Hierarchy in lexical organisation of natural languages

Bruno Gaume, Fabienne Venant, Bernard Victorri

Hiearchy in Natural and social Sciences, Methodos series, vol 3, Springer, 2006 - Draft

Hierarchy in lexical organisation of natural languages

Bruno Gaume, Fabienne Venant, Bernard Victorri

Introduction

Words have often been compared to living organisms. One of the first authors to develop this idea was the French linguist Arsène Darmesteter who wrote in 1887 a book entitled *La vie des mots étudiée dans leurs significations* (Darmesteter 1887). He described the evolution of the meanings of words as a 'struggle for life' (*concurrence vitale* in his own words). To stay alive, words need to occupy as much 'semantic ground' as possible, particularly by taking over 'territories' in new semantic domains, creating thus the important phenomenon of polysemy (see also Bréal, 1897). Some words enjoy real 'success stories', expanding their meanings in many different directions, in a rather monopolistic way, while others decline and eventually die.

Even though Darmesteter was more concerned by the analogy with biology rather than with social sciences, it is interesting to notice that his conception of interactions between words as a 'struggle for life' could also be applied to many social, political and economical human interactions. As a matter of fact, words are organized in the lexicon as a complex network of evolving semantic relations. It is not surprising that such a system shares many important properties with complex systems of social relations as well as complex systems of biological relations.

In this paper, we focus on the comparison between lexical systems and social structures. It has been discovered very recently that several graphs of semantic relationships between words belonged to the class of what is called 'small world' graphs, i.e. they first characterized graphs of social relationships. This result opens new perspectives in lexical semantics. It suggests that lexical graphs contain a rich amount of information concerning the semantic structure of the lexicon. In particular, we can expect that analyzing these graphs will enable a better understanding of its hierarchical organization.

We present here a mathematical model in which each word is associated with a region in a global semantic space. In this representation, polysemy is taken into account by the size of the regions: words with many different meanings are represented by very large regions, while words with unique precise meaning are represented by very small (point-like) regions. If the regions associated with two words intersect, these two words share one or several meanings. Said differently, overlaps of regions in the semantic space correspond to (partial) synonymy between words.

Thus this model brings an interesting light to the similarity between graphs of lexical and social relationships. The semantic space plays the same role for words that the geographical space does for humans. Words meet in the semantic space like people meet in the world. Each meeting between two words means that there is a place in the semantic space, i.e. a precise meaning, that is common to both of them, exactly as a meeting between two persons shows that there is a place which belongs to the geographical fields of activity of both people.

As we know, fields of activity are not homogenously distributed on the map. There is a scaling structure from big cities to small villages corresponding to a scaling distribution of the density of fields of activity on the geographical map. Small world properties of many social networks are clearly related to the scaling structure of the underlying geographical space.

Since lexical graphs are also small world graphs, we assume that the same holds for their underlying space. Meanings must have a scaling distribution in the semantic space, from places of high density (covered by many words) to the equivalent of villages, i.e. meanings that are poorly covered by the lexicon.

To test this hypothesis, we designed several methods to build a semantic space from a graph of synonymy. We present here these methods, illustrating them with the example of the French verb lexicon. As we will see, the different methods lead to rather similar results, showing that they reveal intrinsic properties of the semantic structure of the lexicon.

Small world graphs

Watts and Strogatz (1998) defined small world graphs as graphs combining two features: a high 'clustering coefficient' and a short 'characteristic path length'.

The clustering coefficient is a measure of how tightly the neighbors of a node in the graph are connected to each other. Numerically, it is defined as the proportion of pairs of nodes linked with one another among all the neighbors of a node[1]. In social terms, it measures how many of one's acquaintances know each other. So, it is not surprising that social networks have a high clustering coefficient (most of my friends are friends of each other).

The characteristic path length is a measure of how far two nodes are situated one from the other in the graph. The distance between two nodes is defined as the minimum number of edges traversed to get from one of them to the other. The characteristic path length is the average of the distance over all pairs of nodes. In the social context, a short characteristic path length means that there is generally a small number of go-betweens in the smallest chain which connects two people. This is the popular notion of "6-degrees of separation" (Guare 1990) resulting from the famous experiments devised by Stanley Milgram who introduced the term of "small world" (Milgram 1967).

A third property of small world graphs was put forward after Watts and Strogatz's work. It concerns the distribution of the number of edges among the nodes. It was discovered that the degree of a randomly selected node (the number of its neighbors) follows a power-law distribution[2]. The power-law was first verified on the Web network, which is also a small world graph (Barabási *et al.* 2000, Huberman & Adamic 1999, Kleinberg *et al.* 1999), but it also holds for social networks (Newman 2001, Barabási *et al.* 2002). An important consequence is that small world graphs have a "scale-free" topology. Roughly speaking, it means that the ratio of very connected nodes to the number of nodes in the rest of the network remains constant as the network changes in size.

As shown by Ravasz & Barabási (2003), the two features, high clustering coefficient and scale-free topology, determine an original combination of modularity and hierarchical organization. As the authors put it, "we should not think of modularity as the coexistence of relatively independent groups of nodes. Instead, we have many small clusters, which are densely interconnected. These combine to form larger, but less cohesive groups, which combine again to form even larger and even less interconnected clusters. This self-similar

---

[1] More precisely, it is computed as follows. Let p be a node, k its degree (number of its neighbors) and n the number of edges among them. The clustering coefficient at node p is $c(p) = 2n/k(k-1)$. It is easy to check that c(p) lies between 0 and 1. It equals 0 if there is no edge linking any pair of neighbors of p, and 1 if all neighbors are connected with one another. Then the clustering coefficient C of the graph is the average of c(p) over all nodes.

[2] The probability P(k) that a randomly selected node has k links follows the law $P(k) \sim k^{-\lambda}$ where λ is a constant for the given graph.

nesting of different groups or modules into each other forces a strict fine structure on real networks".

So, hierarchy appears as an emergent feature of the network. It is not a simple pyramidal organization. No node can be viewed as dominating other nodes. The hierarchy is made of groups of nodes, with small clusters at the bottom and very large groups at the top. Moreover, groups of nodes may overlap at any level. A group (or a part of a group) of the lower level can be included in more than one group at the higher level, since it can belong to several different groupings having approximately the same clustering coefficient.

As far as social networks are concerned, such a hierarchical structure can, in many cases, be related with the underlying geographical space. For instance, acquaintance relationship is highly correlated with geographical proximity. So we can expect a duality relation between the hierarchical organization of a graph of acquaintance and the hierarchical structure of the geographical distribution of humans. Each person (node of the graph) is associated with her spatial zone of activity, which may be a very large area for some individuals. Then, small clusters of strongly interconnected people correspond to relatively small areas where few people often meet, such as villages and districts in cities (notice that the same individual can belong to several different clusters, corresponding for instance to his home and his workplace). As we climb up the hierarchy on the graph by considering larger and larger groups (less and less interconnected), we obtain a smaller number of more densely occupied places. At the top level, the largest groups correspond to the centres of the largest cities.

Now if we consider other types of small world graphs, we can assume that there is always an underlying space with a dual hierarchical structure, even though most of the time the nature of this space is more abstract than a geographical map. This is the main idea that we will develop here to study the semantic structure of the lexicon. But before focusing on lexical graphs, we have to remark that the approach could be applied to any 'semantic' graph. For instance, let’s consider, in the Internet universe, the small world graph whose nodes are all the web pages and whose edges indicate the presence of a hypertext link. Clearly the geographical factor is not relevant. But if we build an abstract semantic space whose dimensions are the different topics that a website may deal with, every website can be conceived as occupying a region of the space. Generalist sites will be represented by rather large areas, whereas more specialized ones will occupy smaller regions. We can expect that some places of the space will play the role of big cities in being densely covered by many sites, and others the role of countryside in being rarely broached on the web. Studying the hierarchical organization of the semantic space and its evolution could bring interesting insights of what is going on on the web: what are the hottest topics, which ones are growing up and which ones are declining. Of course, the two most important problems with this approach is first to design the abstract semantic space (how to choose the relevant dimensions and the relevant metric on the space), and second to compute automatically the region associated with each website. The methods we present here provide the beginnings of a solution to both problems since they allow to derive the whole geometrical representation from computations on the initial graph, which is (relatively) easy to obtain.

