# Mise en évidence de rapports synonymiques par la méthode des cooccurrences.

William MARTINEZ

Université de la Sorbonne nouvelle Paris 3 - martinez@msh-paris.fr

**Abstract**

The purpose of this article is to demonstrate how the study of syntagmatic lexical relations through collocation may allow a more thorough study of word meaning by way of a more elaborate technique for disambiguation. Collocations which can be described as sequences of lexical items that habitually occur together were considered by Firth [1957] as elements of lexical meaning on the syntagmatic level. While the paradigmatic axis comprises words of the same class which can be substituted for one another in a given context, the syntagmatic axis represents a word's ability to combine with other words inside a sentence. As the intensity of these lexical relations can be measured statistically a word may be characterized by its collocational features. By comparing their lexical environments it is then possible to associate as synonyms those words which share similar characteristics.

**Résumé**

L'article suivant montre comment grâce à une application particulière de la méthode des cooccurrences il est possible d'étudier les liens syntagmatiques entre formes lexicales en vue de la désambiguïsation de ces dernières. L'approche du texte est double, au niveau syntagmatique pour en étudier les combinaisons lexicales et au niveau paradigmatique pour y discerner des commutations possibles entre formes. Si l'on considère les cooccurrences comme des associations lexicales habituelles et spécifiques à chaque forme, on peut alors caractériser un mot à partir de celles-ci. En réunissant en classes de vocabulaire les formes à caractéristiques cooccurrentielles communes, on obtient les éléments pour une interprétation précise des formes lexicales en mettant en évidence même certains rapports synonymiques entre ces formes.

**Mots clefs :** lexicométrie, cooccurrence, axe paradigmatique, axe syntagmatique, synonymie.

## Introduction

L'analyse des cooccurrences connaît actuellement des développements visant à repérer, au delà des associations caractéristiques de mots dans un corpus, des réseaux de récurrences de type sémantique. En dévoilant la dimension collocative du langage les approches cooccurrentielles procurent une mesure statistique des attirances et des répulsions que les formes d'un texte peuvent exercer entre elles. Dans des domaines aussi variés que la lexicographie, le traitement du langage naturel ou encore l'enseignement des langues, l'exploitation des collocations par des approches linguistiques et statistiques[1] permet la désambiguïsation de formes en puisant dans leur entourage contextuel les éléments de cette désambiguïsation. Notre travail est consacré à l'étude distributionnelle des formes dans un discours politique[2] et à leurs particularités collocationnelles afin de regrouper en classes

1 Cf. Church & Hanks [1993], Biber [1993].

2 Le corpus est constitué de 300 allocutions parlementaires du ministre des affaires étrangères français sur une période de 11 ans (1986 à 1996) et compte 442 707 occurrences pour 16 585 formes. Quatre ministres se succèdent dans cette fonction, J.- B. Raimond, R. Dumas, A. Juppé et H. de Charette.

sémantiquement homogènes celles qui partagent des caractéristiques cooccurrentielles communes. Sur le plan linguistique, notre méthode exploite l'articulation double de l'énoncé et s'inscrit dans une perspective d'analyse structurale des langues. Jakobson [1963] pose que l'interprétation de toute unité linguistique met en œuvre deux mécanismes intellectuels[3] indépendants. Une unité lexicale s'interprète en rapport avec les unités coexistantes sur l'axe syntagmatique et avec les unités semblables sur l'axe paradigmatique. Chaque occurrence d'un mot se situant à la croisée de ces deux axes qui représentent pour l'un la combinaison lexicale et pour l'autre le choix lexical, il se caractérise par un ensemble de traits particuliers. Une exploitation double, du lexique environnant et du lexique latent, pour permettre la définition d'une forme appelle une méthode d'appréhension du texte qui soit double elle aussi. La méthode de classement présentée ici exploite cette combinaison et notamment les commutations possibles entre éléments lexicaux. Nous montrerons comment, à partir de leurs caractéristiques cooccurrentielles et de leur distribution syntagmatique, on peut regrouper les formes en classes paradigmatiques par un traitement automatisé.

