

ÉTUDE SIMULTANÉE DE LA SYNONYMIE ET DE L'ANTONYMIE EN CONTEXTE PAR ANALYSE DISTRIBUTIONNELLE CONTRAINTE

Jean-Luc MANGUIN (1) ; Itsuko FUJIMURA (2) ;
Guillaume JACQUET (3) ; Fabienne VENANT (3)

(1) CRISCO, CNRS & Université de Caen ;

(2) Université de Nagoya ; (3) LATTICE, CNRS & Université de Paris 61.

OBJECTIFS DE L'ÉTUDE

Cette étude se propose d'analyser et de représenter le sens d'un adjectif du lexique par un objet géométrique visualisable, dont les différentes régions correspondent aux différents sens pris en contexte nominal. Nous avons déjà montré (Manguin, à paraître) qu'il est possible d'établir une cartographie sémantique d'une telle unité lexicale à partir de ses propres cooccurrences nominales et de celles de ses synonymes, recueillies dans un corpus, pourvu que l'on applique à celles-ci certaines restrictions décrites dans le paragraphe suivant ; la représentation obtenue permettait d'observer le déploiement du sens de l'adjectif étudié, et de constater quels étaient les contextes associés à ces différentes acceptions. Cette fois, nous voulons répondre à la question de la stabilité de cette représentation en fonction du corpus d'observation, en particulier nous désirons savoir si les contextes nominaux (ou peut-être les « classes » de contextes) sont répartis de la même manière. Et d'autre part, nous souhaitons diminuer les contraintes de cette analyse distributionnelle particulière, en incluant les antonymes de l'adjectif étudié dans l'ensemble paradigmatique employé lors de la première partie de cette étude.

2. MÉTHODE ET CORPORA

2.1. Méthode

Notre méthode d'analyse des cooccurrences est un processus en plusieurs étapes, dans lequel les résultats obtenus à un certain stade doivent être transmis au traitement suivant. Comme nous l'avons dit, une fois que nous avons choisi l'adjectif à étudier, nous recherchons dans le corpus quels sont les noms dont il peut être épithète, autrement dit nous devons repérer dans le corpus les occurrences « NOM + ADJ » ou « ADJ (+ ADV) + NOM », où ADJ est notre adjectif dans toutes ses formes possibles ; cette manière de procéder implique que le corpus soit au moins balisé, avec indication de la catégorie et du lemme ; le balisage fonctionnel n'est pas nécessaire, mais d'autre part les erreurs éventuelles de l'analyseur qui pose ces balises nous obligent à vérifier les exemples recueillis pour les débarrasser des fautes de catégorisation automatique. À l'issue de cette étape, nous possédons la liste des noms qui, dans le corpus étudié, coexistent avec notre adjectif ; nous ne gardons de cette liste que les éléments les plus fréquents, soit normalement ceux dont les exemples représentent plus de 0,5 % du total des cooccurrences. Nous voyons déjà que nous avons imposé dès cette première étape une contrainte à l'analyse distributionnelle que nous effectuons : cette contrainte est

catégorielle, puisque nous imposons au cooccurrents de notre unité d'appartenir à la catégorie nominale.

L'étape suivante voit l'apparition de la seconde contrainte, cette fois d'ordre paradigmatique : nous allons à présent relever les cooccurrences de même structure que précédemment, mais en exigeant que NOM soit un membre de la liste constituée à la première étape, et que ADJ soit un synonyme ou un antonyme de l'adjectif étudié ; c'est dans cette deuxième condition que réside la contrainte supplémentaire de notre analyse distributionnelle. Le choix des synonymes et des antonymes se fait au moyen d'un dictionnaire élaboré au laboratoire Crisco, et déjà décrit dans de précédentes publications. Le principe sous-jacent à cette contrainte est que la substituabilité de l'unité étudiée par un de ses synonymes est liée au contexte nominal ; en outre, le fait d'inclure les antonymes permet d'élargir la substitution synonymique au remplacement par un antonyme précédé d'un adverbe de négation ou d'affaiblissement.

