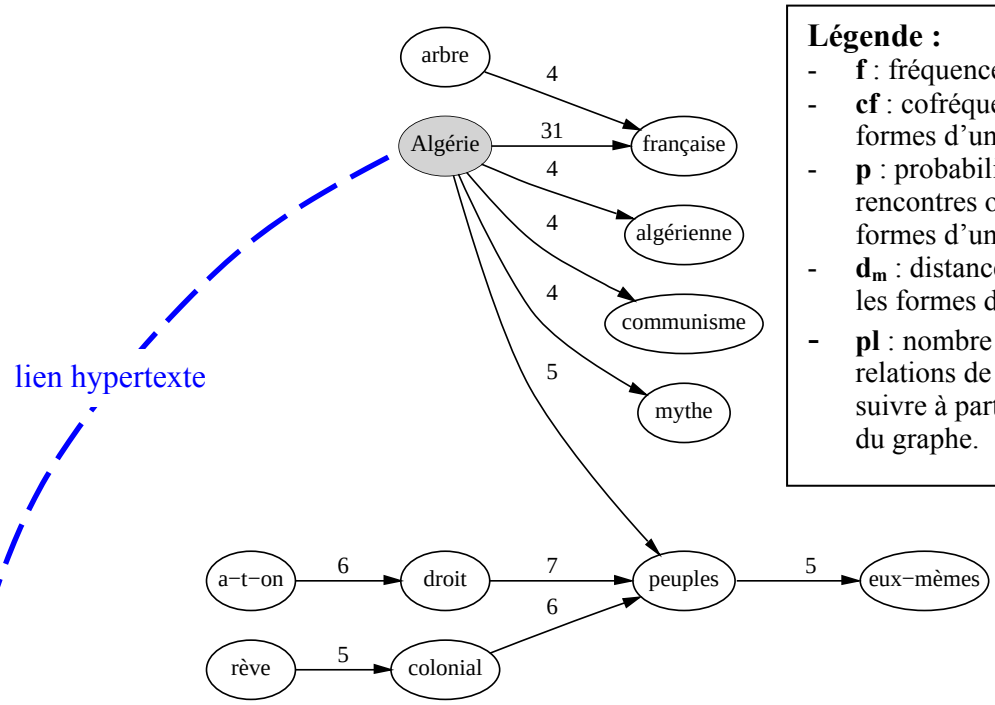
Enfin, chaque cooccurrent au pôle générant son propre espace de cooccurrence, nous avons ménagé un lien hypertextuel vers le calcul de son propre lexicogramme à partir de sa forme. Nous proposons, finalement, une synthèse du parcours récursif de l'ensemble des lexicogrammes du vocabulaire accessibles à partir d'un pôle sous la forme de graphes dessinés automatiquement (voir figure 1). Dans ces graphes chaque sommet correspond à une forme et chaque arc, à une relation de cooccurrence éventuellement étiqueté par l'ordre de grandeur de la probabilité. Chaque sommet correspond au pôle d'un lexicogramme particulier, un lien hypertextuel le relie directement au calcul de son lexicogramme détaillé. Nous présentons une application de cette navigation hypertextuelle à travers les quatre niveaux de synthèse successifs disponibles : les graphes de formes cooccurrentes ↔ les lexicogrammes autour d'une forme pôle ↔ les concordances de couples de formes cooccurrentes ↔ l'édition en ligne du texte.

4. Valeurs sémantiques des Npr étudiés

Il serait illusoire d'envisager une définition "complète" des deux entités qui sont porteuses d'autres valeurs non identifiées par le programme car non exprimables concrètement par les mots et le discours. Nous ne prétendons donc, en aucun cas, avoir exploré le "sens total" des deux mots mais seulement quelques significations à travers les couples dégagés par l'analyse. Le lexicogramme de *Algérie* met en évidence, parmi les termes intéressants qui accompagnent ce mot, soit à droite soit à gauche, les termes *française* et *guerre* . La violence et la guerre se sont imposées comme une matière première dans les romans de Jury. Par ailleurs, l'adjectif *française* se situant en tête de liste, de part et d'autre du pôle, est à très faible distance à droite; il est quasiment collé à *Algérie*. Cette forme s'avère, de ce fait, très impliquée (première dans la liste hiérarchisée selon l'ordre des probabilités) dans la constitution du couple lexical, dans le fonctionnement discursif et, par là dans le système référentiel, voire dans le contenu sémantico-thématique du corpus. Le couple reflète une réalité historique. L'Algérie n'avait pas, en effet, d'autonomie fonctionnelle en politique, à cette époque. Le lexicogramme du pôle "*France* " permet de constater que le slogan *"vive la France"* s'inscrit dans les deux romans, dénotant un climat de conflit et de trouble. Les autres mots qui précèdent *France ,* sur le plan syntagmatique sont des termes à valeur locative ou socio-politique qui lui assignent une image de liberté.
Bien que les lexicogrammes de *France* et *Algérie*, ne soient pas très riches en cooccurrents, nous pouvons dire que la méthode a pu mettre en évidence quelques valeurs sémantiques relatives à l'aspect historique, culturel et social des mots étudiés. Valeurs que l'œil nu n'aurait peut-être pas vues de façon aussi rapide. Le programme a permis également de montrer, grâce aux cooccurrences et aux contextes, que les deux noms de lieux ont des valeurs sémantiques totalement opposées. *France* donne l'idée d'un pays authentique, réel, libre, autonome, agréable, tandis que *Algérie* suggère l'idée d'un espace irréel, mythique, double, déchiré, en guerre, sous la violence.