## Méthode

La première étape de notre méthode explore le syntagme et applique le principe de la 'signification par collocation' de Firth [1957] en exploitant le milieu ambiant d'une forme en vue de son interprétation. Recourant au modèle hypergéométrique[4] on relève les associations lexicales qui s'opèrent autour des formes d'un corpus. Le tableau 1 présente les résultats du calcul[5] pour la forme pôle *mesures* (fréquence=227) dans différents espaces de rencontre[6]. Ces cooccurrents permettent une description précise de l'univers lexical du pôle. Le choix de la phrase comme unité d'exploration contextuelle fait apparaître parmi les principaux cooccurrents du pôle des formes syntaxiquement et sémantiquement proches de *mesures* : adjectifs (*nouvelles*, *unilatérales*, *concrètes*, *nécessaires*, *sensibles*) et verbes (*prises*, *prendre*, *adopter*, *décidé*).

La deuxième étape de notre méthode s'appuie sur ces résultats pour rechercher d'autres formes dans le corpus partageant certaines de ces caractéristiques collocationnelles. En

---

3 Pour Jeandillou [1997] l'étymologie du mot *texte* qui dérive du verbe *texere* qui signifie *tisser*, doit nous amener à considérer 'le texte comme un entrelacs structuré de mots et de phrases, où se croisent une chaîne (syntagme) et une trame (paradigme). Il tire sa cohésion de la concaténation de ses composantes, et de leur correspondance in absentia.' Pour F. de Saussure [1916] chaque terme, suivant une combinatoire linéaire dans le syntagme, 'n'acquiert sa valeur que parce qu'il est opposé à ce qui précède ou ce qui suit, où à tous les deux'. Alors que ce rapport syntagmatique repose sur des termes in presentia, le rapport paradigmatique permet de rattacher l'énoncé à un système latent de la langue, un fonds où l'on puise 'des termes in absentia dans une série mnémonique virtuelle'.

4 Cf. Lafon [1984]. Une comparaison s'effectue entre l'ensemble du corpus (*T*) et l'échantillon des contextes contenant la forme pôle (*t*). En fonction de la fréquence totale d'une forme (*F*) et de sa fréquence locale (*f*), on affecte un indice de spécificité (*sp*) au cooccurrent. Le modèle fournit un diagnostic sous la forme *+Ex* où le signe indique s'il y a sur-emploi ou sous-emploi de la forme et la valeur indique son degré de spécificité.

5 Ces résultats sont obtenus à partir d'un module de cooccurrences développé autour du programme LEXICO2 avec C. Lamalle et A. Salem.

6 Couramment on choisit la phrase comme fenêtre d'exploration (ex : Lafon [1984]) car elle représente une expression de l' organisation syntagmatique facilement identifiable par l' homme et par la machine à partir des ponctuations fortes. Parfois on opte pour une fenêtre déterminée à la fois par une ponctuation et par un nombre de mots maximal pour chaque fenêtre (ex : Reinert [1993]). Chaque approche a ses avantages et ses inconvénients. Tantôt on saisit un syntagme intègre mais de taille très variable, tantôt on extrait des tronçons de taille fixe mais en interrompant l'expansion syntagmatique. Ce choix influe considérablement sur les résultats : l'exploration d'un contexte réduit (fenêtre de 10 mots) révèle des formes liées syntaxiquement au pôle tandis que l'examen de contextes plus larges (phrase ou paragraphe) met à jour des liens sémantiques.

