Semantic associations and confluences in paradigmatic networks

Bruno Gaume
ERSS, 5 allées A. Machado,
F-31058 Toulouse cedex 1,
France
gaume[t]irit.fr
http://Prox.irit.fr

Karine Duvignau
Lab. Jacques Lordat,
5 allées A. Machado,
F-31058 Toulouse cedex 1,
France
duvignau[t]univ-tlse2.fr

Martine Vanhove
LLACAN,
7, rue Guy Môquet - BP 8,
94801 Villejuif Cedex,
France
vanhove[t]vjf.cnrs.fr

Summary

In this article, we hypothesize that some of the structural properties of paradigmatic graphs of the hierarchical small world type are to be found in all natural languages. Within this hypothesis of the universal structure of paradigmatic graphs, we explore a method for the automatic analysis of semantic groupings in order to distinguish, on typological and cognitive levels, which groupings are universal, and which are more limited geographically, genetically or culturally.

1. Introduction

Lexical semantics stems from a very long tradition, which underwent important developments with advances in cognitive studies, notably in the domain of metaphors (for example Lakoff 1987, Lakoff and Johnson 1980, Duvignau 2002), in work on semantic primitives (Goddard and Wierzbicka eds. 1994, Wierzbicka 1992), in historical linguistics (Wilkins 1996), as well as in studies on polysemy (Victorri and Fuchs 1996). However, linguistic typology has taken an interest in lexical semantics only recently (Viberg 1984, Koch 2001) because of long-standing suspicion of the object, the lexicon, which appeared both too vast to be grasped in its entirety, and too idiosyncratic in its organization, especially as regards polysemy. The distribution of semantic associations across languages or language families is nonetheless a particularly relevant linguistic phenomenon for inter-language comparative studies, even more so because polysemy is a universal phenomenon: all of the world's languages have terms, roots or stems, with or without expansions (derivational or qualifier morphemes, etc.) which may, each, express several different semantic notions. For example MOUTH and DOOR on one hand, and CHILD and FRUIT on the other, are expressed by the same word in many African languages. Our aim is to make an inventory of these semantic groupings, to analyze their structures, to categorize them and to measure their linguistic distribution: which languages group together MOUTH and DOOR? Which ones group together CHILD and FRUIT? What are the universal groupings shared by all languages, and which are more specific, and to which language families? In fact, recent advances in graph theory and corpus linguistics (Watts and Strogatz 1998, Gaume 2004, Gaume et al. 2004) make it possible to envision exploiting lexical data bases obtained by field linguists in order to study a given corpus in a unified manner, to measure the semantic proximity between lexical terms and to compare the semantic networks in languages. It is within this framework that the present article proposes a method for the automatic analysis of semantic groupings crosslinguistically.

In section 2.1, we will summarize the structural properties shared by most field graphs, so that in section 2.2 we may focus on lexical graphs, which will bring us in section 2.3 to voice a universality hypothesis concerning the structure of paradigmatic graphs. In section 3, using a stochastic flow approach in paradigmatic graphs, we will define the notion of confluence, and will then, in section 4, show how the notion of confluence in paradigmatic networks makes it possible to quantitatively measure the strength of semantic groupings between lexical units for a given language, which will lead us, in section 5, to imagine a robust automatic method for the analysis of semantic groupings across languages in order to determine which groupings are universal and which are more limited geographically, genetically or culturally. We will conclude in section 6 with the analysis of the advantages of and limits to the proposed approach.

2. The structure of French dictionary graphs

Graphs are widely used as a medium for presenting knowledge in (almost) all sciences. Created in the 18th century by Léonard Euler, graph theory was boosted by the arrival of computers, and is now picking up speed. In effect, machine calculation capacity makes it possible today to manage the large field[1] data graphs provided by human and social sciences (acquaintance networks, economic networks, geographical networks, semantic networks, etc.) as well as by engineering sciences (internet networks, electrical networks...) and by life sciences (neural networks, epidemiological networks, protein networks...). These graphs can contain up to several billion vertices and hundreds of billions of edges (Watts 1999, Newman 2003a, Newman 2003b).

