Quand les mots s'organisent en réseaux

Bernard Victorri

L'Archicube, 8, mai 2010, p.53-59. –Version préliminaire

Quand les mots s'organisent en réseaux

Réseaux de mots

Tout le monde a joué, un jour ou l'autre, à l'un de ces jeux où l'on doit bâtir une chaîne de mots, chaque mot étant relié au précédent par une règle donnée : on peut ainsi effectuer de longues promenades à travers le lexique d'une langue. L'aspect ludique de ces exercices provient de la diversité et de la densité des relations qu'entretiennent les mots entre eux, qui expliquent le côté imprévisible, surprenant et toujours inédit de ces parcours. Ces relations peuvent être de deux niveaux bien distincts : niveau phonologique, par exemple quand on passe d'un mot à un autre par des substitutions de phonèmes (*livre* → *vivre* → *vitre* → *votre* → *cotre* → …), et niveau sémantique, quand c'est par « association d'idées » que les mots sont reliés entre eux, soit parce qu'ils apparaissent souvent ensemble dans la parole (relation syntagmatique : *livre* → *lire* ou *livre* → *page*), soit au contraire parce qu'ils pourraient être mis l'un à la place de l'autre (relation paradigmatique : *livre* → *bouquin* ou *livre* → *journal*). Il faut noter que d'autres types de jeux, plus élaborés, s'appuient sur des relations bien plus arbitraires : c'est le cas du célèbre jeu oulipien où l'on remplace chaque mot lexical d'un texte par le septième mot de même catégorie grammaticale (nom, verbe ou adjectif) qui le suit dans le dictionnaire (cf. le poème de Raymond Queneau, *La cimaise et la fraction : La cimaise ayant chaponné tout l'éternueur / se tuba fort dépurative quand la bixacée fut verdie…*).

Ce sont aux relations sémantiques, et plus particulièrement à celles de type paradigmatique, que nous allons nous intéresser ici, car elles font l'objet de recherches assez intensives ces dernières années. Comme souvent dans les recherches contemporaines en linguistique, ce sont les nouvelles technologies qui sont à l'origine de cet engouement. Grâce à elles se sont constituées des bases de données lexicales recensant toute sorte de relations sémantiques dans de nombreuses langues. On peut ainsi citer pour l'anglais WordNet, pionnier en la matière, et pour le français le dictionnaire électronique des synonymes consultable sur le site Web de l'Université de Caen, dont nous allons reparler ci-dessous. Citons encore l'atlas sémantique de l'Institut des Sciences Cognitives, qui a l'intérêt de fournir un réseau sémantique lexical mixte français-anglais. Ces ressources informatiques auraient sans doute beaucoup intéressé

Saussure, puisque, comme on va le voir, elles permettent d'étudier la langue avec des méthodes purement structuralistes : le sens de chaque mot est défini par les relations qu'il entretient avec tous les autres mots du lexique, ce qui correspond bien à la conception différentielle de Saussure pour lequel « la valeur d'un mot ne sera jamais déterminée que par le concours des termes coexistants qui le limitent » et « dans la langue chaque terme a sa valeur par son opposition avec tous les autres termes ».

La relation de synonymie est particulièrement intéressante à cet égard. Précisons tout de suite qu'il s'agit ici de synonymie partielle : deux mots sont considérés comme synonymes si l'un peut remplacer l'autre dans un certain nombre de contextes (pas nécessairement tous) sans modification notable de sens. Pour ne prendre qu'un exemple, *défendre* et *interdire* sont synonymes parce qu'ils sont à peu près équivalents dans des énoncés tels que *défendre de fumer* et *interdire de fumer*. De même, *défendre* et *soutenir* sont synonymes parce que *défendre les droits de l'homme* et *soutenir les droits de l'homme* ont sensiblement le même sens. En revanche, *interdire* et *soutenir* ne sont pas synonymes car il n'existe pas de contexte dans lequel on puisse les permuter sans modifier considérablement le sens de l'énoncé qui subit la transformation. On en déduit que *défendre* n'a pas le même sens dans tous les énoncés puisqu'il est remplaçable tantôt par *interdire* et tantôt par *soutenir*, qui n'ont, eux, jamais le même sens. Ainsi, on a pu établir une propriété sémantique de *défendre* (sa polysémie) uniquement à partir de relations dans le graphe de synonymie : c'est sur ce genre de considérations, purement structuralistes, que se basent les travaux actuels sur les réseaux lexicaux.