Lexical graphs

Lexical graphs have been a more and more important topic for the last few years, following the tremendous development of electronic linguistic resources (dictionaries and large corpora). The most famous example is WordNet, a very rich lexical database for English (cf. Fellbaum 1998) comprising more than 150 000 words and many different relations between them. There are different types of lexical graphs, depending on the semantic relation used to

build the graph. This relation can be a paradigmatic relation such as synonymy, hyperonymy or translation (when more than one language is involved). It also can be a syntagmatic one, when, for instance, two words are linked if they appear in a same sentence in a given corpus. It can also be a more general semantic proximity relation, mixing syntagmatic and paradigmatic dimensions, as is the case when two words are linked if one appears in the definition of the other in a given general dictionary (cf. Gaume et al., 2002).

The structures of many lexical graphs of all sorts have been studied (see, among others, Ferrer & Solé 2001, Sigman & Cecchi 2002, Ravasz & Barabási 2003, Gaume et al. 2001, Gaume 2003). All the studies lead to the same conclusion. It seems that every lexical graphs have a small world structure, whatever the nature of the semantic relation involved. This result is important, since it shows that what is at stake is an intrinsic property of the semantic organization of the lexicon in natural languages. It sustains the idea of an underlying semantic space whose hierarchical topological organization could explain why different semantic relations share a same small world graph structure.

The graph we worked on for the present study, Synoverbe, is typical of these lexical graphs. It is a synonymy graph of French verbs which has been extracted from a general dictionary of French synonyms[3] by one of us (Bruno Gaume). Synoverbe has roughly 9000 nodes and 50,000 links. It has the three characteristic features of small world graphs. Its characteristic path length is small, around 4, which is the order of magnitude that can be expected from a random graph with the same number of nodes and links. Its clustering coefficient is very large, around 0.3, five hundred times higher than a random graph[4]. As shown Figure 1, the distribution of the degree of the nodes follows a power-law distribution.

[3] The general dictionary of French synonyms is managed by J.L. Manguin at the CRISCO research laboratory in linguistics, at the University of Caen. It is available on the Web (http://www.crisco.unicaen.fr/).

[4] For a random graph of n nodes and p links, the characteristic path length is $L = \log(n)/(\log(p)-\log(n))$ on average, and the clustering coefficient is $C = p/n^2$ on average. In our case ($n = 9000$, $p = 50{,}000$), the computation gives $L = 5.31$ and $C = 0.0006$. The precise figures for Synoverbe are $L = 4.17$ and $C = 0.318$.

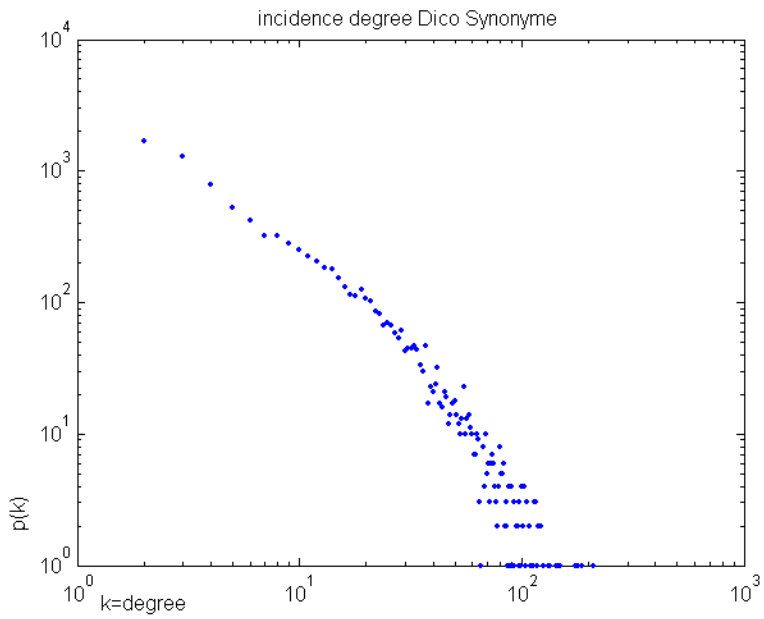

Figure 1. *Synoverbe : Log-log plot of the distribution of the degrees (number of links by node).*

A more detailed description would give a more concrete idea of what this distribution means. While the average degree is less than 12 among the 9000 nodes, about 1000 nodes have more than 30 links and about 100 more than 80 links. Furthermore, nearly 90% of the nodes are directly linked to at least one of the 1000 most connected nodes, and nearly 50% are directly linked to at least one of the 100 most connected ones. In other words, among the nearly 10,000 French verbs, we can extract a subset of 1000 verbs which covers virtually all the meanings covered by the entire set, since nearly all the French verbs are synonyms of verbs of the subset. Moreover, a subset of only a hundred verbs covers half of the verb meanings. These verbs are of course the most highly polysemic ones, since each of them has a hundred or so synonyms. The most connected ones, like *faire* (translations: make, do …) and *prendre* (translations: take, get …) have even more than 200 synonyms. They have two other interesting properties: (1) they are the most frequently used by French speakers, and (2) they are the first to be acquired by children. No doubt that they are the winners in the 'struggle of life' evoked by Darmesteter ! It is also worth noticing that they are rather tightly interconnected. In fact, the subgraph composed by these 100 verbs has basically the same properties as the whole graph: an average of 6 links by node, a characteristic path length of the same order of magnitude as the one of a random graph of the same size and connectivity, and a clustering coefficient markedly higher than the one of a random graph.

Thus we can then describe the structure of the French verb lexicon as a hierarchical structure with three levels:

- at the top, a first subset of 100 verbs, each with several general meanings. It represents the basic vocabulary for the verb semantic domain. With these 100 verbs, one can express most actions and events, but vaguely and without accuracy.

- at the second level, a subset of 1000 verbs presenting a rather important degree of polysemy (more than 30 synonyms each) cover all verb semantics, quite sufficient to describe any

action or event in everyday life. In fact, the verb lexicon used by most people in production is only a part of this subset.

- at the third level, the entire set of nearly 10,000 verbs, permitting very precise descriptions, subtle uses of qualifications, and different styles and levels of language (formal, technical, poetic, slang, etc.).

Even though it gives a first idea of the structure of the French verb lexicon, the above description is neither accurate nor satisfying enough. The problem comes from the arbitrary nature of the choice of our levels. Why three levels, rather than four or five? As a matter of fact, the hierarchy is not discrete, with well identified intrinsic levels: it is a continuous scaling. Therefore, we need mathematical tools suited to continuous representations in order to model the lexical hierarchy in a more appropriate way. Here is the main reason why we turned to geometrical tools and quantitative measures such as the notion of proxemy, that we introduce in the following paragraphs.

Proxemy: a measure of semantic nearness

Bruno Gaume, defined a new measure of the nearness of nodes in a graph that takes into account the density of the graph along the different paths linking them. This measure, that he called *proxemy* (Gaume 2002, 2003, 2004), is well suited for small world graphs because it relies on the structural properties of the graph. We know that somehow two nodes are never far from one another in a small world graph, since the characteristic path length is small. But for the same minimal path length, two nodes may be very loosely related by only one path linking two separate dense regions, or they may belong to the same dense region with many different paths of minimal length connecting them. Obviously, the nodes must be qualified as "closer" in the latter case than in the former one: this is exactly what the measure of proxemy does.