Références

Bergounioux A. et al. (1982). *La parole syndicale*. Paris, PUF, pages 187–219.
Chetouani L. (1998). Mémoires dans des jeux de miroirs. In *Des mots en liberté,* ENS-Édition, Fontenay/Saint-Cloud, pages 291–303.
Heiden S. (1999). *Lexploreur : Manuel Utilisateur,* v2.3, UMR 8503, http://diderot.lexico.ens-fcl.fr/doc/lexploreur/, 100 pp.
Heiden S. et Lafon P. (1998). Cooccurrences : La CFDT de 1973 à 1992. In *Des mots en liberté,* ENS-Édition, Fontenay/Saint-Cloud, pages 65–83.
Lafon P. (1984). *Dépouillements et statistiques en lexicométrie*. Genève-Paris, Slatkine-Champion
Tournier M. (1987). Cooccurrences autour de "*travail"*. In *Mots n° 14*, pages 89–123.



Légende :
- **f** : fréquence d'une forme ;
- **cf** : cofréquence des deux formes d'un couple ;
- **p** : probabilité de cf rencontres ou plus entre les formes d'un couple ;
- **d_m** : distance moyenne entre les formes d'un couple ;
- **pl** : nombre maximum de relations de cooccurrence à suivre à partir de la racine du graphe.

lien hypertexte

Figure 1
Lexicogramme récursif autour du pôle "Algérie" dans le corpus tala-péché.
Seuils : f ≥ 3, cf ≥ 2, p≤ 0,04%, d_m ≤ 1000.0 occ, pl ≤ 1000 arcs.

Algérie
(72)

cooccurrents gauches

	f	cf	p	d_m
française	26	3	2.443e-03	7.0
grand	48	3	1.367e-02	12.3
guerre	90	4	1.416e-02	3.2
armée	50	3	1.525e-02	10.0

cooccurrents droits

	f	cf	p	d_m
française	26	18	3.705e-31	1.1
peuples	9	3	9.041e-05	3.3
algérienne	12	3	2.312e-04	0.0

Figure 2
Lexicogramme du pôle "Algérie" dans le corpus tala-péché.
Seuils : f ≥ 3, cf ≥ 3, p≤ 5%, d_m ≤ 1000.0

1	aux musulmans est une invention de réactionnaires françaouis ; c' était bon pour certains politiciens des années 50 . en	**Algérie , tous vivaient accrochés au mythe de l' Algérie entièrement française**	- c' est ce qui leur avait plu dans l' élan oratoire qui joignait Dunkerque à Tamanrasset - . l' idée ultérieure de regrouper européens
1	se doivent d' être de la partie . d' où cette confidence faite un jour à Tipasa , si inutile . cette confidence-piège . l'	**Algérie a d' abord été pour lui une " moisson " de cheveux blonds et des yeux riant au ciel , une française**	un peu folle , qui courait les douars avec des camionnettes de lait et que j' imagine belle comme sa générosité . en vertu d' une
1	réactionnaires françaouis ; c' était bon pour certains politiciens des années 50 . en Algérie , tous vivaient accrochés au mythe de l'	**Algérie entièrement française**	- c' est ce qui leur avait plu dans l' élan oratoire qui joignait Dunkerque à Tamanrasset - . l' idée ultérieure de regrouper européens
1	, l' air hébété . un croisement de voix sourdes , de bruits de bennes et de klaxons obstinés , s' agaçant au rythme d' "	**Algérie française**	! " , bourdonne à nos oreilles . il a rapproché sa chaise pour me parler . pour la seconde fois de ma vie , je me trouve dans un café en
1	de vivre ? des furieux m' ont injuriée plusieurs fois derrière les grilles . ils crient en défilant , mains levées , "	**Algérie française**	! " ; des armes noires courent sur le dessus du cortège ; des poings sont venus tambouriner à la porte d' en-bas , comme

Vers la page correspondante de l'édition en ligne

Figure 3

Premières des 18 lignes de la concordance du couple (Algérie, française) dans le corpus tala-péché.
Les lignes sont d'abord triées par leur pivot, puis par le contexte droit, puis par le contexte gauche.