Le traitement automatique des exemples recueillis permet de disposer de paires adjectivo-nominales et de leurs fréquences absolues dans le corpus, que nous pouvons présenter sous forme d'un tableau NOM vs ADJ, dans lequel le contenu de chaque case sera la fréquence de la paire correspondante ; nous transformons alors l'ensemble du tableau en fréquences relatives, en divisant chaque fréquence absolue par la fréquence absolue (dans le corpus) de l'adjectif impliqué dans la paire. Ceci nous permet de nous affranchir des fréquences adjectivales, et de comparer les adjectifs uniquement par les proportions d'emploi avec les noms de la liste.

La dernière étape consiste à appliquer à ces données une « métrique du chi² », puis à effectuer sur ces nouvelles valeurs une analyse factorielle des correspondances, ce qui nous donne les coordonnées des différents points de la représentation. Ce traitement final s'inspire directement du processus décrit dans Jacquet & Venant (2005).

2.2. Corpora

Comme nous venons de l'indiquer, nous utilisons des corpus balisés, assez larges afin de pouvoir recueillir un nombre suffisant d'exemples. Le premier corpus est la base Frantext catégorisée, qui s'étend de 1830 à nos jours, et dont nous avons exclu les textes poétiques ; il demeure 1711 textes, représentant un total de 120 millions de mots. Le second corpus est un ensemble de textes de journaux et de magazines contemporains : *Le Monde* (1987 – 1992), *Le Monde Diplomatique* (1984 – 1998), *Libération* (1995 – 1999) et *Le Point* (1995 – 2000) ; les textes comptent 208 millions de mots au total, et ils sont tous balisés par le logiciel TreeTagger. Ce corpus a déjà été utilisé dans d'autres publications (voir Fujimura, Uchida & Nakao, 2004).

3. RÉSULTATS

Rappelons que cette étude prolonge les travaux déjà présentés en 2003 lors des précédentes journées de linguistique de corpus de Lorient ; c'est la raison pour laquelle nous avons décidé de continuer sur le même adjectif, en l'occurrence *curieux*, qui possède la particularité déjà signalée (voir Bat-Zeev Schyldkrot, 1997 ; François, Victorri & Manguin, 2003) de changer de sens en changeant de position par rapport au substantif dont il est l'épithète ; ce phénomène se traduit assez clairement, comme nous le verrons, sur la forme de son espace sémantique.

3.1. Contextes nominaux

Nous donnons ci-dessous les contextes nominaux obtenus pour chacun des corpus, suivant la méthode exprimée plus haut ; nous avons indiqué en gras les contextes partagés par les deux ensembles de textes, afin de les conserver comme points de comparaison.

Les contextes issus de Frantext (25) :

bête, cas, chose, conversation, détail, esprit, exemple, expérience, fait, figure, histoire, livre, mélange, objet, oeil, page, personnage, phénomène, regard, spectacle, spécimen, travail, type, tête, étude.

Les contextes issus des journaux et magazines (55) :

affaire, alchimie, alliance, ambiance, atmosphère, attitude, bête, chassé-croisé, chose, cocktail, comportement, conception, couple, coïncidence, effet, esprit, exercice, façon, film, forme, gens, habitude, histoire, homme, idée, impression, lecteur, livre, manière, mélange, méthode, nom, objet, oeil, opération, paradoxe, pays, personnage, phénomène, pratique, public, rapport, regard, renversement, retour, retournement, roman, résultat, sensation, sentiment, situation, spectacle, système, tour, visiteur.

Nous remarquons dès à présent que de nombreux contextes trouvés par Frantext apparaissent dans la liste de ceux trouvés dans le corpus de journaux et magazines, et que parmi ces contextes communs, certains comme *regard* permettent à *curieux* de changer de sens suivant sa position (*un curieux regard* vs *un regard curieux*).

3.2. Représentation avec synonymes

Dans la suite du processus, nous recherchons dans nos corpus les cooccurrences des différents noms avec les synonymes de *curieux* pris dans notre dictionnaire des synonymes du CRISCO ; dans le cas présent, les synonymes retenus sont :

amusant, anxieux, attachant, attentif, avide, bizarre, chercheur, drolatique, drôle, déconcertant, extraordinaire, extravagant, fureteur, incompréhensible, inconcevable, incroyable, indiscret, inouï, inquisiteur, intéressant, intéressé, investigateur, original, paradoxal, piquant, pittoresque, plaisant, rare, singulier, soucieux, surprenant, unique, étonnant, étrange.