Promenades aléatoires dans les réseaux de mots

Bruno Gaume, chercheur CNRS de Toulouse, a été le premier en France à se plonger dans l'étude de ces graphes sémantiques lexicaux. Il a pu montrer que la plupart de ces graphes sont des réseaux complexes, tels que les définit Alain Barrat dans sa contribution à ce numéro : ils ont une structure de "petit-monde" (*small world graphs*) et ils sont "sans échelle" (*scale-free*). L'aspect petit-monde repose sur deux propriétés qui sont caractéristiques des réseaux sociaux humains : d'une part une distance moyenne très petite relativement à la taille du graphe (on peut atteindre n'importe qui sur la planète avec très peu d'intermédiaires) et d'autre part une structure locale très riche (les amis de nos amis sont – le plus souvent, mais pas toujours ! – aussi nos amis). Quant à l'aspect sans échelle, ce sont sans doute les cartes routières qui en

donnent l'image la plus simple à comprendre : que l'on regarde à l'échelle d'un continent, d'un pays, d'une région, ou d'un département, on observe une structure analogue, avec quelques centres qui "irradient" dans toutes les directions et un plus grand nombre de villes qui semblent plus modestement connectées mais qui, à une autre échelle, peuvent se révéler elles-mêmes très centrales par rapport à des localités encore plus périphériques.

Bruno Gaume a aussi et surtout mis au point une mesure de proximité entre sommets qui tient compte de la structure globale du graphe. Cette mesure, qu'il a appelée "proxémie", s'est avérée très pertinente pour caractériser la distance sémantique entre deux mots. La méthode de calcul n'est, au fond, qu'une systématisation des jeux évoqués au début de cet article : Partant d'un sommet A donné, on simule des promenades aléatoires suivant les arêtes du graphe, et l'on calcule la probabilité de se retrouver au sommet B au bout d'un certain temps (en fait, cette mesure n'est pas une distance au sens mathématique car elle n'est pas symétrique : la proxémie du sommet A relativement au sommet B n'est pas égale à celle de B relativement à A).

Grâce à la proxémie, Bruno Gaume a pu représenter ces réseaux de manière géométrique, mettant en évidence les dimensions sémantiques qui les structurent. Prenons l'exemple d'un des graphes qu'il a beaucoup étudié, Synoverbe, l'ensemble des verbes français muni de la relation de synonymie telle qu'elle est donnée dans le dictionnaire électronique des synonymes de l'Université de Caen. Ce graphe comporte quelque 9000 sommets (chaque sommet représentant un verbe) et près de 50 000 arêtes (chaque arête représentant une relation de synonymie entre deux sommets). Bien entendu, il existe une grande disparité entre les sommets : certains verbes sont très centraux, au sens où ils sont au centre de zones très densément connectées, tandis que d'autres ne sont rattachés au reste du graphe que par quelques liens. On trouvera figure 1 la représentation en trois dimensions, grâce à la proxémie, des deux cents verbes les plus centraux. On observe qu'ils s'organisent suivant quatre axes qui forment une sorte de tétraèdre conceptuel du lexique verbal du français. Au bout du premier axe, noté A sur la figure, on trouve des verbes exprimant la fuite et le rejet (*partir*, *fuir*, *disparaître*, *abandonner*, *sortir*). Autour de B on a des verbes de production et de croissance comme *exciter, exalter, animer, soulever, transporter, provoquer, agiter, augmenter.* Le troisième axe C est caractérisé par l'idée de lien et de communication (*assembler, joindre, accorder, fixer, établir, indiquer, montrer, révéler, exposer, marquer, dire, composer*). Enfin, la région D correspond à des verbes de destruction et de dégradation tels que *briser, détruire, anéantir, abattre, affaiblir, ruiner, épuiser, écraser, casser,*

dégrader. Il faut souligner que l'on passe d'une région à une autre par des changements sémantiques graduels : ainsi on passe de B à D par la série de verbes *exciter, enflammer, agiter, tourmenter, troubler, ennuyer, bouleverser, fatiguer, ruiner, détruire, anéantir*.

