Représentation géométrique d'un paradigme lexical

Bernard Victorri, Jean-Luc Manguin

▶ **To cite this version:**

Bernard Victorri, Jean-Luc Manguin. Représentation géométrique d'un paradigme lexical. Conférence TALN 1999, Jul 1999, Cargese (Corse), France. hal-04520029

HAL Id: hal-04520029
https://hal.science/hal-04520029v1

Submitted on 25 Mar 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial - ShareAlike 4.0 International License

Conférence TALN 1999, Cargèse, 12-17 juillet 1999

Représentation géométrique d'un paradigme lexical.

Jean-Luc Manguin, Bernard Victorri

Laboratoire ELSAP
CNRS – Université de Caen
Esplanade de la Paix
14032 Caen Cedex
manguin@elsap.unicaen.fr
http://www.elsap.unicaen.fr/

Laboratoire LTM
CNRS – ENS Montrouge
1 , rue Maurice Arnoux
92120 Montrouge
Bernard.Victorri@ens.fr
http://www.ltm.ens.fr

1. Principes et méthode

L'étude des paradigmes lexicaux est d'un grand intérêt pour la compréhension des relations existant entre des termes d'un même champ sémantique ; ces relations peuvent s'exprimer en faisant appel à des notions de hiérarchie, de proximité, et même d'agencement topologique. Un certain nombre de travaux (cf. entre autres Hindle D. 1990, Grefenstette G. 1994, Veronis J. & Ide N. 1990, Warnesson I. 1992 - pour une présentation générale, cf. Habert B. *et al*. 1997) ont cherché à obtenir ces relations de manière automatique en s'appuyant soit sur des occurrences des termes lexicaux dans des corpus, soit directement sur des dictionnaires informatisés de synonymes. Le problème essentiel auquel se heurtent ces différentes approches est celui de la polysémie. En effet chaque unité lexicale pouvant prendre une pluralité de sens, les relations entre unités ne peuvent pas être décrites de manière simple : deux unités peuvent à la fois être très proches dans certains de leurs emplois, et très opposées dans d'autres contextes. Par exemple, les verbes *mettre* et *prendre* sont presque synonymes dans des énoncés comme *mettre son chapeau*, *prendre son chapeau*, et presque antonymes dans *mettre de l'argent*, *prendre de l'argent*. Généralement, on cherche à résoudre ce problème en " éclatant " chaque unité lexicale en autant de sous-unités qu'elle peut avoir de sens, ce qui a l'inconvénient de perdre ce que ces différents sens ont en commun.

L'originalité de notre travail consiste à respecter l'intégrité des unités lexicales, en associant à chacune une *région* d'un espace sémantique partagé par tous les termes du paradigme étudié. Ainsi deux unités peuvent se recouvrir sur une partie de l'espace sémantique tout en restant distinctes sur d'autres zones de l'espace. Dans cette approche géométrique, les relations entre unités sont donc représentées par des relations entre régions, ce qui permet de les caractériser précisément sans pour autant faire l'impasse sur le phénomène de la polysémie.

Nous présentons ici une méthode pour construire de telles représentations à partir de dictionnaires informatisés de synonymes. Cette méthode sera illustrée par l'étude d'un paradigme lexical de 16 verbes pouvant évoquer un déplacement d'objet : *apporter, descendre, déplacer, emporter, envoyer, jeter, lancer, lever, mettre, monter, porter, poser, prendre, soulever, tirer, traîner*. Nous avons choisi ce paradigme parce qu'il a fait l'objet par ailleurs d'une expérimentation par des méthodes de psychologie cognitive (cf. Marquant-Thiébaut, 1999) : nous visons, à terme, à comparer ces deux études pour évaluer la pertinence cognitive des relations mises en évidence par une analyse purement linguistique.