Quel que soit le corpus utilisé, les exemples recueillis (plusieurs milliers, vérifiés manuellement) montrent que toutes les associations possibles n'apparaissent pas dans les textes ; par exemple, dans le cas du corpus Frantext, sur les 1750 paires possibles, on n'en observe que 414 qui sont réalisées.

Nous donnons aux figures 1 et 2 les représentations obtenues respectivement avec Frantext et avec le corpus de presse contemporain ; la ressemblance des deux figures est frappante : dans les deux cas, les synonymes se regroupent avec une grande pertinence sémantique, mais en outre, les adjectifs sont clairement dissociés en deux groupes : à droite le sens de curieux « intéressé » et à gauche celui de curieux « intéressant » ; qui plus est, la zone de gauche montre la même évolution géométrique du « surprenant » et de l'« inexplicable » (en haut) jusqu'au « pittoresque » et à l'« attachant » (en bas). Par souci de lisibilité, nous n'avons pas étiqueté la zone centrale, mais les figures complètes de notre article peuvent être consultées sur <http://www.crisco.unicaen.fr/>.

Au point de vue des contextes, les plus agentifs (*œil, regard, voire esprit*) se placent logiquement à droite ; ceux qui se placent à gauche montrent que leur emploi avec *curieux*

les place dans une classe de patients ; ainsi *personnage* n'est pas représenté parmi les agents. De plus, ces contextes « patients » sont associés aux synonymes de la même manière dans les deux corpus : dans la zone de l'« inexplicable », on trouve des contextes génériques comme *phénomène, chose, cas, situation* ou encore *effet*, tandis que les contextes désignant des entités précises se rencontrent dans la partie du « pittoresque », comme *personnage, histoire, livre, détail, ou roman*. L'explication de ces placements est relativement simple : ce qui est *inexplicable, surprenant* ou *paradoxal* est aussi difficilement reconnaissable (peut-être à cause de l'émotion du moment) et par conséquent difficile à nommer ; à l'opposé, ce qui est *plaisant* et *amusant* est facilement caractérisé et donc nommable.

Globalement, nous pouvons dire que dans le cas des cooccurrences avec synonymes seuls, il n'y a pas de différence de représentation en faisant varier le corpus, ce qui en soit constitue un résultat intéressant.

3.3. Représentation avec synonymes et antonymes

Nous avons répété notre analyse en relevant ensuite les occurrences des noms de la liste et d'un antonyme de *curieux* ; les antonymes retenus sont :

anodin, banal, commun, courant, discret, habituel, indifférent, insouciant, ordinaire, quelconque, réservé, usuel, vulgaire.

Les représentations obtenues sont données par les figures 3 et 4, et l'on remarque une certaine déformation de la disposition, sans que l'on puisse parler de bouleversement (les antonymes sont repérés par l'encadrement de leur étiquette).

La visualisation obtenue avec le corpus de périodiques est la moins déformée, et le placement des antonymes est des plus logiques ; par rapport à la figure 2, on remarque que l'axe vertical est en quelque sorte « inversé », et que la plupart des antonymes se trouvent dans la zone de gauche qui correspond aux contextes « patients ». Seul *indifférent* se situe dans la zone des agents, ce qui correspond à son acception habituelle. Le placement détaillé des autres antonymes n'est pas toujours en correspondance avec leurs opposés parmi les synonymes ; on peut toutefois signaler que *réservé* et *courant* sont plutôt logiquement situés.

La représentation issue du corpus Frantext appelle plus de commentaires ; tout d'abord, les antonymes situés dans la zone de gauche, comme *anodin, quelconque, vulgaire, usuel, courant* ou *habituel* possèdent ici un emplacement plus proche de leur(s) opposé(s) parmi les synonymes ; par contre, *réservé* et *insouciant* sont assez excentriques, ce qui peut mettre en doute leur emploi d'antonyme de *curieux* dans leurs contextes préférentiels qui sont *conversation réservée* et *esprit insouciant*. D'autre part, la position de *discret* à mi-chemin des deux zones « patients » et « agents » peut laisser penser que son emploi dans *regard discret* est parfois antonymique de *regard indiscret*, mais l'examen des exemples recueillis infirme cette hypothèse ; nous touchons là les limites de notre méthode représentative. Enfin, la position de *indifférent* peut surprendre : situé lui aussi à mi-chemin des deux zones, il semble qu'il puisse qualifier des agents comme des patients ; cette fois, les exemples révèlent que notre représentation n'est pas trompeuse, puisque des exemples de *choses indifférentes* (c'est-à-dire de choses qui laissent les gens sans réaction) abondent dans la littérature du 19^{ème} siècle, et reflètent un emploi soutenu qui n'est plus d'actualité.