A good idea of the notion of proxemy can be given by considering a particle wandering randomly on the graph, going from one node to any of its neighbors with equal probability. Let the particle be at node A at the beginning of the process. After the first time step, the only nodes that can be reached by the particle are the direct neighbors of A[5], each with a probability of 1/n, where n is the degree of A. After k time steps, any node B located at a distance of k links or less can be reached, the probability of this event depending on the number of paths between A and B, and the structure of the graph around the intermediary nodes along those paths. The more interconnections between these nodes, the higher the probability of reaching B from A will be. In other words the probability for a random particle to go from A to B is a good candidate for the measure we were looking for: we call it the k-proxemy of B with respect to A.

More generally, we define the *k-proxemy of a node with respect to a given subset of nodes* as the probability for a particle to reach it after k time steps if the particle were at time 0 on one of the nodes of the subset (if the subset contains p nodes, each of them is endowed with a probability of 1/p to be the starting point of the particle)[6]. When the subset includes all the nodes of the graph, we will speak of *global k-proxemy*.

---

[5] Including A itself: for technical reasons (property of ergodicity, see note below), it is preferable to consider that the graph is reflexive, i.e. that each node is its own neighbor.

[6] From a mathematical point of view, it is easy to show that the random process we described is a markovian process. If we call $A = (a_{ij})$ the matrix of adjacency of the graph ($a_{ij} = 1$ if nodes i and j are connected, else $a_{ij} = 0$), the markovian matrix $M = (m_{ij})$ associated with the random walk of the particle is given by $m_{ij} = a_{ij}/s_i$ where s_i is the sum of the row i of the matrix A. The k-proxemy of a node n with respect to a subset S can be

It must be emphasized that the value of k plays a crucial role in the definition of proxemy. For very small k, the k-proxemy fails to catch the structural properties of the graph because it is too local: the proxemy of most nodes of the graph is zero with respect to any given node. On the other hand, if k is too large, the k-proxemy of a given node with respect to any subset does not depend on the subset any longer: it tends towards a value that only depends on the degree of the given node[7]. Thus interesting values of k lie between the two extremes. Empirically, it seems that the best results are obtained with values belonging to the interval (L, 2L) where L is the characteristic path length of the graph. For instance, in the case of Synoverbe (L = 4.17), the value k = 6 proved to be the best one. From now on, we drop the "k" prefix in the term k-proxemy, assuming a choice of k in the right interval (and a value of 6 for the examples from Synoverbe).

Using proxemy, a geometrical representation of the graph can be built, which preserves its structural properties (Gaume 2004). To each node A of the graph is associated its *proxemic representation*, a vector whose n^{th} component is the proxemy of the n^{th} node of the graph with respect to the node A. In other words, the proxemic representation of a node gives the probability distribution over the whole graph for the random walk of a particle originating from this node[8]. This means that the proxemic representation takes into account the relations of a node with all the others: it is characteristic of the structural position of the node in the whole graph.

When dealing with a lexical graph, the proxemic representation could be qualified as 'Saussurian', since it fits exactly Saussure's structuralist theory according to which the semantic value of a lexical unit cannot be defined in absolute terms, but only by its relative position in the entire system. As a matter of fact, Karine Duvignau and Bruno Gaume have shown that proxemy is also relevant for psycholinguistic considerations, in particular in studying lexical acquisition and children production (Duvignau 2002, Duvignau & Gaume 2003, 2004) as well as in modeling disambiguation processing.

Here we will focus on the use of proxemy for visualizing the hierarchical organization of the lexicon. To begin with, we must notice that the global proxemy of a node (its proxemy with respect to the entire set of nodes) is a better indicator of its semantic extent than its degree. Whereas the degree of a node only indicates the number of its synonyms, its global proxemy gives more precise information about the more or less central role played by the node in the whole graph. Moreover, the nodes can be located in the same geometrical space thanks to their proxemic representation. Of course, the geometrical space cannot be faithfully visualized because of its high dimensionality, but the use of a classical method of dimension reduction (principal component analysis) allows to obtain a two or three dimensional representation preserving the main geometrical relations between the nodes we choose to visualize. We show figure 2 the proxemic representation of the 200 highest-ranked French verbs according to

---

computed as follows. Let $U=(u_i)$ be the vector associated with the uniform probability density over S ($u_i=1/s$ if the node i belongs to S, else $u_i=0$, s being the number of elements of S). Then the k-proxemy of n with respect to S is given by the n^{th} component of the vector V obtained by applying k times the transformation M to the vector U (in matrix notation: $V = U.M^k$ where U and V are row vectors).

[7] When the graph is reflexive, it can be shown (Gaume 2004) that the markovian process is ergodic. Then, a corollary of the theorem of Perron Froebenius implies that there is a unique stationary probability and that the process converges towards this stationary probability for any initial conditions (see for instance Semata 1981 and Bermann & Plemons 1994). In our case, it is easy to verify that the stationary probability is the vector $W = (d_i/2p)$ where d_i where d_i is the degree of the node i and p is the number of links in the whole graph.

[8] Computationally speaking, the proxemic representation of the node i is the i^{th} row of the matrix M^k, where M is the markovian matrix defined above (see note 6).

their global proxemy (computed from Synoverbe)[9]. Each verb is represented by a sphere whose size is proportional to its global proxemy. As can be seen on the figure, the most general French verbs (the top of the hierarchy) are organized along four semantic axes structuring the whole lexicon as a sort of conceptual tetrahedron. Around the first vertex (labeled A on the figure) can be found verbs expressing escaping and rejecting actions (*partir, fuir, disparaître, abandonner, sortir…* Interestingly, *quitter* is located between *disparaître* and *abandonner*). The zone around vertex B is composed by verbs expressing productive and enhancing actions like *exciter, enflammer, exalter, animer, soulever, transporter, soulever, provoquer, agiter, augmenter* (and *entraîner* between *attirer* and *provoquer*). The third vertex C is characterized by the ideas of connecting and communicating (*assembler, joindre, accorder, fixer, établir, indiquer, montrer, révéler, exposer, marquer, dire, composer…*, *réunir* between *attacher* and *joindre*, and *reveler* between *montrer* and *indiquer*). At last, vertex D corresponds to destructive and damaging actions such as *briser, détruire, anéantir, abattre, affaiblir, ruiner, épuiser, écraser, casser, dégrader…* The verb *tuer* is located there, between *altérer, dégrader* and *supprimer*.

It must be noticed that we can observe gradual semantic changes as we move from one vertex to another. For instance, moving from A to B we find successively *s'enfuir, fuir, partir, sortir, passer, courir, venir, marcher, aller, suivre, avancer, revenir, introduire, faire*. From A to D the gradation involves *s'enfuir, fuir, disparaître, quitter, abandonner, mourir, cesser, perdre, diminuer, supprimer, casser, anéantir, détruire.* Between B and D can be found the series *exciter, enflammer, agiter, tourmenter, troubler, ennuyer, bouleverser, fatiguer, ruiner, détruire, anéantir, briser*, whereas one passes from B to C through e*xciter, exalter, animer, soulever, provoquer, entraîner, augmenter, élever, conduire, déterminer, produire, former, dire, établir, exposer, indiquer, montrer, révéler*. Last example, here is the series from C to D: *fixer, assembler, joindre, réunir, arranger, attacher, retenir, serrer, fermer, arrêter, cesser, rompre, séparer, couper, étouffer, supprimer, diminuer, casser, affaiblir, abattre, anéantir, briser*.