Notre travail se fonde sur une technique mise au point par Sabine Ploux (cf. Ploux S., 1998 ; Ploux S. et Victorri B., 1998). Faute de place, nous nous contenterons d'indiquer que cette

technique repose sur une analyse du graphe de la relation de synonymie fournie par la fusion de plusieurs dictionnaires de synonymes symétrisés : chaque "clique" (sous-graphe complet maximal) du graphe représente par hypothèse un "sens élémentaire" qui est associé à un point de l'espace sémantique en construction, sur lequel on définit une métrique (de type χ^2). A chaque unité est alors associée la région de l'espace englobant toutes les cliques qui contiennent l'unité en question.

2. Analyse globale du paradigme lexical

Nous pouvons tout d'abord remarquer que certains verbes du paradigme contiennent énormément de synonymes, engendrant une quantité impressionnante de cliques ; ceci reflète leur polysémie, et les verbes les plus significatifs à cet égard sont *prendre* (217 synonymes, 486 cliques), *porter* (117 synonymes, 201 cliques), *mettre* (109 synonymes, 207 cliques) et *monter* (105 synonymes, 142 cliques). Ces chiffres sont à comparer avec les moyennes de notre dictionnaire : le nombre moyen de synonymes d'une unité est de l'ordre de 10, et le nombre moyen de cliques entre 15 et 20. De plus, le nombre total de cliques engendrées à partir du paradigme est bien supérieur à la somme des cliques de chacun des termes (3233 cliques pour 783 synonymes) ; ceci parce qu'il existe entre les synonymes de ce groupe des relations indépendantes de celles qui les lient avec l'un des termes du paradigme.

Voici, à titre d'exemple, quelques-unes des cliques que l'on obtient pour notre paradigme :

1. *abattre, descendre, exécuter, tuer*
2. *abandonner, jeter, rejeter, repousser*
3. *abandonner, déposer, poser, quitter*
4. *affecter, jouer, poser, simuler*
5. *affirmer, alléguer, avancer, poser, énoncer*
6. *caser, disposer, mettre, placer, ranger, situer*
7. *demeurer, descendre, loger, résider, séjourner*
8. *destituer, démettre, déplacer, détrôner*
9. *différer, lanterner, retarder, tarder, traîner*
10. *dresser, hausser, lever, monter, relever, soulever, élever*

Comme on peut le constater, chaque clique représente un sens très précis, qui permet de prendre en compte des nuances assez fines (cf. par exemple la différence entre les cliques 2 et 3) et aussi de bien différencier les sens d'une même unité (cf. les cliques 3, 4 et 5, qui représentent chacune un sens différent de *poser*).

Comme on pouvait s'y attendre, l'ensemble des sens révélés par l'analyse des cliques déborde donc très largement des sens associés à un déplacement d'objet. En fait, l'analyse en composantes principales nous montre que l'espace sémantique global n'est pas du tout structuré par les sens de déplacement d'objet mais par des dimensions sémantiques très générales, qui correspondent à des distinctions de sens pouvant servir à discriminer l'ensemble des sens que peuvent prendre des verbes en français, et que l'on peut traduire par des questions du type : le verbe évoque-t-il une action ou non, l'action est-elle détrimentale pour le sujet ou pour l'objet, etc.

Par exemple, les axes 1 et 2 (cf. figure 1) peuvent être qualifiés comme suit :

axe 1 : action - inaction (la partie positive contient les sens de *traîner* correspondant à l'absence d'action)
axe 2 : capture - exposition (du sujet)

Représentation géométrique d’un paradigme lexical

Après quelques restrictions sur les données et un choix judicieux d'axes dans le sous-espace engendré par les axes 2, 3 et 5, on obtient le plan d'observation de la figure 2, qui résume assez bien les oppositions entre quatre groupes : (1) *prendre, emporter*, (2) *jeter, lancer, envoyer*, (3) *monter, lever, soulever* (au sens de soulèvement, excitation) et (4) *descendre* (chute, déclin). Un autre groupe, constitué par *poser, mettre* (fixation, arrangement) peut être mis en évidence sur une autre dimension de l'espace sémantique. Enfin, *apporter* et *porter* correspondent à des régions centrales, peu différenciées, tandis que *tirer* s’étend sur deux zones, correspondant respectivement à des sens proches de *lancer* (*tirer une flèche*) et de *prendre* (*tirer son mouchoir de sa poche*).