In section 2.1 we will study the structure of field graphs in their entirety, and in section 2.2 we will focus on lexical graphs (language dictionaries, synonym dictionaries, thesauruses, semantic networks, large corpuses...) which will bring us in section 2.3 to formulate a universalistic hypothesis on the structure of paradigmatic graphs.

2.1. Properties of field graphs

Most of the large field graphs which are of interest to us here do not resemble random graphs, despite the fact that they are irregular[2]. Large field graphs possess both a rich local structure and a very "tight" global connectedness. This means that these graphs have a very particular topology, in which the relations between the local and global structures are completely different from what one finds with the graphs usually studied in graph theory (either random or regular). This explains the considerable interest that these recent findings have awakened in the scientific communities concerned. Indeed, one may imagine that these characteristics reflect the specific properties of the systems that these large field graphs represent, and that therefore the study of their structures may allow a fuller understanding of the phenomena from which they stem, as well as making it possible to better use the data thus represented: processing, modeling, structuring, indexing, information access, classifying, meaning extraction, visualizing…

[1] Field graphs are those found in practice, they are construed from field data. They are found in all field sciences. For example graphs of scientific collaborations (the vertices correspond to authors of scientific papers, and two authors A and B are linked if they have at least one publication in common).

[2] Regular graphs are what are usually studied in graph theory: all their vertices have the same degree of incidence (the same number of neighbors).

Formally, a ***graph***[3] G=(V,E) is obtained from a set V of ***vertices*** and a set E of pairs of vertices forming ***edges***. The vertices can represent objects and the edges relations of different natures between these objects. One usually illustrates these graphs by representing the vertices by points and by joining two points by a line if the two corresponding vertices form an edge: the only relevant information in such a case is not geometrical (the shape of the vertices or the placement of the points could be entirely different, all the while representing the same graph), but only of a relational type: whether the pairs of vertices constitute an edge or not.

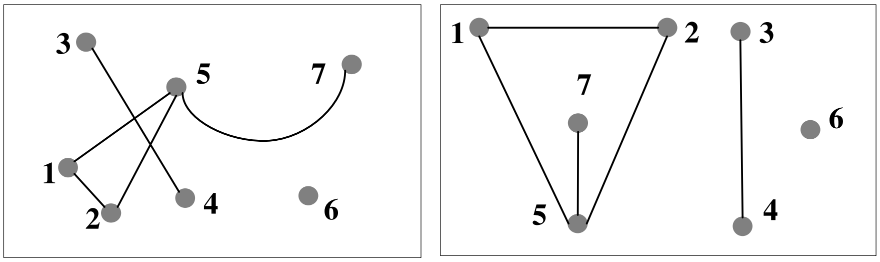


Fig. 1. G=(V,E) where V={1,2,3,4,5,6,7} and E={{1,2},{1,5},{2,5},{3,4},{5,7}}

The fact that the edge joining two vertices v_1 and v_2 is present in G will be written $\{v_1,v_2\}\in E$ (one then says that v_1 and v_2 are two vertices which are neighbors in G), the notation $v\in V$ indicating simply that v is a vertex in G. For any natural integer $m\neq 0$, a *path of length m in G* is an (m+1)-tuple $c = \langle v_0,\ldots,v_m\rangle$ such that $\forall i, 0\leq i<m : \{v_i,v_{i+1}\}\in E$, v_0 being the starting point, and v_m the end point. A graph G=(V,E) is said to be *connected* if $\forall x,y\in V$, there exists a path $\langle x, \ldots, y\rangle$ of finite length in G. The graph in Figure 1 is therefore not connected, and its greatest connected part is in the sub-graph formed by the vertices {1,2,5,7} with the edges {{1,2},{1,5},{2,5},{5,7}}.