| forme | F | fen 10 | fen 20 | phrase | paragraphe |
|---|---|---|---|---|---|
| prises | 119 | 22 +E27 | 25 +E27 | 27 +E20 | 30 +E15 |
| nouvelles | 180 | 21 +E22 | 23 +E19 | 27 +E15 | 29 +E10 |
| prendre | 178 | 21 +E22 | 24 +E21 | 26 +E14 | 31 +E11 |
| des | 7529 | 98 +E16 | 136 +E14 | 254 +E12 | 438 +E12 |
| unilatérales | 16 | 8 +E15 | 8 +E13 | 9 +E12 | 10 +E10 |
| confiance | 50 | 11 +E15 | 11 +E13 | 12 +E10 | 12 +E07 |
| adopter | 36 | 8 +E11 | 9 +E11 | 10 +E09 | 11 +E07 |
| enseignants | 14 | 0 1 | 6 +E07 | 9 +E10 | 14 |
| protection | 81 | 6 +E06 | 6 +E05 | 10 +E06 | 12 +E04 |
| adoption | 55 | 6 +E07 | 8 +E08 | 9 +E06 | 11 +E05 |
| mf | 40 | 3 +E03 | 3 +E03 | 8 +E06 | 7 +E03 |
| concrètes | 35 | 7 +E10 | 7 +E08 | 7 +E06 | 7 +E04 |
| les | 7942 | 84 +E09 | 122 +E08 | 226 +E05 | 419 +E06 |
| décidé | 169 | 3 | 5 +E02 | 13 +E05 | 17 +E04 |
| faveur | 149 | 7 +E05 | 7 +E04 | 12 +E05 | 17 +E04 |
| résidant | 12 | 0 | 1 | 4 +E05 | 5 +E05 |
| de | 23330 | 160 +E04 | 262 +E05 | 571 +E04 | 1008 |
| nécessaires | 117 | 7 +E06 | 8 +E05 | 10 +E04 | 13 +E03 |
| actions | 99 | 0 | 2 | 9 +E04 | 12 +E04 |
| économie | 69 | 0 | 1 | 7 +E04 | 7 +E02 |
| produits | 44 | 1 | 1 | 6 +E04 | 8 +E04 |
| sensibles | 31 | 0 | 1 | 5 +E04 | 6 +E03 |
| série | 28 | 5 +E07 | 5 +E06 | 5 +E04 | 5 +E03 |
| accroissement | 24 | 0 | 1 | 5 +E04 | 4 +E02 |
| conventionnelle | 17 | 1 | 4 +E05 | 4 +E04 | 4 +E03 |

***Guide de lecture du tableau 1*** : Le principal choix méthodologique pour le calcul des cooccurrences concerne l'étendue de la zone à explorer autour de la forme pôle. Notre méthode détermine suivant différentes unités contextuelles (fenêtre de 10 ou 20 mots autour du pôle, phrase ou paragraphe) quels sont les cooccurrents spécifiques d'un pôle. Pour chacun on trouve sa fréquence totale dans le corpus (*F*), sa fréquence locale (*f*) dans la fenêtre explorée et un indice de spécificité. Les résultats, triés sur la colonne 'phrase', montrent par exemple que dans l'ensemble des phrases où apparaît *mesures*, on rencontre exceptionnellement 27 fois la forme *prises*.

*Tableau 1 : Cooccurrents spécifiques de la forme* ***mesures****.*

limitant par exemple la recherche à ses deux cooccurrents les plus spécifiques, *nouvelles* et *prises*, on retrouve dans le corpus une manifestation de ces cooccurrents indépendante du pôle *mesures* (figure 1). Considérons maintenant ces deux formes comme un pôle unique dont on étudie les associations lexicales : c'est le cooccurrent *décisions* qui émerge du nouveau calcul. On note que ce nom commun, que qualifie l'adjectif *nouvelle* et auquel se rapporte également le participe *prises*, appartient au même champ sémantique que le pôle d'origine et occupe dans ce contexte indépendant une position centrale autour de laquelle gravite une partie du système lexical mis en évidence autour de *mesures*. Ces associations et ce mode de fonctionnement communs laissent apparaître un possible lien de synonymie entre les deux pôles *mesures* et *décisions*.

Dans cet exemple une lecture cursive a suffi pour isoler un nouveau pôle et la recherche motivée d'un synonyme a permis de repérer facilement un candidat satisfaisant dans un contexte limité. Il s'agit d'un premier essai servant à tester le bien-fondé d'une méthode qu'il faut maintenant formaliser afin d'appliquer la procédure au corpus tout entier en automatisant la recherche de synonymes pour chaque forme.