Ainsi, l’analyse du paradigme lexical constitué par ces 16 verbes montre que leur fonctionnement sémantique général ne saurait se déduire de leur comportement dans le champ restreint du déplacement d’objet. Au contraire, leur sens de déplacement d’objet semble ne jouer qu’un rôle secondaire dans leur sémantisme. C’est particulièrement net pour trois d’entre eux : *traîner*, principalement associé à des sens “ d’inaction ”, d’un sujet replié sur lui-même ; *descendre*, pour lequel l’emportent les sens liés à une action détrimentale ; et enfin, de manière plus surprenante *déplacer*, souvent utilisé comme hyperonyme pour ce paradigme, et qui est ici caractérisé plutôt par une action consistant à “ défaire ”, “ désorganiser ” un objet.

3. Analyse restreinte au champ sémantique du déplacement d’objet

A la suite de ces premiers résultats, nous avons essayé de voir si cette approche pouvait aussi permettre de structurer le champ restreint de déplacement d’objet, en cherchant en quelque sorte à faire un “ zoom ” sur la partie de l’espace sémantique global qui contenait les sens de déplacement d’objet.

La méthode retenue consiste à diminuer le nombre de cliques, en se limitant à celles qui contiennent au moins deux termes du paradigme. Le nombre de cliques s’en trouve considérablement réduit (228 au lieu de 3233) de même que le nombre de synonymes (179 au lieu de 783) ; par ailleurs, l’analyse des “ couples ” apparaissant dans les cliques met en évidence l’hétérogénéité du paradigme de départ, si l’on considère le nombre total de cliques que chaque terme forme avec les autres ; pour la plupart des verbes, ce nombre est compris entre 23 et 75, sauf pour *descendre* (6), *déplacer* (2) et *traîner* (6) ; cette constatation explique pourquoi ces trois unités forment sur le diagramme global de départ des “ branches ” qui s’éloignent fortement du noyau central des cliques.

Ceci nous amène à dire d’une part que le paradigme choisi à l’origine n’est pas “ homogène ” : tous les termes ne sont pas aussi “ proches ” les uns des autres (nous définirons plus tard cette “ proximité ”). D’autre part, ces données confirment les remarques formulées à l’égard de *traîner*, *descendre* et *déplacer* à la fin de notre première analyse.

La nouvelle étude avec les 13 verbes donne un espace sémantique qui, dans le cas des cliques contenant au moins 2 termes du paradigme, peut se structurer en attribuant aux axes principaux des sèmes ayant trait le plus souvent au déplacement d’objet ; par exemple (figure 3), le premier plan principal s’explique comme suit :

axe 1 : séparation *vs* prise de l’objet par le sujet
axe 2 : fixation *vs* expulsion de l’objet par le sujet

On peut donner également des interprétations du même ordre aux autres axes, par exemple :

axe 3 : développement *vs* contrainte de l’objet par le sujet
axe 4 : sujet fixe *vs* sujet en mouvement

Cependant, l'examen du nuage de points révèle la complexité de l'agencement des termes du paradigme ; autrement dit, leur représentation par des ellipses englobant les points qui leur sont liés ne reflète qu'imparfaitement la topologie de leur relation ; nous sommes ainsi amenés à définir un critère plus synthétique, à savoir une mesure de la proximité entre les nuages correspondant à chacun des termes. Cette mesure de proximité entre deux ensembles de points A et B peut se définir de la manière suivante :

$$\text{prox}(A,B) = \text{moyenne}_{a\in A}(\min_{b\in B}(\text{dist}(a,b))) + \text{moyenne}_{b\in B}\ (\min_{a\in A}(\text{dist}(a,b)))$$

L'exploitation de cette mesure nous permet alors de représenter graphiquement les distances sémantiques pour les termes du paradigme, pris soit individuellement, soit dans leur ensemble par un « graphe de proximité sémantique » (figure 4). Ce dernier type de représentation nous sera particulièrement utile pour comparer ces résultats à ceux de l'expérimentation psycholinguistique menée sur ces verbes par Marquant-Thiébaut (1999).