Three important comments are worth emphazising:

- First, as shown by these examples, the geometrical distance between spheres presents a very good correlation with the semantic distance between lexical units: close verbs on the figure are also close by their meanings. This proves that proxemic representation actually catches semantic properties of the lexical units.

- Second, the global structure, sort of tetrahedron with its four vertices, is relatively independent of the precise number of top-ranked verbs used to build it: a very similar form is obtained with the first 100 or 300 verbs instead of the first 200. This means that the method is well suited to the continuous aspect of the hierarchical structure of the lexicon. Thanks to our geometrical representation, we do not need to define any 'levels' of hierarchy. The choice of the number of verbs taken into account is not a crucial decision, but a question of convenience: taking more verbs leads to a more accurate representation, but at the same time a less readable figure.

[9] Proxemic representations of different lexical graphs are available on the Web: http://dilan.irit.fr.

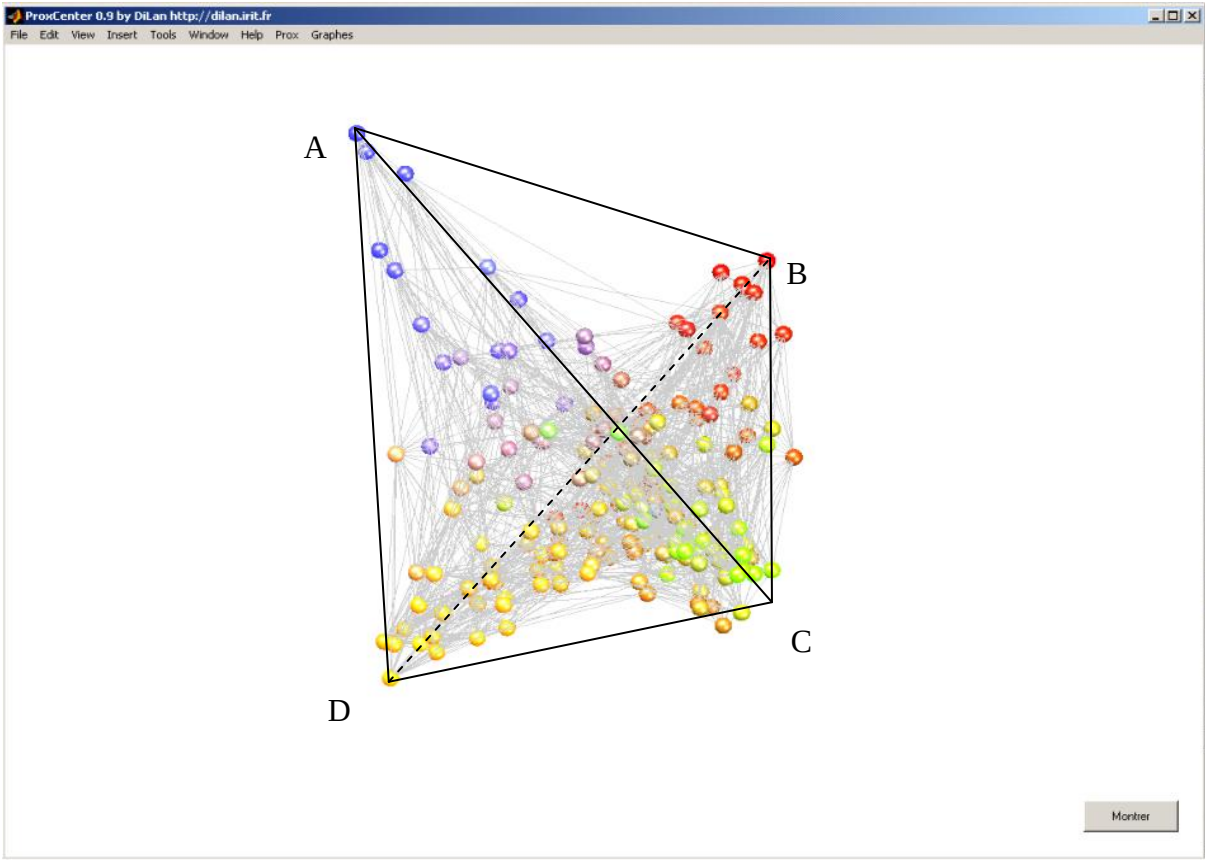

Figure 2. *Representation of the first 200 French verbs with highest global proxemy*

- The third remark is also a consequence of the continuous aspect of the geometrical tools. Once the representation has been built with a small number of top-ranked node, we can represent any of the remaining nodes in the same figure. In other words, the geometrical representation is a global referential frame in which we can locate all the nodes of the graph. As regards our example of Synoverbe, it follows that any French verb can be characterized by its location in the tetrahedron. For instance, if we add the verb *accabler*, which is not among the 200 top-ranked nodes, to the representation, we find that it is located in the D region, between *écraser*, *fatiguer* and *bouleverser*, as could be expected from its meaning. Far from being restricted to the 200 verbs used to build them, the four vertices correspond to four semantic dimensions whose relevance is general all over the French verb lexicon[10].

We can also use the proxemic representation of the nodes to visualize more local parts of the graph. Instead of using global proxemy to choose the nodes to be represented, we can choose to study any subset of nodes of particular interest by representing the verbs having the highest proxemy with respect to the selected subset. We will call such a representation a *proxemic zoom onto* the given subset. Actually, since all the nodes can theoretically be represented in the same high dimensional space, we can consider that we really zoom into a part of this representation when we select some nodes to visualize the relative positions of these nodes in the high dimensional space. Of course, we practically need to use principal component

[10] An important question is whether these dimensions are *universal*, i.e. shared by all human languages. This is one of the issues that we intend to explore in the near future.

analysis to reduce the dimensionality of the space, exactly as we proceeded when we visualized the global structure.

Figures 3 to 5 show such proxemic zooms. In each case, we have chosen a couple of antonyms as subsets defining the proxemy: {*monter*, *descendre*}, {*commencer*, *finir*} and {*aimer*, *haïr*} (respectively go up/go down, begin/end and love/hate). Each time, one of the two verbs is on the 'positive' axis and the other on the 'negative' one. It is interesting to see how the antonyms are connected by relatively short paths through their semantic domain, with semantically very relevant intermediary verbs: *sauter* (to jump) between *monter* and *descendre*, *partir* (to depart) between *commencer* and *finir*, *envier* (to envy) between *aimer* et *haïr*.