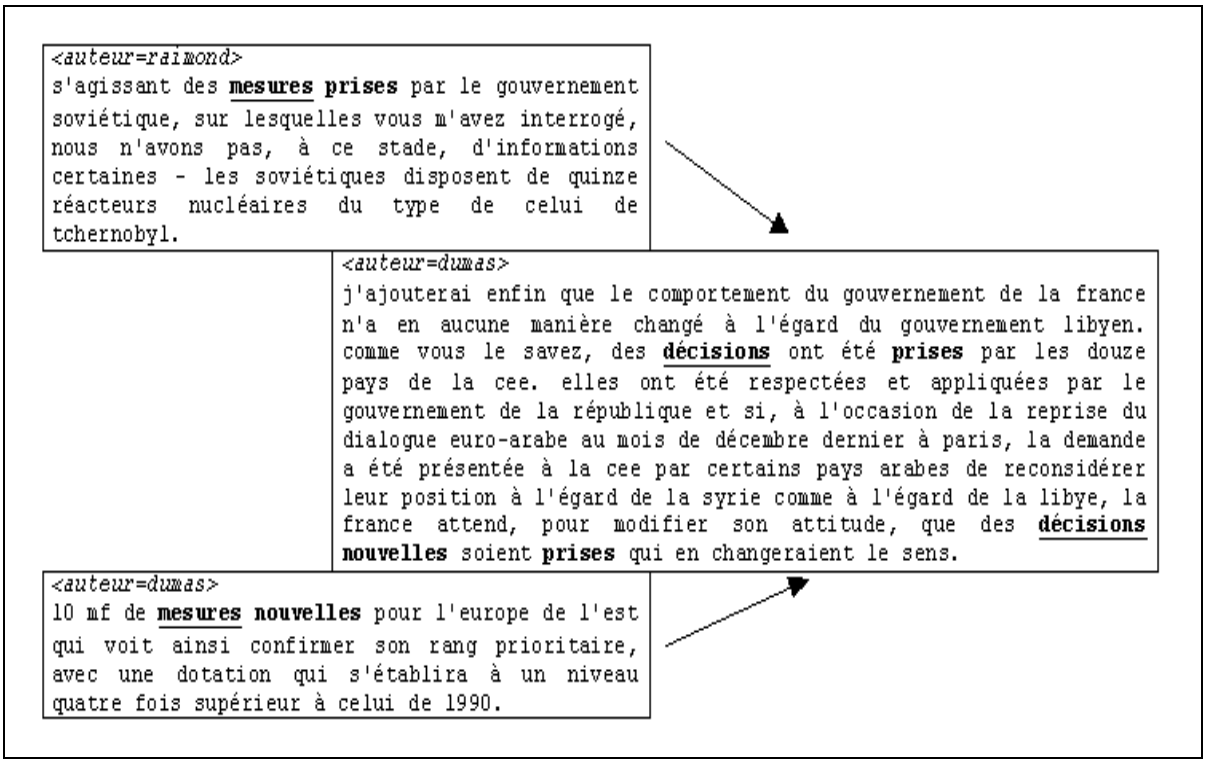



*Figure 1 : Recherche de caractéristiques distributionnelles communes entre formes.*

## Formalisation

Le principe fondamental de cette méthode de recherche de synonymes est la réitération du calcul des cooccurrences. On a vu dans l'exemple précédent que la répétition du comptage s'effectue à partir d'une sélection des résultats obtenus au cours du premier calcul. L'objectif de cette sélection est l'homogénéisation de la liste afin d'obtenir un ensemble caractérisant au mieux le pôle. En effet, les cooccurrents sur lesquels se base le second calcul doivent être choisis parmi les plus spécifiques du pôle initial afin d'éviter un élargissement exagéré du contexte[7] étudié qui produirait des rapprochements lexicaux peu pertinents. Deux filtrages sont donc effectués : un filtrage sur la fréquence pour ne retenir que les dix cooccurrents les plus spécifiques et un filtrage sur les mots outils. La grande variabilité de la phrase en tant qu'unité contextuelle pose en soi un problème d'interprétation auquel il faut ajouter celui de la versatilité des articles *des*, *les* et *de* et autres formes de liaison sans véritable référent qui rendent difficile l'appréciation de leurs rapports au pôle. Après filtrage on considère la liste réduite de cooccurrents spécifiques que nous appellerons 'cooccurrents de 1er niveau' comme un pôle unique dont on étudie à son tour les associations lexicales. A partir de cette transformation en une forme générique[8] (opération schématisée dans le tableau 2) peut s'effectuer la deuxième phase de notre méthode où l'on applique le calcul des cooccurrences au sein des phrases qui contiennent les cooccurrents du pôle sans contenir le pôle lui-même.