En conclusion, la méthode que nous proposons pour l'analyse automatique de paradigmes lexicaux permet d'obtenir deux types de résultats :

- d'une part, la mise en évidence de l'organisation du paradigme en soi, sans référence à un champ sémantique spécifique : les axes obtenus par l'analyse globale permettent de révéler des dimensions sémantiques générales, dont on peut soupçonner qu'elles sont communes à de nombreuses unités de la langue. Ainsi, en ce qui concernent les verbes, les dimensions que nous avons mises en évidence semblent être pertinentes bien au-delà du paradigme que nous avons présenté ici : un certain nombre d'autres études, que nous menons sur d'autres verbes (comme *jouer*, *compter*, *dire*, etc.) nous confortent dans cette idée. D'autres études systématiques sont bien sûr nécessaires pour valider cette hypothèse, et pour définir précisément ces dimensions.

- d'autre part, la mise en évidence de l'organisation d'un champ sémantique spécifique, obtenu par restriction de l'espace sémantique global à des sens plus étroits : en quelque sorte, les axes obtenus par ces restrictions peuvent être considérés comme des projections de ces dimensions sémantiques générales à un domaine particulier de l'expérience.

Références

GREFENSTETTE G. (1994) : *Explorations in Automatic Thesaurus Discovery*, Dordrecht, Kluwer.

HABERT B., NAZARENKO A., SALEM A. (1997) : *Les linguistiques de corpus*, Paris, Armand Colin.

HINDLE D. (1990) : " Noun classification from predicate-argument structures ", *ACL'83*, Berkeley, 268-275.

MARQUANT-THIÉBAUT M. (1999) : *L'organisation sémantique des verbes de déplacement d'objets chez l'enfant*, Caen, Colloque « Sémantique du lexique verbal ».

PLOUX S. (1998) : " Modélisation et traitement informatique de la synonymie ", *Linguisticæ Investigationnes*.

PLOUX S., VICTORRI B. (1998) : Construction d'espaces sémantiques à l'aide de dictionnaires de synonymes, *Traitement automatique des langues*, 39, n°1, pp.161-182.

VERONIS J., IDE N. (1990) : " Word sense disambiguation with very large neural networks extracted from machine-readable dictionaries ", *COLING'90*, Helsinki, 389-394.

WARNESSON I. (1992) : " Lexicographie et informatique, vers une nouvelle génération de dictionnaires ", *Publications scientifiques et techniques d'IBM France*, décembre 1992, Paris, 107-157.