The first explorations concerning large graphs, which were less regular than the laboratory graphs, were carried out by Erdös and Renyi (1960) who introduced and studied the notion of random graphs (a random graph is built starting from a set of isolated vertices, to which one randomly adds a given number of edges between the vertices) as a model for so called *field* graphs: large graphs (several thousand vertices and edges) from biochemistry, biology, technology, epidemiology, sociology, linguistics…

Since then, recent research in graph theory has brought to light a set of statistical characteristics shared by most field graphs; these characteristics define the class of graphs belonging to the *hierarchical small world* type. This is the case for the network of protein interactions for certain types of yeast (Jeong et al. 2001), of the neural network of the worm *Caenorhabditis elegans* (Watts and Strogatz 1998), of the graph of the World Wide Web (Barabasi et al. 2000), of that of a day's telephone calls in the US (Abello et al. 1999), of epidemiological graphs (Ancel et al. 2001), of scientific co-author graphs (Redner 1998), or of cinema collaborations (Watts and Strogatz 1998), or lexical networks taken from WordNet (Sigman and Cecchi 2002) or even of co-occurrences in a corpus of texts (Ferrer and Solé 2001)…

These graphs, like most field graphs, are sparse, which is to say that they have relatively few edges as compared to their number of vertices. In a graph with n vertices, the maximum number of edges is n(n-1)/2, i.e., approximatly $n^2/2$. Generally speaking, the number of edges

[3] For reasons of concision, we will only consider non oriented simple graphs here, which means that between two vertices, either there is no link, or that there is only one, which is non oriented (a link between two vertices is then called an edge).

in large field graphs is in the vicinity of n and not in that of n². For example, the graph of cinema collaborations[4] has 13 million edges, which may seem considerable, but which is quite small compared to the square of its vertices ($225000^2 \approx 5x10^{10}$).
Watts and Strogatz (1998) propose two indicators to characterize a large graph G which is connected and sparse: its L and its C.

– **L**= the average of the shortest paths between two vertices in G

– **C**= the rate of clustering, which is defined in the following way: given that a vertex v has K_v neighbors, whereas there is a maximum of $K_v(K_v-1)/2$ vertices that can exist between its K_v neighbors (which is what one obtains when each of the neighbors of v is connected to all the other neighbors of v). Let E_v be the number of edges between the neighbors of v (this number is thus necessarily lesser than or equal to $K_v(K_v-1)/2$). Let us posit that $C_v= E_v/(K_v(K_v-1)/2)$ which is therefore, for any vertex v, less than or equal to one.
The C of G is the average of the C_v on the vertices of G. The C of a graph is therefore always between 0 and 1. The more the C of a graph is close to 1, the more clusters it forms (zones dense in edges – *my friends are friends amongst themselves*). Applying these criteria to different types of graphs, Watts and Strogatz (1998) observe that:

1) **Field graphs** tend to have a low L (in general there is at least one short path between any two vertices).
2) **Field graphs** tend to have a high C, which reflects the tendency of two neighbors of a same vertex to be connected by an edge. For example in the World Wide Web[5], two pages that are linked to the same page have a relatively high probability of including links from one to the other.
3) **Random graphs** have a low L. When one randomly builds a graph having an edge density comparable to that of large field graphs, one obtains graphs where the L is low.
4) **Random graphs** have a low C: they are not made up of clusters. In a random graph, there is no reason why the neighbors of a same vertex would be more likely to be connected than any other two vertices, whence their low tendency to form clusters.

Watts and Strogatz (1998) propose to call *small world networks*[6], graphs which have these double characteristics (high C and low L) which they find in all the field graphs they have observed.

More recent studies (Ravasz and Barabási 2003) show moreover that most small world graphs have a hierarchical structure. The distribution of the vertices' degree of incidence[7] follows a *power law*. The probability P(k) that a vertex will have k neighbors decreases according to a power law $P(k) = k^{-\lambda}$ (Barabási et al. 2000, Kleinberg et al. 1999, Adamic 1999, Huberman

[4] The 225,000 syndicated American actors are the vertices, and there is an edge between vertex A and B if and only if the actors represented by the vertices A and B acted in the same movie.

[5] The vertices are the 10 billion pages available on the internet, and an edge is drawn between A and B if a hyperlink to B appears on page A or a hyperlink to A appears on page B.

[6] This term echoes that of *small world phenomena* by (Guare 1990; Kochen 1989; Milgram 1967) who studied social graphs in which two people A and B are in relation in the graph if A carries on such or such a type of relation with B (A knows B, A is regularly in touch with B, A worked in the same company as B…). These graphs were popularized by the slogan "six degrees of separation" (Guare 1990): for some of these graphs on a planetary scale, the average length of a path between two humans is around 6, which is very low compared to the billions of humans/vertices.