Ainsi notre méthode permet de mettre en lumière certaines évolutions diachroniques lors des comparaisons entre corpus de différences époques ; cela dit, son « mécanisme » ne

doit pas être perdu de vue, sous peine de conduire à des interprétations fausses en ce qui concerne les liens paradigmatiques (comme dans le cas de *discret*).

4. BILAN ET PERSPECTIVES

À l'issue de cette étude, nous pouvons affirmer que la méthode que nous avons mise en œuvre s'avère donner des résultats stables malgré les changements de corpus ; de plus, la stabilité globale ne gomme pas les variations qui caractérisent les emplois de l'unité étudiée selon les textes retenus pour l'application du traitement. En revanche, l'élargissement paradigmatique opéré n'a pas tout à fait rempli nos espérances, et ceci pour deux raisons : la première vient du fait que l'hypothèse de substituabilité formulée au départ ne se vérifie pas dans la réalité, autrement dit on rencontre assez peu d'adjectifs épithètes précédés d'un adverbe de négation ou d'affaiblissement. Dans notre travail, nous n'avons pas cherché de tels cas de substitution, mais plutôt les emplois antonymiques, cependant il est probable que l'application de notre méthode au cas de l'adjectif attribut serait probablement plus fructueuse, mais les corpus employés nécessiteraient alors un balisage plus élaboré, qui devrait résoudre les problèmes de référence, vu le nombre d'occurrences où l'adjectif est attribut d'un pronom. Par ailleurs, il n'est pas exclu qu'une différence de nature entre la synonymie et l'antonymie conduise, lors de la transformation d'un énoncé en son contraire, à une double substitution du nom et de l'adjectif, par exemple *un homme intelligent* devient *un gars bête* (ou *un type bête*) ; ceci peut expliquer l'absence de certaines occurrences escomptées.

La deuxième raison réside dans la polysémie des adjectifs, qui est un problème déjà rencontré avec *sec* dans un traitement automatique proche du nôtre (voir Venant, 2004). Pour le résumer, disons simplement qu'un adjectif synonyme de *sec* pour un contexte donné peut se rencontrer avec un autre contexte sans qu'il y ait cette fois de synonymie ; par exemple : *accueil sec / accueil glacial*, comparé à *temps sec / temps glacial*. Dans notre cas, nous pouvons rencontrer *personnage commun*, mais aussi *histoire commune*, où cette fois l'antonymie avec *curieux* est loin d'être certaine.

Ainsi, les résultats prometteurs obtenus dans la direction de la comparaison de corpus au travers d'une unité lexicale devront être complétés, dans le domaine de l'étude antonymique, par une réflexion méthodologique et des outils plus élaborés.

5. RÉFÉRENCES

- Bat-Zeev Schyldkrot H.**, 1997. 'Synonymie et polysémie, le cas de CURIEUX comme parcours sémantique d'un mot', *Langages*, 128, p. 113-125.
- François J., Victorri B. & Manguin J.-L.**, 2003. 'Polysémie adjectivale et synonymie : l'éventail des sens de curieux', in Soutet O. (ed.) *La polysémie*, Paris : Presses de l'Université de la Sorbonne.
- Frantext, base textuelle catégorisée*, 1999. CNRS, ATILF (Analyse et traitement informatique de la langue française), UMR CNRS-Université Nancy2, <http://www.inalf.fr/atilf>.
- Fujimura, I., Uchida, M., Nakao H.**, 2004. 'De vs des devant les noms précédés d'épithète en français : le problème de petit', in *Le poids des mots, Actes des 7^{èmes} Journées internationales d'Analyse statistique des Données Textuelles*, Louvain La Neuve.
- Jacquet G., Venant F.**, 2005. 'Construction automatique de classes de sélections distributionnelles', *12^{ème} conférence annuelle sur le traitement automatique des langues* (TALN 05), Dourdan.