Représentation géométrique d'un paradigme lexical

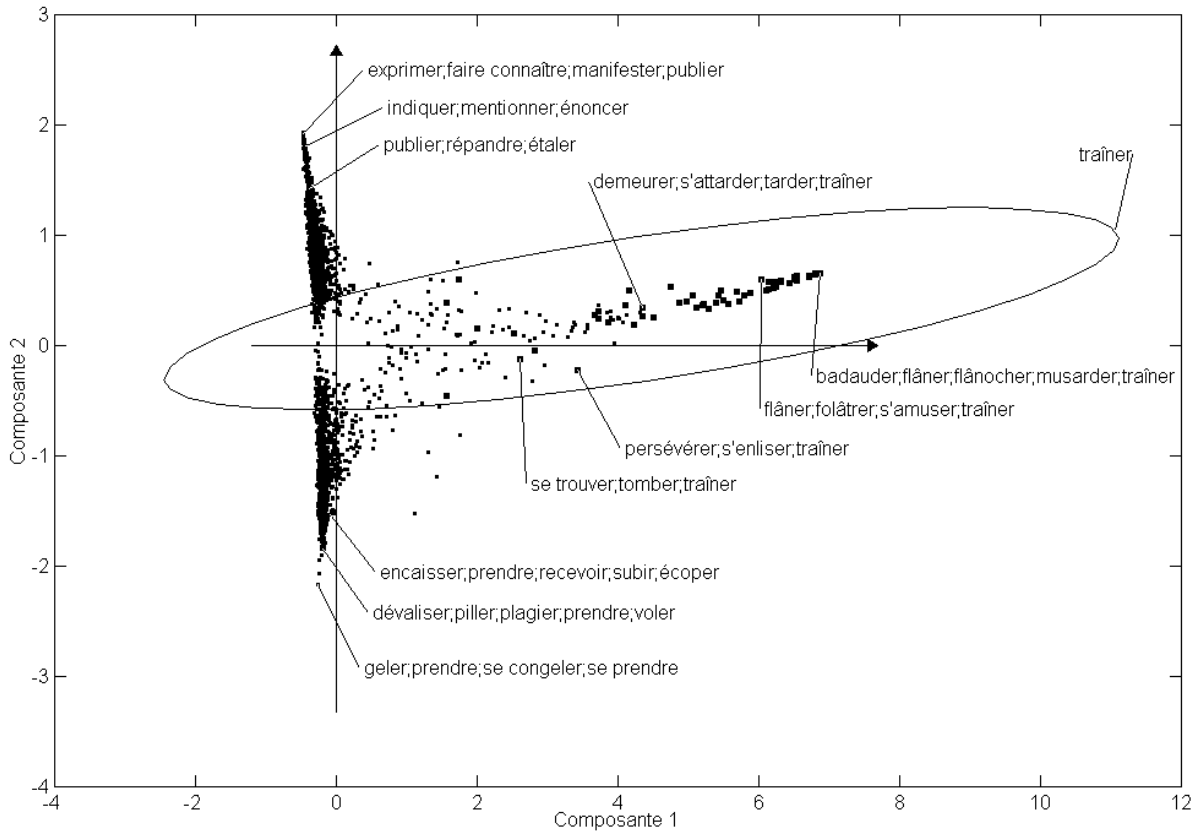

Figure 1 (Remarque : points = cliques ; ellipses = unités lexicales)