[7] The degree of incidence d(r) of a vertex $r \in V$ is the number of neighbors of the vertex r.

and Adamic 1999) where λ is a constant characteristic of the graph, whereas in the case of random graphs, it is Poisson's law which applies.

Table 1 below summarizes the four fundamental properties of field graphs.

GRAPHS	**global edge density**	**L: average measurement of the shortest paths** ***Global structure***	**C: measurement of the tendency to have edge dense sub-zones** ***Local structures***	**P(k) : Distribution of degrees** ***Incidence curve***
Random graphs	density here is an input parameter of the construction process	short paths *low L*	no clusters *low C*	Poisson's Law *the degree of the great majority of vertices is close to the degree average*
Field graphs	**P_1** sparse *few edges*	**P_2** short paths *low L*	**P_3** clusters *high C*	**P_4** Power law *without a scale: there is no significant average*

Table 1. The four fundamental properties of field graphs

In Table 1, the properties P_1, P_2, P_3, P_4 are extremely favorable for the low space time complexity of the processing algorithms. Furthermore, the property P3 expresses the communitarian character of field graphs whereas the property P_4 reflects their hierarchical organization. The properties P_3 and P_4 reveal the fundamental properties that these structures stem from, thus allowing greater understanding and usefulness of the data represented by the networks.

2.2 Lexical graphs

Following the works of Watts and Strogatz (1998), many articles appeared where the structures of the different field graphs are analyzed in an extremely wide array of domains (social sciences, life sciences, engineering sciences), but graph studies of linguistic origin remain very rare. We believe however that graphs from linguistics could help to better understand the structural properties of lexicons as well as comparative studies across languages.

There are several types of lexical networks, depending on the nature of the semantic relations which define the edges of the graph (the vertices represent the lexical units of a language – from some tens of thousands to some hundreds of thousands of elements, depending on the language and coverage of the corpus used). The three main types of relations are as follows:
– ***Syntagmatic* relations**, or rather of cooccurrence; one creates an edge between two words if one finds them near each other in a large corpus (typically at a maximum distance of two or three words (see Ide and Véronis 1998, Karov and Edeman 1998, Lebart and Salem 1994).
– ***Paradigmatic* relations**, notably synonymy; using a lexical database, such as the famous WordNet (Fellbaum 1999), one builds a graph in which two vertices are linked by an edge if the corresponding words show a synonymy relation (Ploux and Victorri 1998).
– ***Semantic proximity* relations**; these are less specific relations which may be taken into account both by the paradigmatic axis and by the syntagmatic axis. We created a graph of the French lexicon, defining the vertices in the following manner: an edge was created between the words A and B when one was found in the definition of the other in a general dictionary.

As general dictionary entries show the word's grammatical category (Verb, Noun, Adjective…) and also often definitions, examples, synonyms, and even antonyms, the vertices were therefore labeled according to their lexical category and the edges were labeled according to the type of relations they represented: it is therefore possible, according to one's needs, to limit the graph to certain lexical categories and/or combinatory relations: syntagmatic, paradigmatic and even logical-semantic relations (Gaume 2004).
All these graphs clearly belong to the hierarchical small world network type (P_1: Few vertices (sparse graph), and P_2: the average distance between two vertices is very small in the whole graph (low L), and P_3: community structuring (high C), and P_4: a hierarchical structure (the distribution of degrees of incidence ≈ power law). We will limit ourselves here to the study of paradigmatic graphs.

Generally speaking, if the dictionary definitions bear meaning, it is minimally through the network that they weave between the words constituting the entries. The idea of using this network (considered simply as a structured text source) was applied by Ide et al. (1990) through a neural network for removing ambiguities[8]. Our aim is to use this sort of hierarchical small world network by putting to work the hypothesis according to which zones which are dense in vertices ($P_3 \Rightarrow$ the communities) identify zones where meanings are close in their semantic capitals ($P_4 \Rightarrow$ the strongly connected vertices). We will illustrate our approach using two types of dictionaries: a standard dictionary, the Grand Robert[9] and DicoSyn[10], a dictionary of synonyms compiled from seven standard dictionaries (Bailly, Benac, Du Chazaud, Guizot, Lafaye, Larousse and Robert) from which the synonymic relations were extracted.
The dictionaries are represented by graphs whose vertices and edges can be defined in multiple ways. One of which consists in taking the dictionary entries as the graph's vertices, and in admitting the existence of an arc from a vertex A to a vertex B if and only if the entry B appears in the stemmed definition[11] of entry A. This is the starting position which we adopted. Indeed, this simple procedure makes it possible to extract from a standard dictionary[12] what we will henceforth call the **graph of the dictionary** in question.