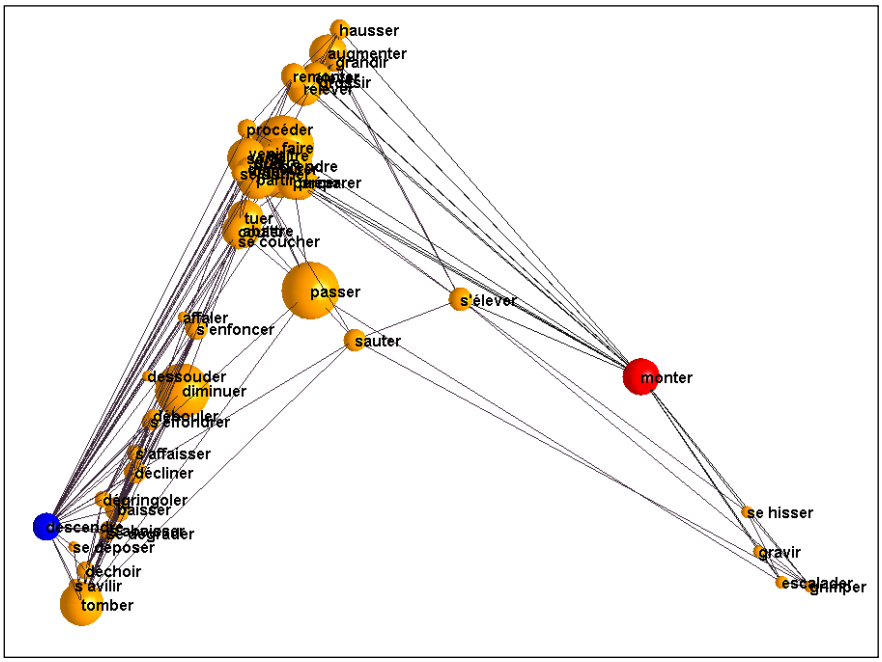

Figure 3. *Proxemic zoom onto {*monter, descendre*}*

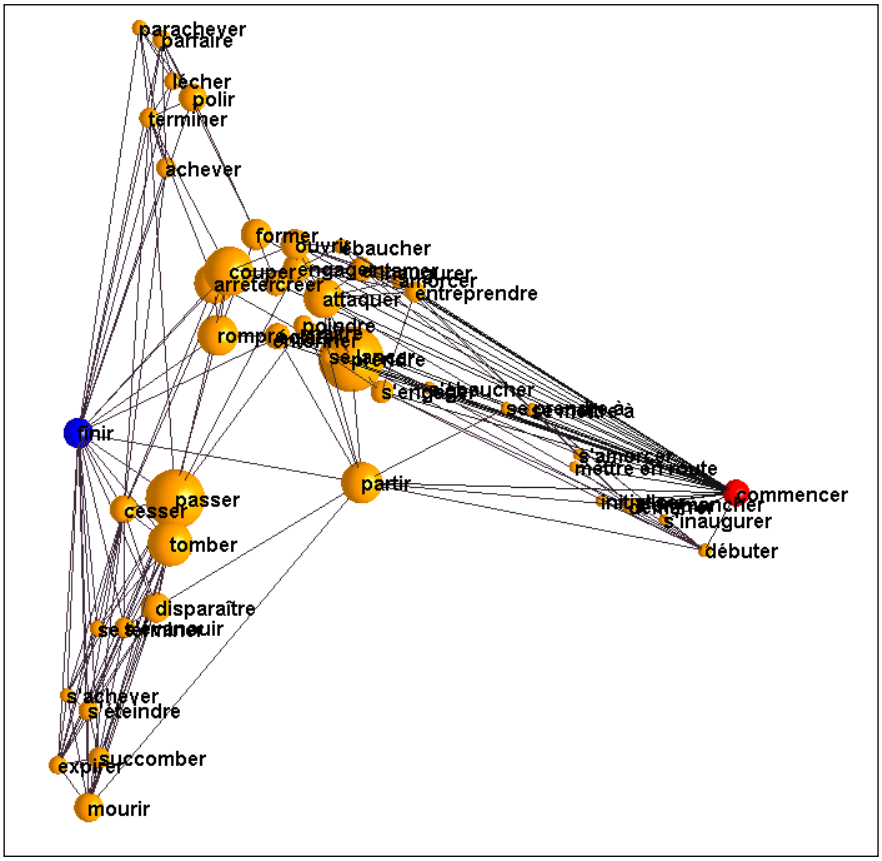

Figure 4. *Proxemic zoom onto {*commencer*,* finir*}*

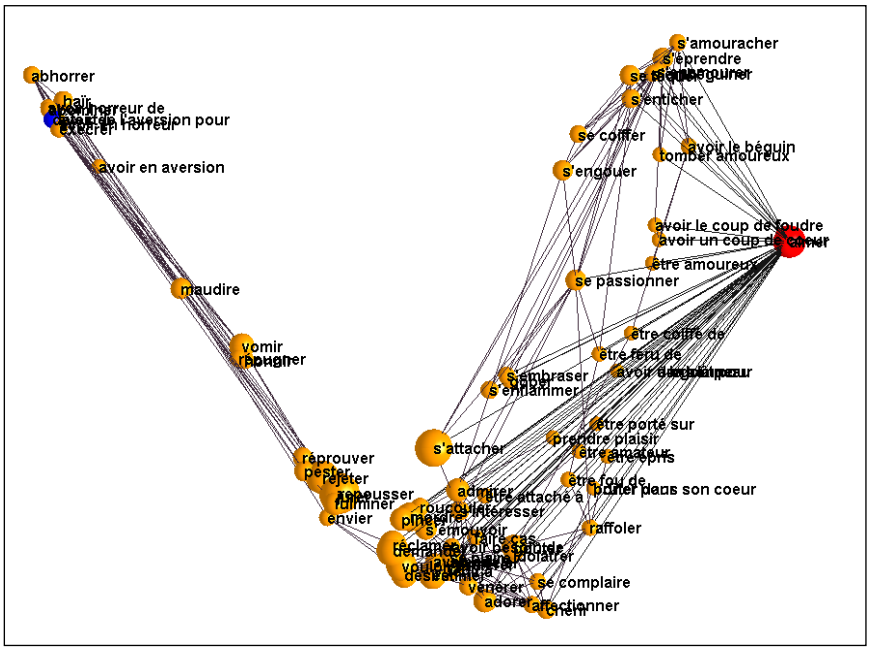

Figure 5. *Proxemic zoom onto {*aimer, haïr*}*

Proxemic zooms can also be used to study the polysemic organization of a single verb. We just have to zoom with respect to the subset reduced to this verb. For instance, figure 7 shows the representation around the verb *jouer* (to play). As can be seen on the figure, four specific meanings of the verb emerge: *s'amuser* (have fun), *risquer* (*jouer de l'argent*: to gamble), *tromper* (*se jouer de quelqu'un*: to deceive somebody), and *imiter* (*jouer les victimes*: to play the victim), the center of the representation corresponding to the more general meaning *pratiquer* (*jouer aux cartes, au tennis, du piano*: to play cards, tennis, piano).

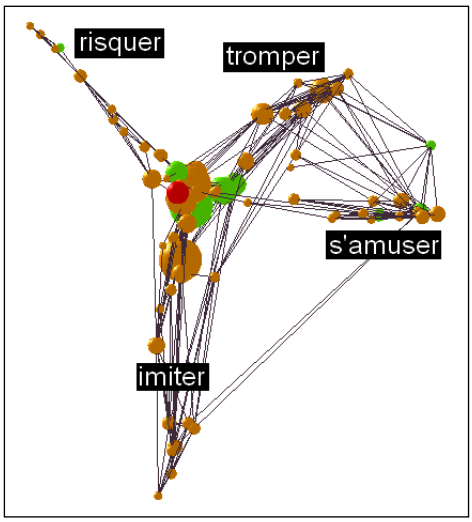

Figure 6. *Proxemic zoom onto* jouer

Thus, proxemy offers an interesting method to study a lexical graph at different scales, from the most global structure to the most detailed meanings of a word.

Semantic spaces

Can the geometrical figure obtained by the proxemic method be considered as the abstract semantic space we were looking for? As we said at the beginning of the paper, lexical units must be represented by regions rather than points in the semantic space, if we want to take into account their polysemy and the overlap of meanings characterizing partial synonymy between several units. In order to maintain coherence in our model, we must consider that the different meanings of a lexical unit are scattered over an area surrounding the proxemic vector representing the unit. As a matter of fact, this is exactly what we just did when we looked at the proxemic zoom onto the verb *jouer*. In the figure 6, the sphere labeled by *jouer* is located in the middle of the representation, where we found the more general meaning *pratiquer*, but all the other meanings are spread and situated relatively far from the sphere of *jouer*, which plays a role similar to a center of gravity. This can be better visualized in the figure 7, where *jouer* and its proxemic neighbors are represented in the global conceptual tetrahedron of the French verbs.