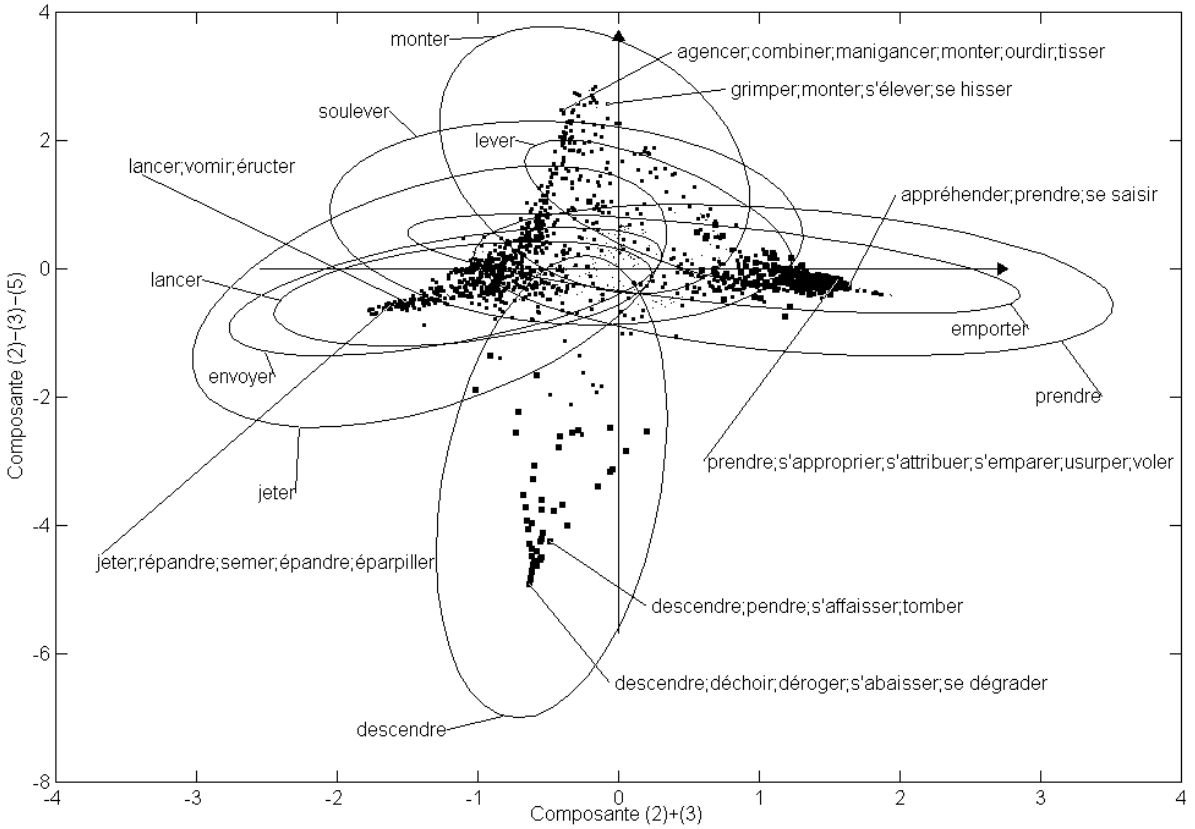

Figure 2

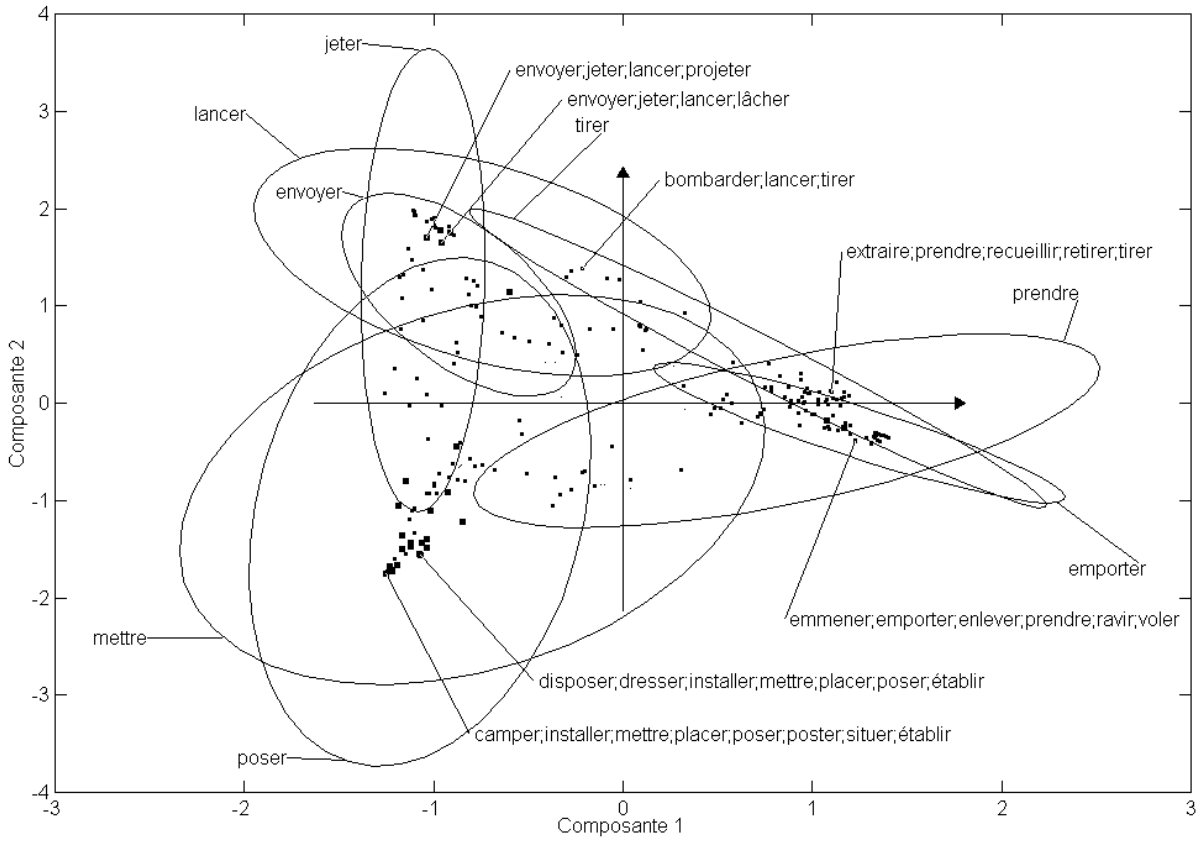