Illustration around the vertex ÉCORCER [TO BARK]:

ÉCORCER [ekóRse] v. tr.; Dépouiller de son écorce (un arbre). Décortiquer, peler (le grain, le fruit)
TO BARK tr. v. To strip of its bark (a tree). Decorticate, peel (grains, fruit).

Fig. 2. Definition of ÉCORCER (to bark) after stemming – ROBERT –

[8] Recognizing a word's meaning from among several given in a dictionary for example, or distinguishing a word from among its various homographs.

[9] We had to undertake the considerable task of typing in, stemming and XML formatting in order to encode the graph extracted from the Grand Robert.

[10] This initial fusion task, carried out at the Institut National de la Langue Française (now ATILF: http://atilf.inalf.fr) produced a series of files, the data from which was assembled and homogenized through largely correcting the final file at the CRISCO laboratory.

[11] To label and stem the dictionary definitions we used Treetagger: http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html

[12] By constructing graphs from dictionary definitions, quantitative and structural studies seem apt for highlighting paradigmatic type relations (dictionary definitions being founded on meaning): if word A and word B belong to a same community (or to a same zone dense in edges), then replacing A by B in a sentence will only slightly change the meaning of the sentence 'the lumberjack strips the tree' → 'the lumberjack undresses the tree', even if the class of the predicative arguments is not always respected, thus creating semantic tensions.

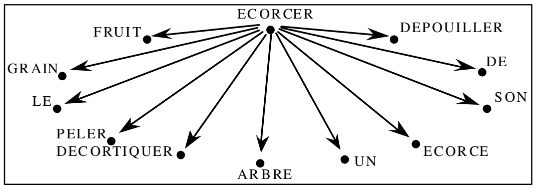


Fig. 3. Extract from the verb graph, around ECORCER (to bark) – ROBERT –

écorcer*: to bark;* fruit*: fruit;* grain*: grains;* le*: the;* peler*: peel;* décortiquer*: decorticate;* arbre*: tree;* un*: a;* écorce*: bark;* son*: its;* de*: of;* dépouiller*: strip*

By repeating this construction for each of the dictionary entries, one obtains the graph of the dictionary in question. If one extracts from the graph the sub-graph formed by the vertices which are verbs, this is what we obtain 'around'[13] the vertex denoted by the verb ECORCER (to bark):

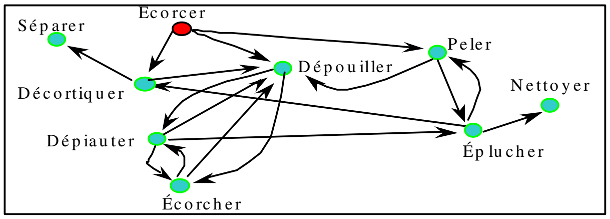


Fig. 4. Extract from the verb graph, around ECORCER (to bark) – ROBERT –

écorcer*: 'to bark';* séparer*: 'to separate';* décortiquer*: 'to decorticate';* dépiauter*: 'to skin';* écorcher*: 'to scrape';* dépouiller*: 'to strip';* éplucher*: 'to peel, pare';* nettoyer*: 'to clean';* peler*: 'to peel'.*

The definitions of DECORTIQUER (to decorticate), DEPOUILLER (to strip), PELER (to peel), SEPARER (to separate) … refer to other verbs absent from our schema for reasons of legibility (if one continues, one rapidly attains all the verbs in the dictionary). We therefore plotted on Figure 4 the vertices at distance 1 of ECORCER (to bark) and part of its vertices at distance 2 and 3. Once this oriented graph is obtained, our algorithms are applied to what we have called an *anonymous graph*[14], which is the non oriented version.