- Manguin J.L.**, (à paraître). 'Utilisation d'un corpus catégorisé pour l'étude et la représentation de la synonymie en contexte' in Williams G. (ed.), *Actes des 3^{èmes} journées de linguistique de corpus*, Paris : L'Harmattan.
- Venant F.**, 2004. 'Polysémie et calcul du sens' in *Le poids des mots, Actes des 7^{èmes} Journées internationales d'Analyse statistique des Données Textuelles*, Louvain La Neuve.

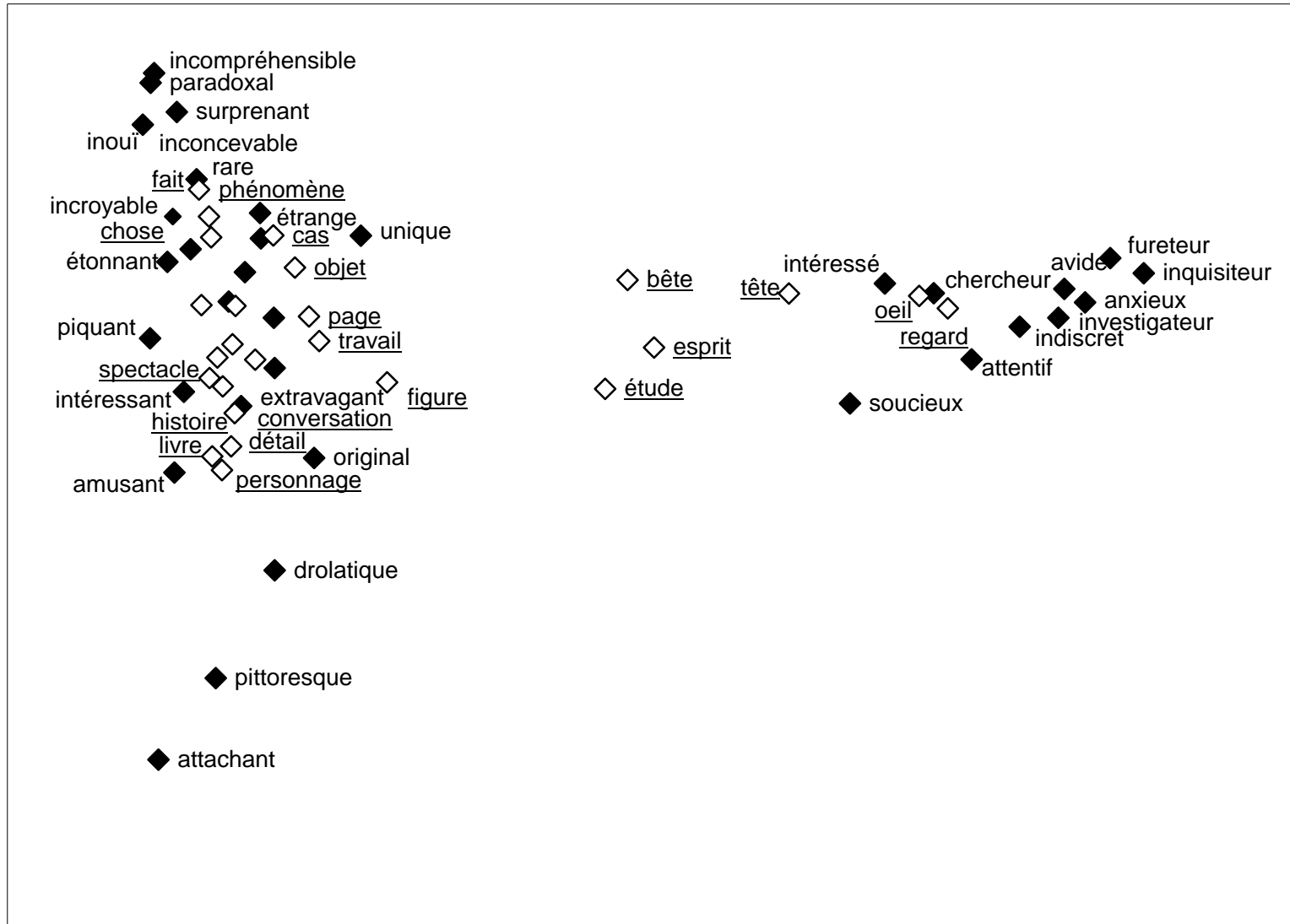


Figure 1 : espace sémantique obtenu avec les synonymes et les cooccurrences dans le corpus Frantext