Figure 3

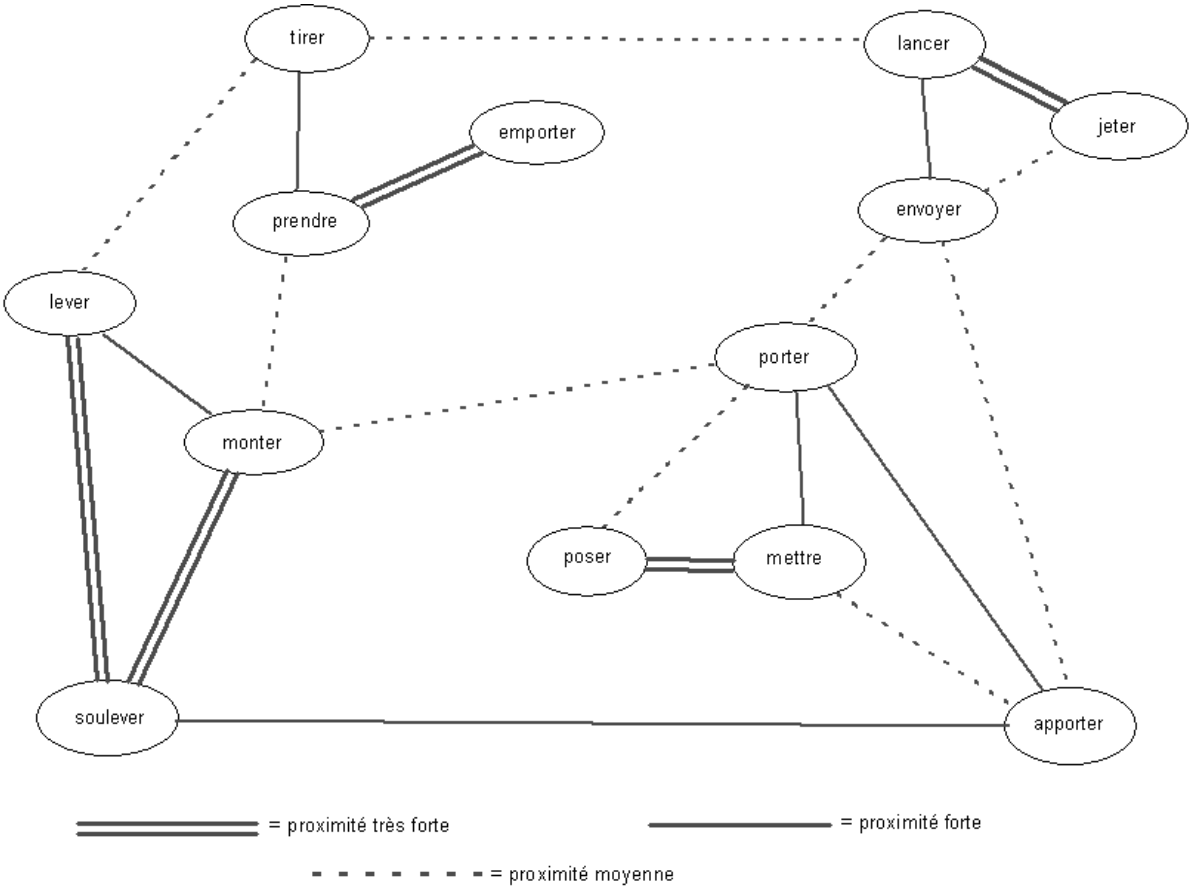

Figure 4