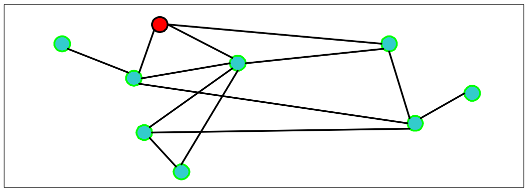


Fig. 5. Extract from the anonymous verb graph, around the vertices associated to ECORCER (to bark) – ROBERT –

[13] Which here is a 'topological around', i.e. the vertices linked to écorcer (to bark) by 'short' paths, topologically speaking = 'having few edges'.

[14] We use the term anonymous graph to insist on the fact that our algorithms apply only to this structure. For example, would it be possible, among several anonymous graphs, to distinguish their origins (standard dictionary, dictionary of synonyms, Internet, Protein network…)?

The graphs thus obtained are typical hierarchical small world graph networks. The hierarchical aspect with the presence of strongly connected vertices is a consequence of the hyperonymy role associated to the polysemy of certain vertices, whereas the high C (existence of zones dense in edges) reflects the role of the cohyponymy (Duvignau 2002, Duvignau et al. 2005b). For example in a standard dictionary (the GRAND ROBERT in our example), the verb CASSER (to break) is found in numerous definitions (ÉMIETTER (to crumble), FRAGMENTER (to fragment), DÉTÉRIORER (to deteriorate), RÉVOQUER (to revoke), ABROGER (to abrogate)…) whence the high incidence of the vertex CASSER (to break). Furthermore, one notes that there are numerous triangles, for example {CASSER, ÉMIETTER, FRAGMENTER} (BREAK, CRUMBLE, FRAGMENT), {CASSER, RÉVOQUER, ABROGER} (BREAK, REVOKE, ABROGATE) … which favor edge dense zones, or more precisely a high rate of C clustering. It is these edge dense zones which bring together the cohyponyms[15].

This is also valid for synonym dictionaries, for example, DicoSynVerbe[16] has 9043 vertices, it has 50,948 edges. On its greatest connected part (8,835 vertices), its L equals 4.17 and its C equals 0.39, which is typically a small world graph. The curve representing the incidence degree distribution of its vertices (Fig. 6) is characteristic of hierarchical small world networks (Ravasz and Barabási 2003) (in log-log it approximately forms a segment whose directing coefficient is equal to -2.01 with a determination coefficient of 0.96).

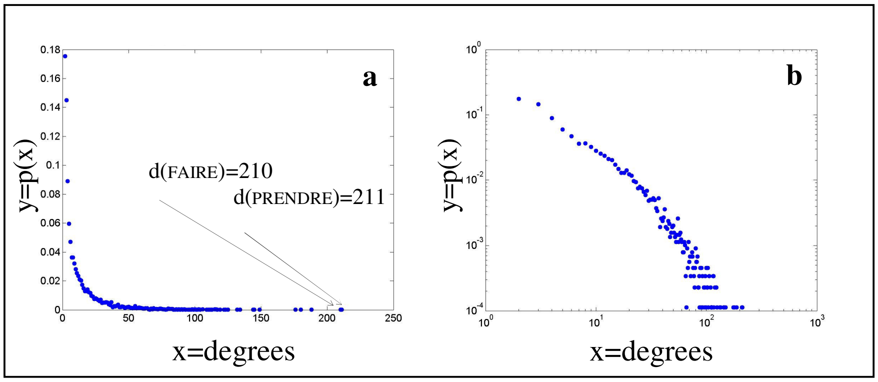


Fig. 6.a Incidence curve of the DicoSynVerbe vertices: 9,043 vertices; **.b** log-log

In Figure 6, the x axis represents degrees of incidence, while the y axis represents the incidence probability (the probability Y that by tracing a random vertex in an equiprobable manner, the vertex will have the incidence X). One also notes (Fig. 6a) that in DicoSynVerbe (as with all hierarchical small worlds), there are numerous vertices with low incidence, slightly fewer with rather higher incidence, fewer again with slightly higher incidence… with some high incidence vertices (the two words with the highest incidence in DicoSynVerbe are PRENDRE [TAKE] with d(PRENDRE)=211 and FAIRE [DO] with d(FAIRE)=210).