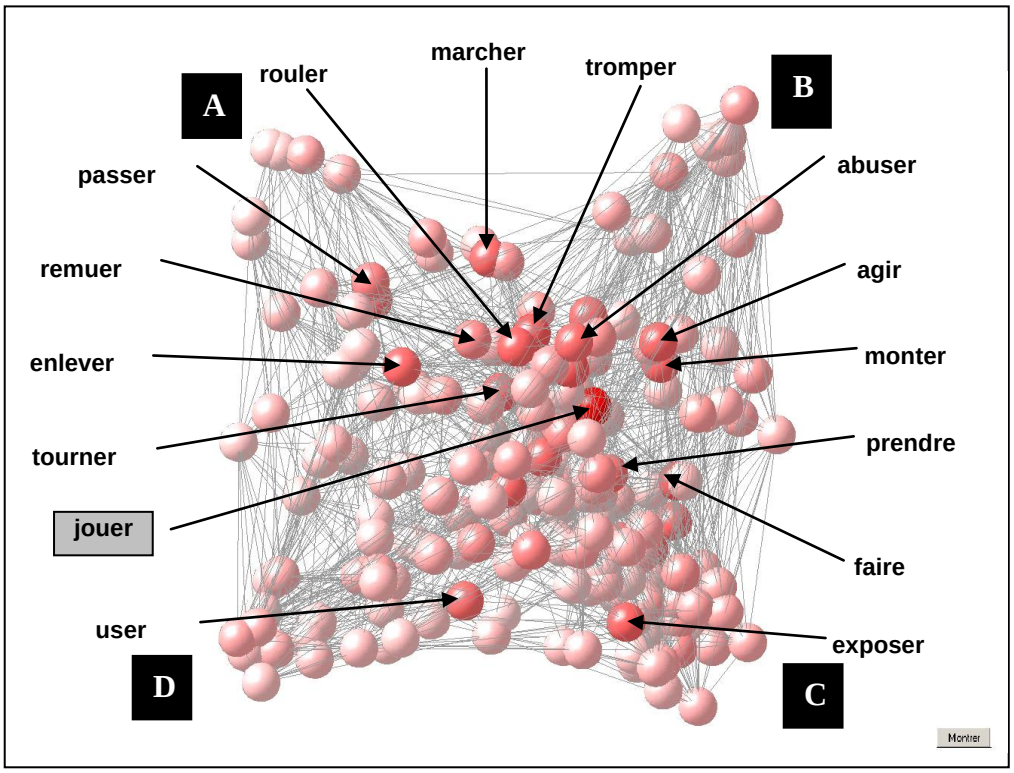

Figure 7 : *Localisation of* jouer *and its proxemic neighbors in the global representation of French verbs (hue indicates the proxemy with respect to* jouer*).*

Therefore, we shall define the semantic area associated with a given unit as the region containing all the units having a high proxemy with respect to it. With this definition, we can assume that proxemic representation gives a good approximation of the semantic space needed in our model.

In order to check this assumption, we used a completely different method of construction of a global semantic space. This method has been used for several years by one of us, Bernard Victorri, to build local semantic spaces associated with polysemic lexical units (Ploux &

Victorri1998). Fabienne Venant (2004) and Nabil Abdellaoui (2004) have extended very recently this method so as to apply it to the building of global semantic spaces.

The main idea of the method consists in associating points of the semantic space with the *cliques* of the lexical graph. The cliques of a graph are its maximal completely interconnected subsets of nodes, i.e., in our case, maximal sets of lexical units that are all synonyms for one another. The cliques define very precise meanings that can be considered as the intersection of the meanings of all the units belonging to the clique[11]. It is worth noticing that the "synsets" of WordNet (Fellbaum 1998) are analogous to the cliques in that they are also sets of words designed to represent the different meanings of a lexical unit.

Let us take an example to illustrate this point. As we saw, the French verb *jouer* displays a rather extended polysemy, with a large number of synonyms (precisely 94). Of course, most of its synonyms are far from being synonyms for one another. For instance, if we look at the synonyms that we presented above to characterize the different parts of the proxemic zoom of figure 6 (*s'amuser*, *risquer*, *tromper*, *imiter*, *pratiquer*), they convey very different meanings. On the opposite, the cliques containing *jouer* evoke a unique nuance of meaning of *jouer.* For instance, we find, among others, the three following cliques:

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

{*jouer*, *miser, boursicoter*}

{*jouer*, *miser, parier, ponter*}

All of them can be considered as instances of the '*risquer*' meaning of *jouer*, but each of them enhances a precise determination of this meaning (the first one evokes venturing and hazarding, the second speculating, and the third gambling and betting). It is then sensible to assume that each clique has to be represented by a point in the semantic space. As the number of cliques containing *jouer* is also rather large (precisely 98), we have enough points to design what we called the semantic space associated to *jouer*

In order to build the semantic space of a given unit, we compute a distance between the cliques containing the unit. We use the chi-square distance, a metric which is well known in statistical analysis, intensely used to compute correspondences between subsets of individuals and subsets of qualitative characteristics[12]. As usual, principal component analysis is applied to reduce the dimensionality of the space (for all the technical details and a thorough discussion of the model, see Ploux & Victorri 1998).

As can be seen in figure 8, the semantic space of *jouer* obtained by this method is strikingly similar to the proxemic zoom we presented above, as far as structure is concerned: four branches for the same specific meanings (*s'amuser*, *risquer*, *tromper*, *imiter*), with the general '*pratiquer*' meaning in the center. It must be emphasized that a synonym can appear in different regions of the semantic space. For instance, one can see figure 8 that the verb *rouler*,

---

11. The algorithm used to compute the cliques can be found in Reingold *et al*, 1977. For a similar approach using also a graph of synonymy, see Warnesson, 1985.

12 More precisely, let u_1, u_2, …, u_n be the synonyms of the given unit, c_1, c_2, …, c_p the cliques containing the unit, and x_{ki} the coordinates of the cliques over the synonyms: $x_{ki} = 1$ if $u_i \in c_k$ and $x_{ki} = 0$ si $u_i \notin c_k$. Then the distance $d(c_k, c_l)$ between two cliques is given by the following formula :

$$d^2(ck, cl) = \sum_{i=1}^{n} \frac{x}{x_{\bullet i}} \left(\frac{x_{ki}}{x_{k\bullet}} - \frac{x_{li}}{x_{l\bullet}} \right)^2 \text{ where } x_{\bullet i} = \sum_{j=1}^{p} x_{ji} \quad x_{k\bullet} = \sum_{i=1}^{n} x_{ki} \text{ , and } x = \sum_{i=1}^{n} \sum_{j=1}^{p} x_{ji} .$$

a highly polysemic French verb, is present in two regions: in the center, with the meaning ‘to swing’, ‘to oscillate’, and in the *tromper* branch, with the meaning ‘to deceive’, ‘to trick’. This is one of the main qualities of the model: as expected (cf. introduction), words meet in the semantic space at different places, each place corresponding to a precise meaning.