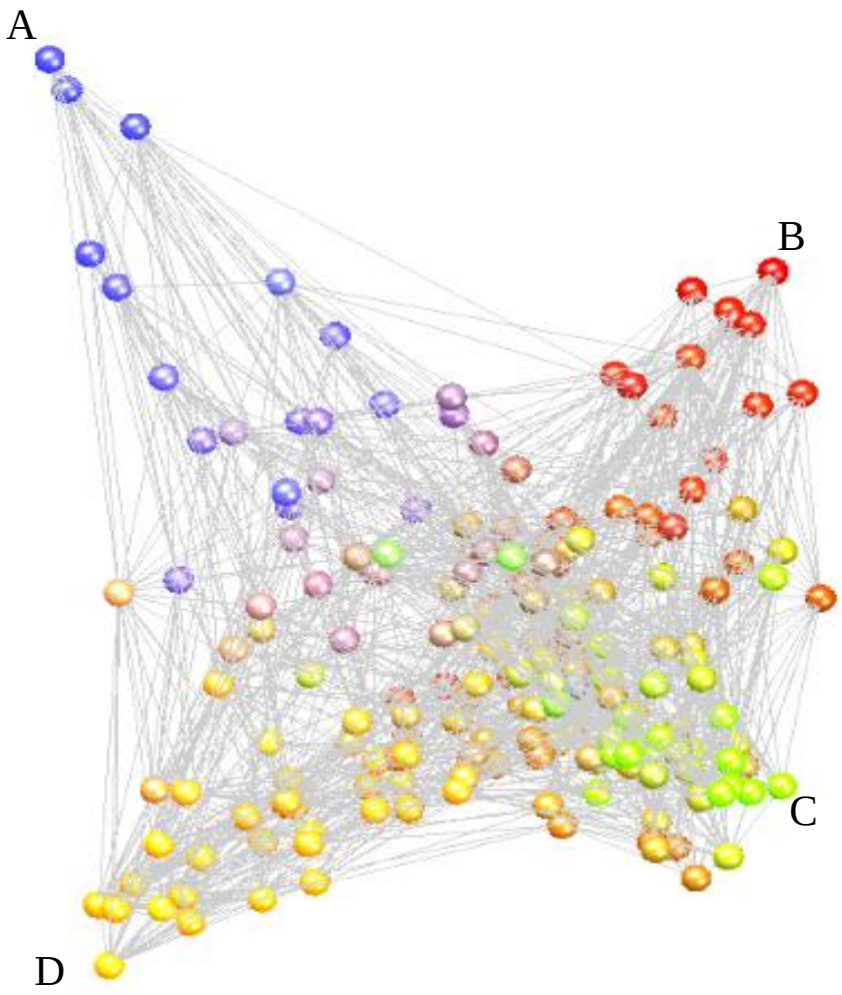

Figure 1. Représentation du lexique verbal français à l'aide de la mesure de proxémie

Notons aussi que, même si seuls les deux cents verbes les plus centraux sont dessinés sur la figure, cette représentation peut potentiellement contenir tous les verbes du graphe : pour ne donner qu'un exemple, si l'on ajoute le verbe *accabler*, qui ne fait pas partie des 200 premiers verbes, il vient se placer entre *écraser*, *fatiguer* et *bouleverser*, comme on pouvait s'y attendre. Ainsi, les quatre axes représentent quatre dimensions sémantiques qui organisent l'ensemble du lexique verbal. Et il n'est pas inintéressant de constater que ces quatre dimensions

correspondent, peu ou prou, à quatre types de comportement du répertoire de base des espèces animales : croissance et reproduction (B), prédation et agression (D), fuite et évitement (A) et coopération (C).

Cliques et espaces sémantiques

Dans notre laboratoire nous avons développé une autre méthode pour exploiter l'information présente dans les graphes lexicaux. Elle est basée sur la notion de clique. Une clique d'un graphe est un ensemble de sommets du graphe qui sont tous interconnectés deux à deux. Nous nous sommes en effet rendu compte que les cliques du graphe de synonymie correspondaient à des sens très précis des unités lexicales et qu'elles pouvaient donc servir à décrire avec précision les différents sens d'un mot.