Les dix cooccurrents du tableau 1 ont été rassemblés du fait de leur associations avec le pôle *mesures*. Toutefois, plutôt que de les caractériser ainsi, on pourrait adopter une perspective

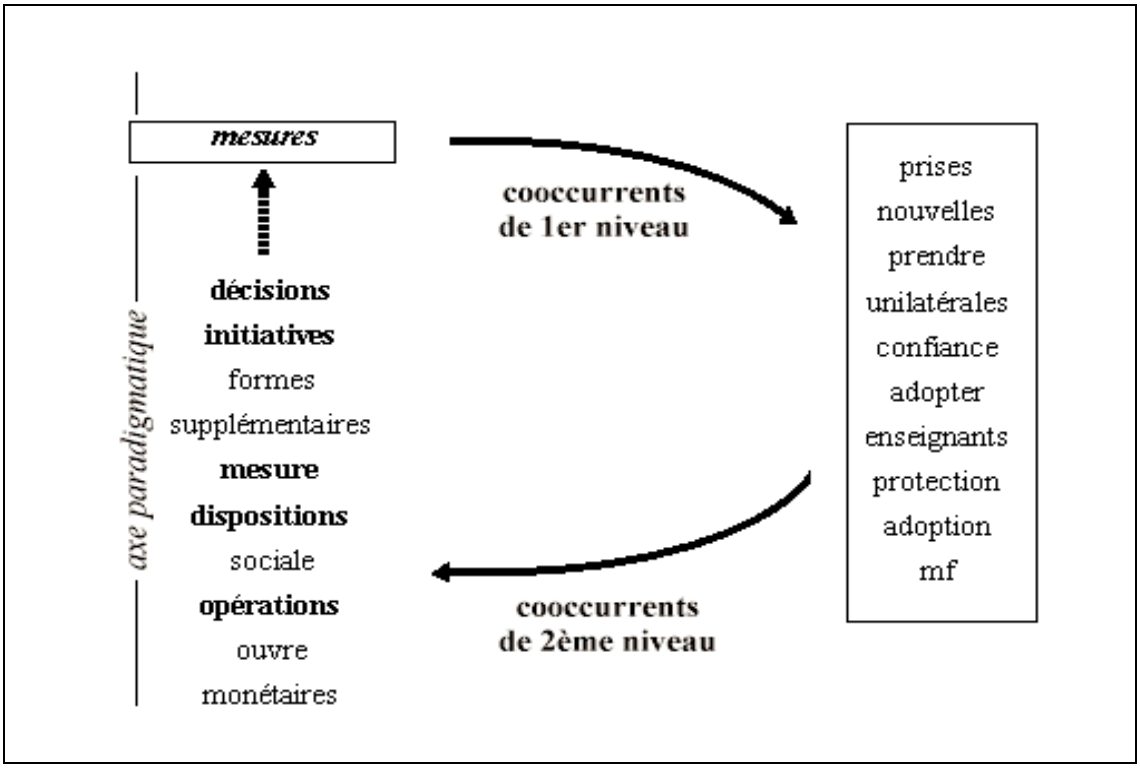
---

7 Pour donner un ordre d'idée, dans notre exemple les contextes du pôle *mesures* représentent 9 288 occurrences tandis que les contextes de ses dix premiers cooccurrents totalisent 30 106 occurrences.

8 La forme générique hérite aussi de la nature grammaticale de chacun des cooccurrents de 1er niveau et occupe les mêmes positions sur l'axe syntagmatique privilégiant ainsi un certain type d'association lexicale.

| Forme pôle | | --> Cooccurrents 1er niveau | | --> Forme générique | |
|---|---|---|---|---|---|
| | F | | F | | F |
| | | prises | 119 | | |
| | | nouvelles | 180 | | |
| | | prendre | 178 | | |
| | | unilatérales | 16 | | |
| mesures | 227 | confiance | 50 | forme générique | 769 |
| | | adopter | 36 | | |
| | | enseignants | 14 | | |
| | | protection | 81 | | |
| | | adoption | 55 | | |
| | | mf | 40 | | |

*Tableau 2 : Cooccurrents et forme générique.*




***Guide de lecture de la figure 2*** : La mise en évidence de liens synonymiques par la méthode des cooccurrences compte deux étapes. Au cours de la première, l'exploration autour du pôle *mesures* au sein de l'unité contextuelle de la phrase révèle une première série de 10 cooccurrents. La seconde phase consiste alors à rechercher les phrases où une ou plusieurs de ces 10 formes apparaissent sans le pôle initial *mesures*. Au nouvel échantillon de texte ainsi défini on applique une deuxième fois le calcul des cooccurrences qui rend compte de phénomènes cooccurrentiels autour des premières collocations. Cette réitération du calcul qui revient donc à chercher les cooccurrents des cooccurrents d'un pôle donné permet de discerner par leurs similarités distributionnelles certaines formes synonymes dans le corpus.