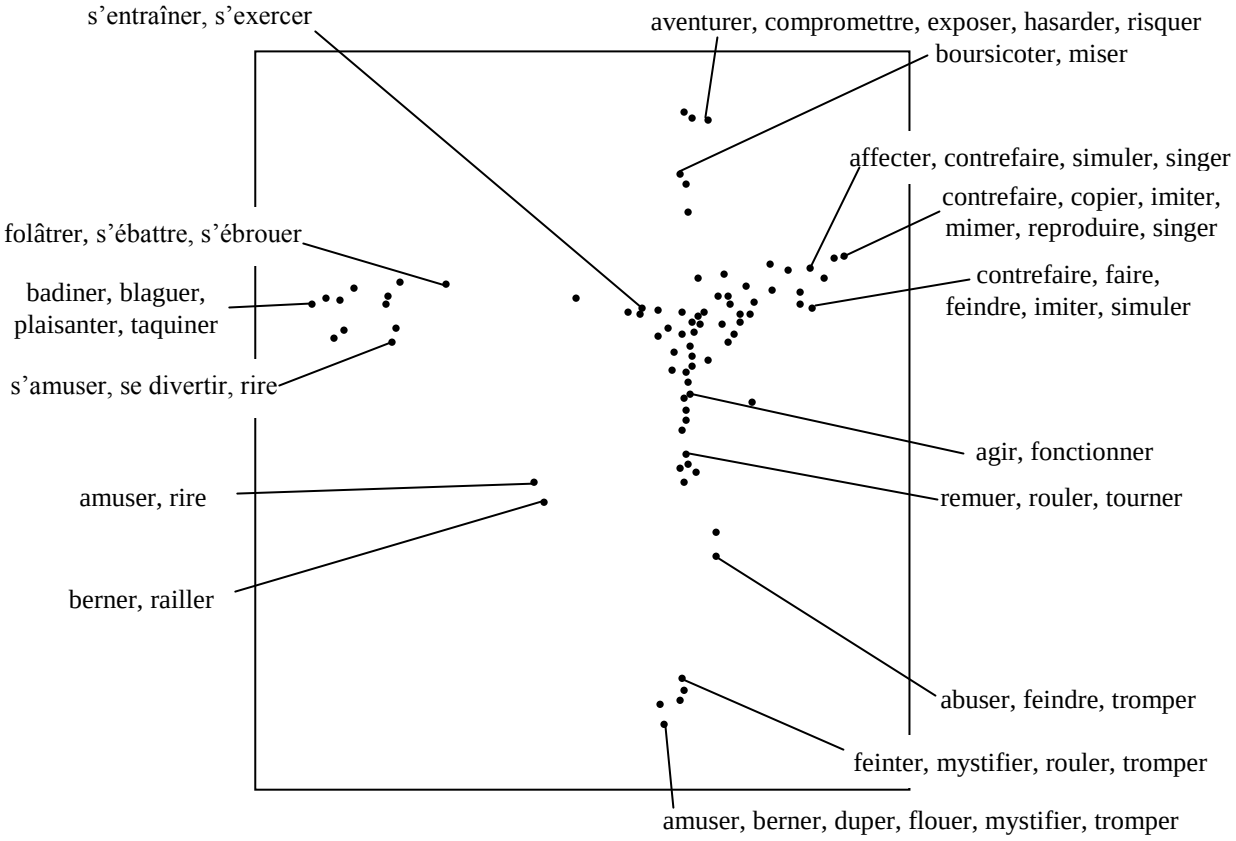

Figure 8. *The semantic space associated with* jouer

Thus, it is interesting to see if this method also gives the same general semantic dimensions as the proxemy method, when applied at the global level. Using the same method, we can obtain all the cliques of the complete graph Synoverbe (there are more than 25,000 cliques for less than 10,000 nodes) and compute the distance between any couple of them. Obviously, we cannot visualize a so large semantic space with the simple technique we used for local semantic spaces.

To solve this problem, we first build all the balls of a given radius centered on a clique, and we select the first hundred ones that have the highest density. In other words, we select the largest groups of strongly interconnected units, analogically corresponding to the centers of the largest cities (see the first part of this paper). Each high-density ball is assimilated to its center, which is associated with the whole set of synonyms corresponding to the union of the cliques included in the ball. Principal component analysis is then applied to the centers.

The results can be visualized on figure 9 and figure 10. Once again, the similarity with the results obtained with global proxemy is striking. We observe the same four main semantic axes structuring the meanings of French verbs. Figure 9 shows the projection of the global semantic space onto the first two dimensions. Three semantic zones appear, corresponding to three semantic axes that can be called 'positive' (*exciter, provoquer, produire*), 'negative' (*détruire, enlever, affaiblir*), and 'expressive' (*dire, montrer*). Figure 10 is a three-dimensional representation of the same space. It reveals the fourth semantic axis that can be called 'repulsive' (*disparaître, quitter, partir, sortir*).

On each figure, some information is given for a few representative balls, namely the content of the clique which is at the center of the ball, and the number of cliques and synonyms belonging to the ball. It can be observed that high density of cliques is not necessarily correlated with high density of synonyms. It means that this method actually brings out semantic zones where lexical units are highly interconnected. Moreover, as we already saw for the local semantic space of *jouer*, highly polysemic units cover very large regions of the global space. For instance the verb *sortir* extends over a large part of the 'expressive' zone and a large part of the 'repulsive' zone. Thus, this model gives a method to classify polysemic units, depending on the size of the associated region in the global representation: the most highly polysemic unit is not the most connected one (i.e. the node of highest degree in the graph), but rather the most extended one in the global semantic space.

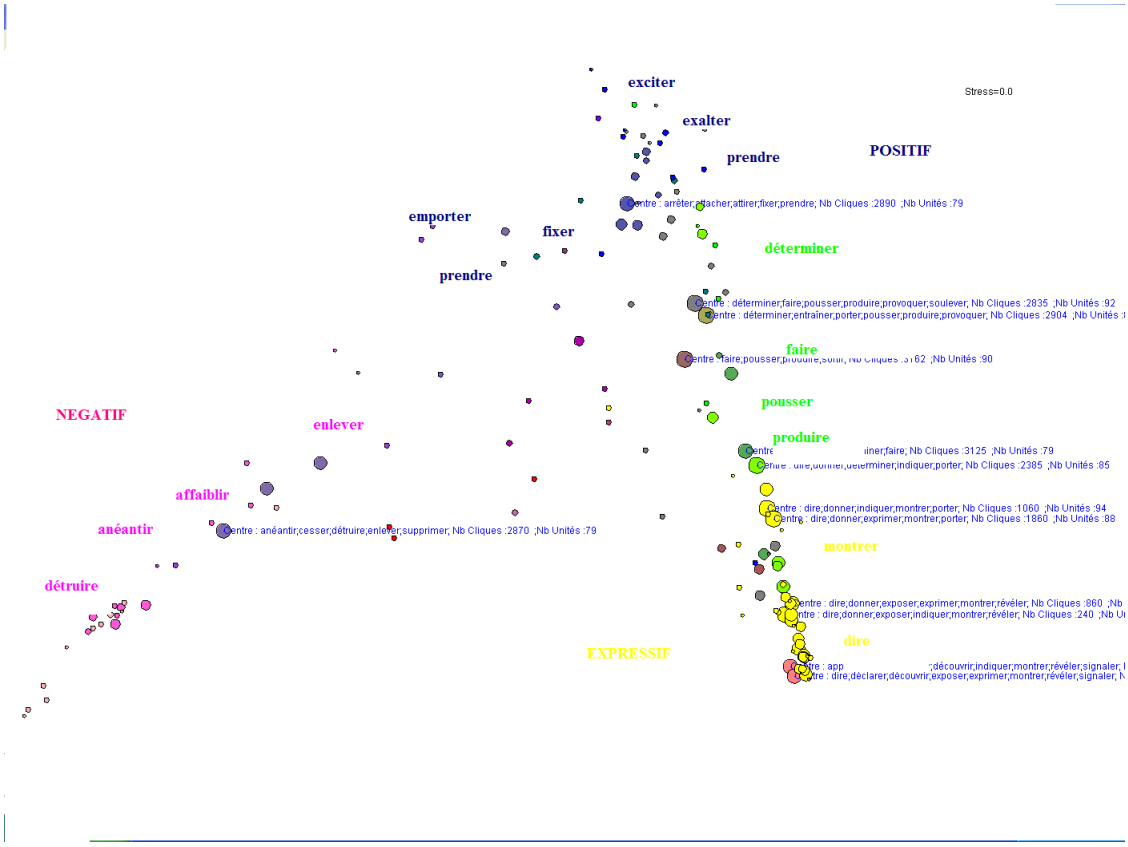

Figure 8. *Two-dimensional representation of the global semantic space of Synoverbe*