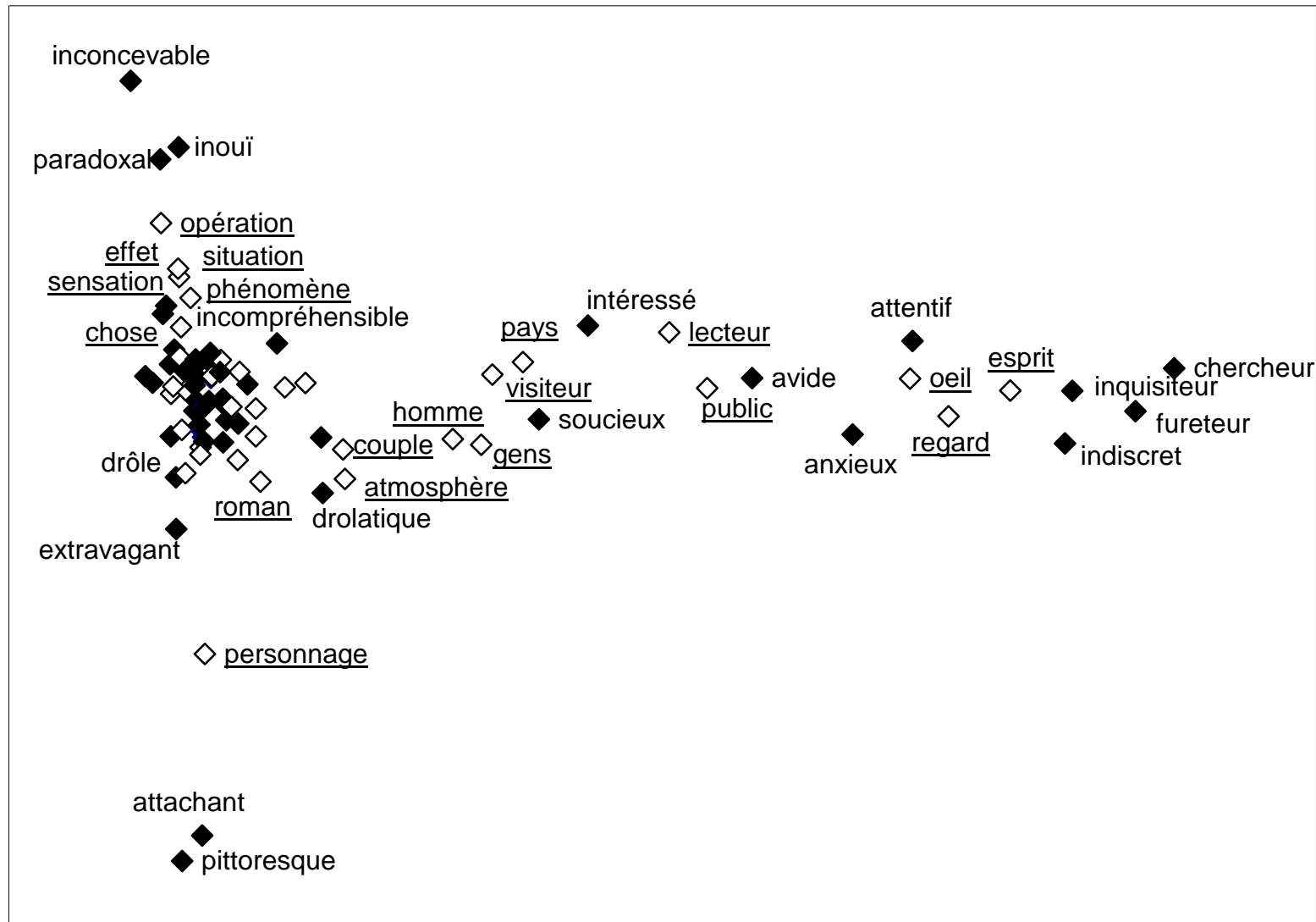


Figure 2 : espace sémantique obtenu avec les synonymes et les cooccurrences dans le corpus des périodiques