Prenons l'exemple de verbe *jouer*. Ce verbe est fortement polysémique : il prend des sens différents dans différents contextes, comme le montrent les exemples suivants :

Les enfants jouent dans la cour.

Pierre joue aux échecs.

Pierre joue en bourse.

Marie joue Andromaque.

Marie joue du piano.

Marie joue des coudes.

Pierre joue les innocents.

Pierre nous a joué un sale tour.

La porte joue sur ses gonds.

La barque joue sur son ancre.

etc.

Bien entendu, ces sens sont tous apparentés les uns aux autres, avec des proximités plus ou moins grandes entre eux. Comment peut-on rendre compte à la fois de la diversité de ces sens et de leurs similitudes ? Dans le dictionnaire électronique des synonymes, *jouer* a près d'une centaine de synonymes (94 pour être précis), ce qui n'est pas étonnant mais qui ne peut pas beaucoup nous aider car la plupart de ces synonymes sont eux-mêmes très polysémiques, comme entre autres *faire*, *pratiquer*, *rouler*, *exposer*, *agir*, *reproduire*, etc. En revanche, si l'on regarde les cliques du graphe qui contiennent *jouer* (il y en a aussi une centaine, 98

exactement), on s'aperçoit que chacune décrit un sens bien déterminé de *jouer*, comme le montrent les quelques cliques suivantes :

{ *jouer*, *s'entraîner*, *s'exercer* }

{ *jouer*, *aventurer*, *compromettre*, *exposer*, *hasarder*, *risquer* }

{ *jouer*, *contrefaire*, *copier*, *imiter*, *mimer*, *reproduire*, *singer* }

{ *jouer*, *folâtrer*, *s'ébattre*, *s'ébrouer* }

Certaines cliques sont très proches les unes des autres, ne se différenciant que par de légères nuances de sens. Pour ne prendre qu'un exemple, outre { *jouer*, *aventurer*, *compromettre*, *exposer*, *hasarder*, *risquer* } que nous venons de citer, on trouve aussi { *jouer*, *miser*, *boursicoter* } et { *jouer*, *miser*, *parier*, *ponter* }. Nous avons mis au point un calcul de distance qui rend compte de ces proximités. Cela nous a permis d'associer à chaque unité lexicale polysémique un "espace sémantique" dans lequel sont représentées par des points toutes les cliques contenant cette unité, et qui donne une bonne idée des liens entre les différents sens du mot. On trouvera figure 2 l'espace sémantique associé à *jouer*.