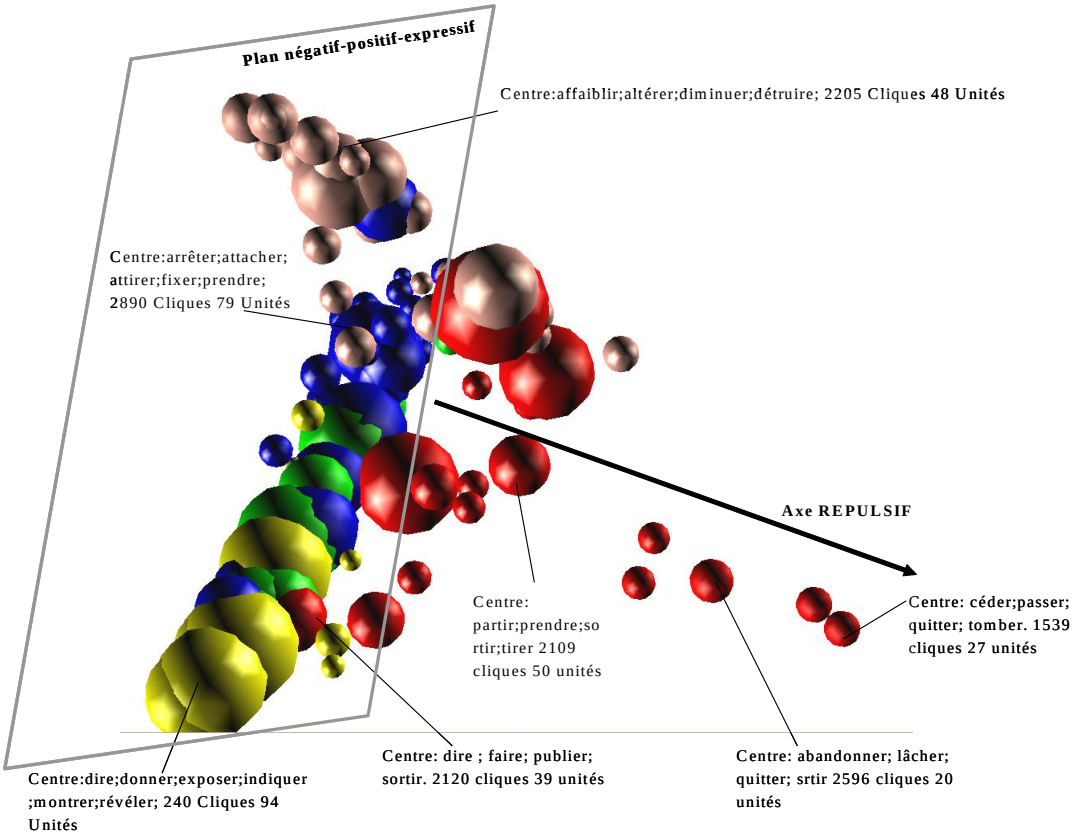

Figure 9. *Three-dimensional representation of the global semantic space of Synoverbe*

The convergence of two independent methods seems to prove that the features revealed by both of them are really intrinsic to the structural properties of the French verb lexicon. We have then at our disposal two tools to explore the hierarchical structure of small world graphs. As we said, the construction of 'semantic spaces' is not only interesting for lexical systems: it can prove very valuable for other 'semantic' graphs like the Web, as well as social graphs where the involved relationship depends more on conceptual factors than on geographical ones.

References

Abdellaoui N., Géométrisation et exploration du sens, *Mémoire de DEA,* EHESS, Paris, 2004.

Barabási A.-L., Albert R., Jeong H., and Bianconi G., Power-Law Distribution of the World Wide Web, *Science*, 287:2115a, 2000.

Barabási A. L., Jeong H., Neda Z., Ravasz E., Schubert A., and Vicsek T., Evolution of the social network of scientific collaboration, *Physica A*, 311(3-4):590-614, 2002.

Bermann A., Plemons R.J., *Nonnegative Matrices in the Mathematical Sciences*, Siam, 1994

Bréal M, *Essai de sémantique science des significations*, Hachette, 1897, rééd. 1983.

Darmesteter A., *La vie des mots étudiés dans leurs significations*, Paris, Éditions Champ Libre, 1887, rééd. 1979.

Duvignau, K., La métaphore, berceau et enfant de la langue. La métaphore verbale comme approximation sémantique par analogie dans les textes scientifiques et les productions enfantines (2-4 ans). *Thèse Sciences du Langage*. Université Toulouse Le-Mirail, Toulouse, 2002.

Duvignau K., Gaume B., Linguistic, Psycholinguistic and Computational Approaches to the Lexicon: For Early Verb-Learning. *Cognitive Systems,* 6 (1), 2003.

Duvignau, K., Gaume, B. First Words and Small Worlds: Flexibility and proximity in normal development. *Proceedings of the interdisciplinary conference* "*Architectures and Mechanisms for Language Processing*", Aix en Provence, 2004.

Fellbaum C. (ed.), *WordNet, an Electronic Lexical Database*, MIT Press, 1998.

Ferrer, R., Solé, R. V. The small world of human language. *Proceedings of The Royal Society of London. Series B, Biological Sciences*, 268(1482):2261-2265, 2001.

Gaume B., Duvignau K., Gasquet O., Gineste M-D., Forms of Meaning, Meaning of Forms, *Journal of Experimental and Theoretical Artificial Intelligence*, 14(1):61-74, 2002.

Gaume B., Analogie et Proxémie dans les réseaux petits mondes, *Regards Croisés sur l'Analogie, Revue d'intelligence Artificielle*, n° 5-6, Hermes, 2003.

Gaume B., Balades aléatoires dans les petits mondes lexicaux, Cepadues, 2004. http://dilan.irit.fr/.

Guare J., *Six degrees of separation : A play*, Vintage Books, New York, 1990

Huberman B. A., et Adamic L.A., Growth dynamics of the world-wide web, *Nature* 401:131, 1999.

Kleinberg J. M., Kumar R., Raghavan P., Rajagopalan S., Tomkins A. S., The Web as a Graph: Measurements, Models and Methods, *Lecture Notes in Computer Science*, 1627:1-17, 1999.

Milgram S., The small world problem, *Psychol. Today*, 2:60-67, 1967.

Newman M. E. J., The structure of scientific collaboration networks, *Proc. Natl. Acad. Sci. USA*, 98:404-409, 2001.

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

Ravasz E., Barabási A.L. Hierarchical Organization in Complex Networks, *Phys. Rev. E*, 67, 026112, 2003.

Reingold E.M., Nievergelt J., Deo N., *Combinatorial Algorithms, Theory and Practice*, Prentice-Hall, 1977.

Senata E., Nonnegative Matrices and Markov Chains, , Springer (2nd ed.), 1981

Sigman M., Cecchi G.A., Global organization of the Wordnet lexicon, *Proc. Natl. Acad. Sci.* 99(3):1742-7, 2002.

Venant F., Géométriser le sens, *RECITAL Conference*, Fès, Morocco, 2004.

Warnesson I., Applied Linguistics : Optimization of Semantic Relations by Data Aggregation Techniques, Journal of Applied Stochastic Models and Data Analysis, vol.1/2:121-143, 1985

Watts D.J., Strogatz S.H., Collective dynamics of 'small-world' networks. *Nature* 393:440-442, 1998.