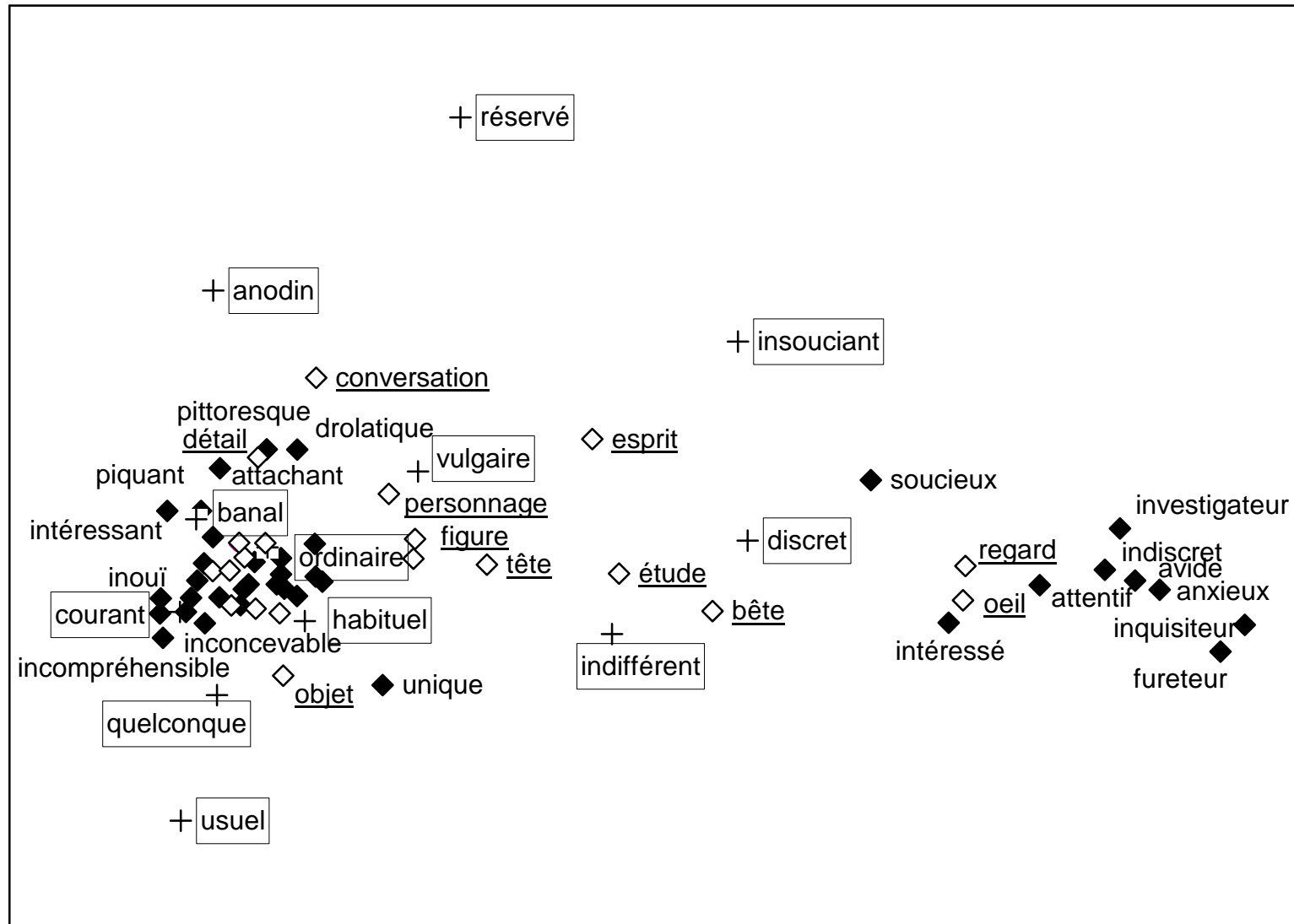


Figure 3 : espace sémantique obtenu avec les synonymes, les antonymes et les cooccurrences dans le corpus Frantext