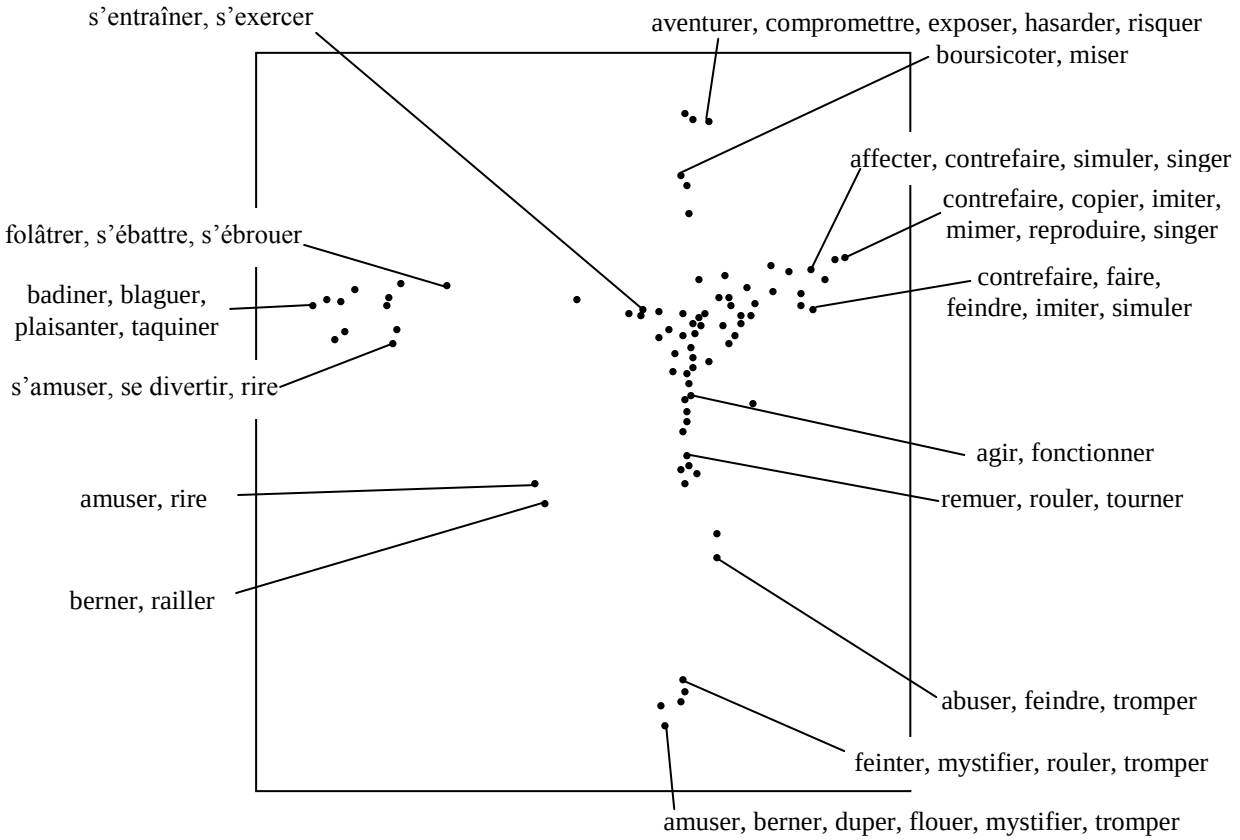

Figure 2. Espace sémantique associé à *jouer*

L'analyse de cet espace sémantique révèle plusieurs caractéristiques intéressantes du sémantisme du verbe *jouer* :

- On remarque d'abord qu'il n'y a pas de "dégroupement" notable des sens : on passe d'un sens à l'autre de manière continue, ce qui milite en faveur d'une unité sémantique de ce verbe.
- Les diverses significations de *jouer* s'organisent selon deux axes (cf. figure 3). L'axe vertical fait passer progressivement de valeurs où *jouer* désigne une activité que l'on exerce aux dépens d'autrui (*berner, duper, mystifier*) à des valeurs où, au contraire, l'activité s'exerce à ses propres dépens (*hasarder, risquer, miser*). Sur l'axe horizontal, on passe de valeurs où l'activité est centrée sur le sujet (*se divertir, rire, folâtrer*) à des valeurs opposées où l'activité consiste à se projeter sur autrui (*copier, imiter, simuler*).