*Figure 2 : Mise en évidence de liens synonymiques par la méthode des cooccurrences.*

différente et affirmer que c'est la forme *mesures* qui se déduit de l'ensemble des dix cooccurrents. Suivant cette perspective, recherchons si d'autres formes partagent une partie de ces signes distinctifs au sein du corpus. Existe-t-il des manifestations de cette classe de dix formes indépendamment du pôle qui a contribué à créer cette même classe ?

La figure 2 présente les résultat du calcul des cooccurrents de *mesures* au 1er et au 2ème niveau[9]. La première liste filtrée montre les dix cooccurrents les plus spécifiques de *mesures* qui lorsqu'ils évoluent indépendamment du pôle attirent dans leur proche voisinage un nombre de cooccurrents spécifiques qui figurent dans la seconde liste. On trouve parmi les cooccurrents des cooccurrents notamment six noms : *décisions*, *initiatives*, *formes*, *mesure*, *dispositions* et *opérations*. La majorité de ces cooccurrents appartiennent donc à la même catégorie grammaticale que le pôle d'origine. On s'aperçoit que la méthode du calcul récursif des cooccurrences met en évidence une relation sémantique entre une forme pôle de départ et un ensemble de formes qui ne sont pas nécessairement cooccurrentes entre-elles.

L'extension de cette méthode à l'ensemble des formes du corpus permet la mise en évidence d'autres rapports de synonymie comme ceux présentés dans les figures suivantes pour d'autres pôles. Ainsi, à partir de la forme *cadre* (figure 3) le calcul des cooccurrents communs a rapproché les formes *domaine*, *domaines*, *contexte* et *perspective*. Au verbe *souhaite* (figure 4) la méthode associe, entre autres, les formes *voudrais*, *veux* et *tiens*.

Ces exemples montrent que suivant une norme endogène notre méthode s'appuie entièrement sur les caractéristiques du corpus pour déterminer automatiquement les associations lexicales. Cette technique de rapprochement des formes par leurs caractéristiques cooccurrentielles communes nous semble dépasser les précédentes méthodes d'analyse de la signification par collocation pour ouvrir la voie à une désambiguïsation par collocation plus élaborée et plus précise.

## Références

Biber D. (1993). Co-occurrence Patterns among Collocations : a Tool for Corpus-Based Lexical Knowledge Acquisition. *Computational Linguistics*, vol. 19 n°3.

Church K. et Hanks P. (1993). Word Association Norms, Mutual Information and Lexicography. *Proceedings 13th International Conference on Computational Linguistics*.

De Saussure F. (1916). *Cours de linguistique générale*. Payot.

Firth J. (1957). A Synopsis of Linguistic Theory 1930-1955. *Studies in Linguistic Analysis*. Philological Society.

Jakobson R. (1963). *Essais de linguistique générale*. Seuil.

Jeandillou J. (1997). *L'analyse textuelle*. Armand Collin.

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

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

Reinert M. (1990). Alceste, une méthodologie d'analyse des données textuelles et une application : Aurélia de Gérard de Nerval. *Bulletin de méthodologie sociologie*, n°26.

9 On note que la forme *des*, mot outil, a été écartée de cette liste afin d'éviter l'exploration de contextes peu pertinents.