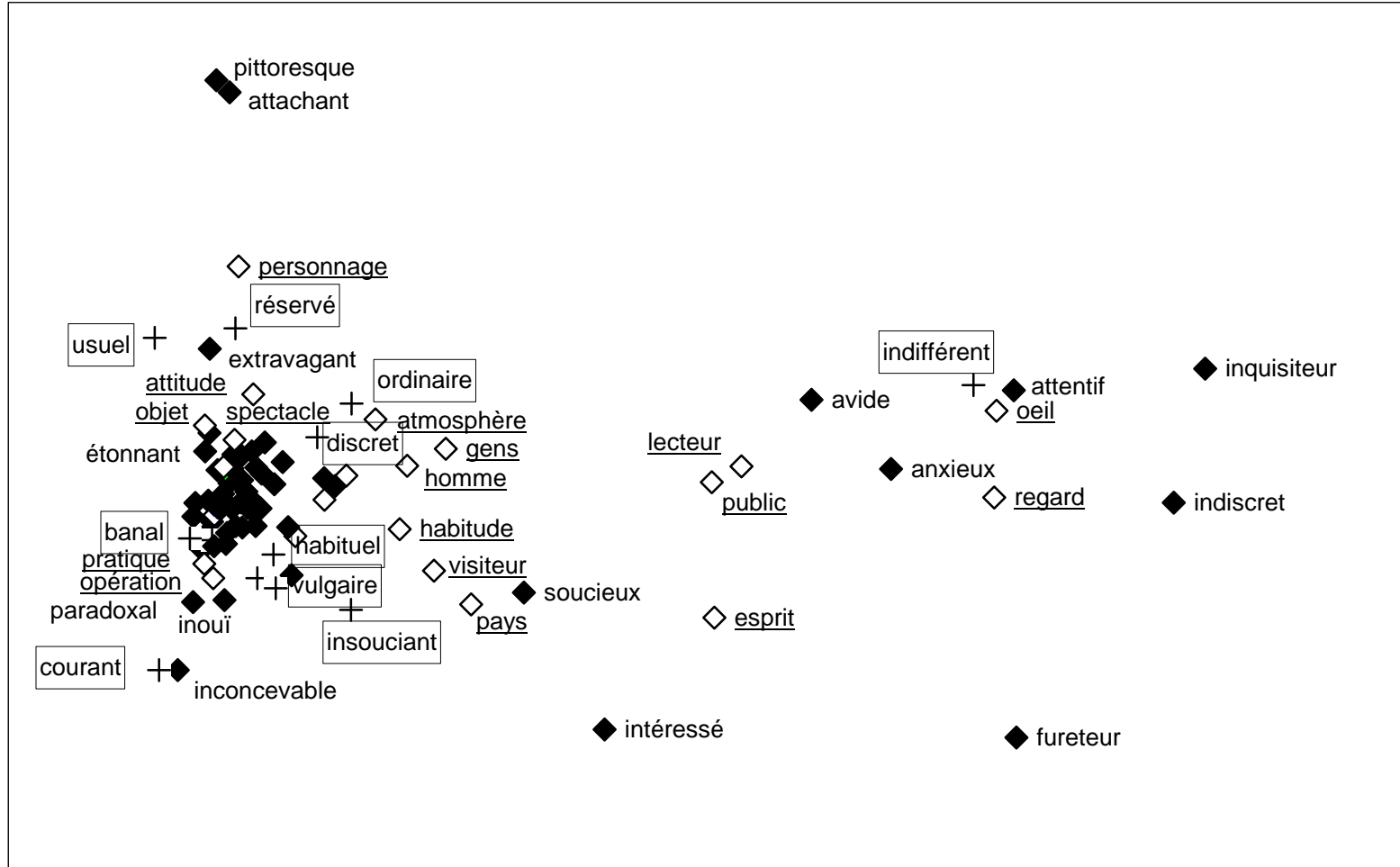


Figure 4 : espace sémantique obtenu avec les synonymes, antonymes et cooccurrences dans le corpus des périodiques