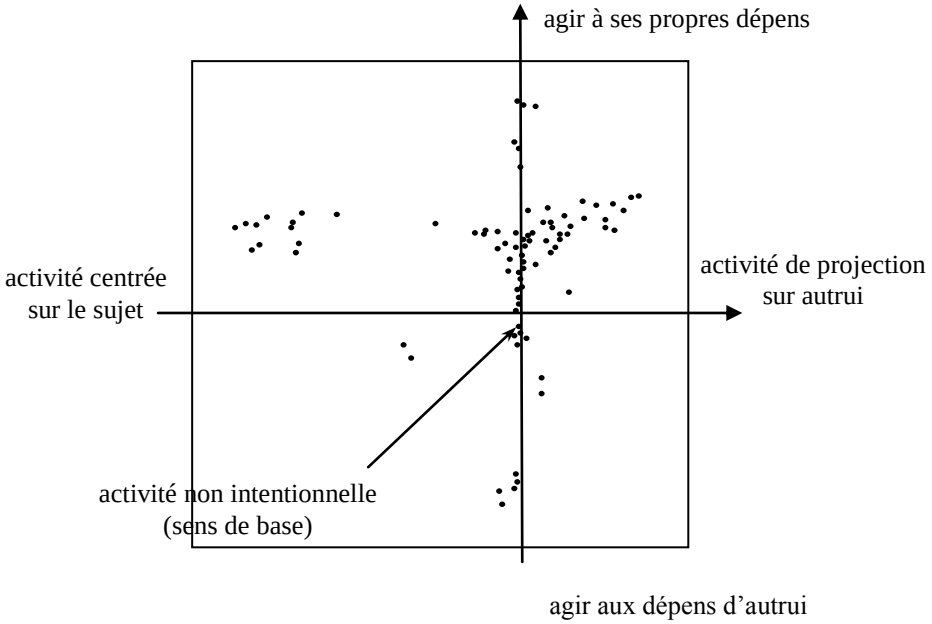

Figure 3 : Structure de l'espace sémantique associé à *jouer*

- Au centre de l'espace sémantique, on trouve des valeurs "neutres" pour les dimensions subjectives portées par les deux axes. C'est notamment dans cette zone que l'on trouve les sens de *jouer* dans lesquels il désigne une activité non intentionnelle (*La porte joue sur ses gonds*, *La barque joue sur son ancre*, etc.).

Cette dernière remarque peut nous permettre de découvrir ce qui fait l'unité du sémantisme de *jouer*. En effet, on peut faire l'hypothèse que ces valeurs neutres représentent un sens de base

dont les traits essentiels seraient partagés par tous les sens du verbe. L'étude de ce sens de base peut donc se révéler très précieuse pour l'analyse sémantique du verbe.

Or dans ce type d'emplois, *jouer* désigne une activité qui s'exerce selon des degrés de liberté "non standards", non prévus par le dispositif en question. Une porte est faite pour tourner autour d'un axe. Quand elle joue sur ses gonds, cela signifie qu'à ce mouvement de rotation "normal" se superpose un mouvement non prévu de translation verticale qui représente une liberté supplémentaire pour le mécanisme.

Il est clair que cette caractéristique est présente dans tous les sens du verbe *jouer*. Dans tous les cas, *jouer* désigne une activité qui s'oppose à une activité régulière, programmée, en ouvrant des degrés de liberté sur lesquels cette nouvelle activité peut s'exercer. Cela ne veut pas dire que cette nouvelle activité ne soit pas elle-même régulée (que l'on joue aux échecs, au football, au théâtre, ou encore aux courses). Mais les règles de cette activité de jeu, quand elles existent, s'inscrivent dans un espace de liberté qui n'a plus rien à voir avec le fonctionnement de l'activité normale à laquelle elle s'oppose.

Ainsi, cet exemple montre que les représentations obtenues à partir du réseau de synonymie sont pertinentes pour l'analyse sémantique de la polysémie. Bien entendu, ce n'est qu'un outil qui ne saurait se substituer au travail du linguiste. Mais c'est un outil puissant, qui fait clairement ressortir la structure sémantique d'une unité lexicale à partir des relations paradigmatiques que l'unité étudiée entretient avec les mots de sens voisins. A ce titre, il constitue une aide très précieuse pour les sémanticiens et les lexicographes.

Au-delà de la linguistique

Les méthodes que nous avons présentées rapidement ici pourraient sans doute être appliquées à d'autres réseaux complexes, dans d'autres domaines que la linguistique. Tout réseau complexe impliquant, d'une façon ou d'une autre, une "sémantique" pourrait bénéficier de ce type d'analyse, qui a l'avantage de révéler, sans a priori théorique, les dimensions structurantes des phénomènes étudiés. On peut penser notamment aux réseaux de sites Web, ou encore aux réseaux de chercheurs dans une discipline donnée…

Pour en savoir plus

Voici deux articles, accessibles en ligne, qui développent plus en détail ce qui a été présenté ici :

Gaume B., Venant F., Victorri B. Hierarchy in lexical organization of natural language, in D. Pumain (éd.), *Hierarchy in natural and social sciences*, Methodos series, vol 3, Springer, 2006, p. 121-142. http://halshs.archives-ouvertes.fr/halshs-00009918

Ploux S. et Victorri B., Construction d'espaces sémantiques à l'aide de dictionnaires de synonymes, *Traitement automatique des langues*, 39, n°1, 1998, pp.161-182. http://halshs.archives-ouvertes.fr/halshs-00009433

Et plusieurs adresses de sites Web présentant des réseaux lexicaux :

- WordNet : http://wordnet.princeton.edu/
- La proxémie : http://erss.irit.fr/prox/
- Le dictionnaire électronique de synonymes : http://www.crisco.unicaen.fr/dicosyn/
- Les atlas sémantiques : http://dico.isc.cnrs.fr/