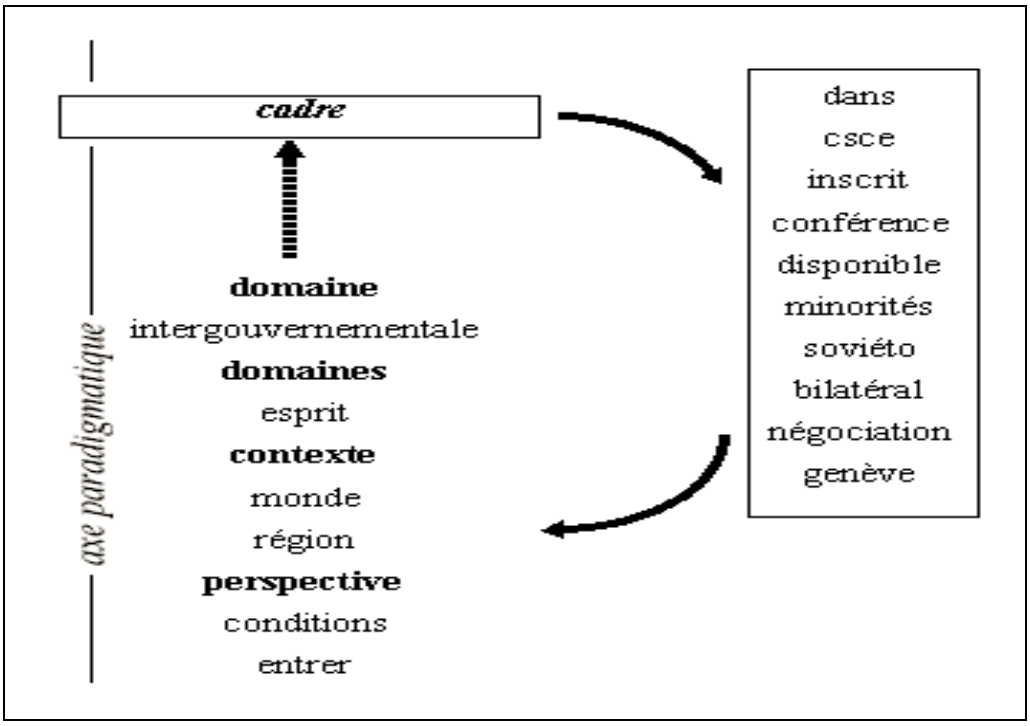



*Figure 3 : Mise en évidence de liens synonymiques autour de **cadre**.*

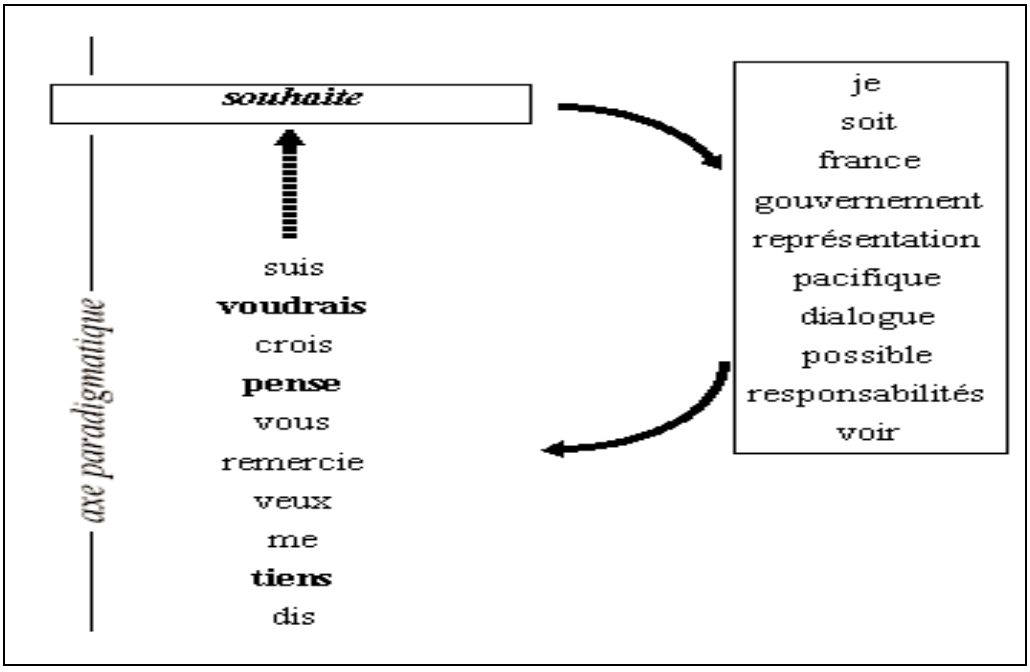



*Figure 4 : Mise en évidence de liens synonymiques autour de **souhaite**.*