2.3 Hypothesis: the paradigmatic graphs of all natural languages are hierarchical small worlds

We formulate here a universality hypothesis on the structure of paradigmatic graphs:

[15] Cohyponyms: several words sharing a same meaning kernel with a common hyperonym: DÉSHABILLER (undress) and ÉPLUCHER (peel) are two interdomain cohyponyms of the hyperonym DÉPOUILLER (strip) whereas ÉPLUCHER (peel), PELER (peel, pare) are intra-domain cohyponyms (the domain of vegetables).

[16] DicoSynVerbe is the graph of verbs extracted from DicoSyn: there is an edge {A,B} if and only if the verbs represented by the vertices A and B are given as synonyms in DicoSyn.

Hypothesis (H1): the paradigmatic graphs of all natural languages are hierarchical small worlds

We are led to formulate this hypothesis (H1) for the following two reasons:

(1) As we saw in section 2.1, most field graphs resemble each other by their hierarchical small world structures[17].

(2) As we saw above in section 2.2, the language paradigmatic graphs that we built from standard dictionaries (the digitized Trésor de la Langue Française, Le Grand Robert), or from synonym dictionaries (Bailly, Benac, Du Chazaud, Guizot, Lafaye, Larousse, Robert, WordNet) or even from large corpuses (10 years of *Le Monde* daily newspaper) are typical hierarchical small worlds.

Moreover, studies on lexical acquisition by young children as well as on certain language pathologies (Duvignau et al 2004a-b, 2005a-b) point in the same direction as the hypothesis (H1). These studies show for example that the lexical approximations of young children (2-4 years old) of the type:

*« je **déshabille** l'orange »* 36 mois (l'enfant pèle une orange) [**PELER/DESHABILLER**]

*"I'm **undressing** the orange"* 36 months (the child is peeling the orange) **[PEEL/UNDRESS]**

*« maman, tu peux **coller** les boutons ? »* 36 mois (les boutons sont décousus, il faut les coudre) [**COUDRE/COLLER**]

*"Mommy, can you **glue** on the buttons?"* 36 months (the buttons are loose, they need to be sewn) **[SEW/GLUE]**

*« le livre est **cassé** »* 26 mois (le livre est déchiré) [**DECHIRER/CASSER**]

*"the book is **broken**"* 26 months (the book is ripped) **[RIP/BREAK]**

*« il faut la **soigner** la voiture »* 38 mois (il faut réparer la voiture) [**REPARER/SOIGNER**]

*"the car needs to be **treated**"* 38 months (the car needs repairing) **[REPAIR/TREAT]**

not only respect the edge dense zones which render vertex communities present in the dictionary graphs (peler↔déshabiller (peel↔undress) are in a common edge dense zone; the same is true of coudre↔coller (sew↔glue); déchirer↔casser (rip↔break); réparer↔soigner (repair↔treat)) but, furthermore, they respect the hierarchical aspect of these graphs (in general, children use those words which have the highest incidence: d(CASSER) (break) =192>d(DECHIRER) (rip) =72; d(COLLER) (glue) =74>d(COUDRE) (sew) =27; the number of neighbors of the child's word is generally higher than the number of neighbors of the word chosen by an adult without any pathologies for describing the same event, even if such is not

[17] The omnipresence of these structures in large field graphs of all origins (life sciences, human and social sciences, technology…) is all the more remarkable for the fact that the hierarchical small world structure is very rare as compared to the set of possible graphs (here rare is taken with its meaning in measurement theory: if all graphs are equiprobable, then by randomly choosing a graph among all possible graphs, the probability of obtaining a hierarchical small world is very close to zero).

always the case: d(DESHABILLER) (undress) =18=d(PELER) (peel) =18; d(SOIGNER) (treat) =49<d(REPARER) (repair) =69. To describe the same events, the average incidence of children's words is 117 whereas that of adults is 60.
Moreover, this phenomenon is found in several languages: French, Chinese, Portuguese, Korean, Ukrainian, as well as among patients with the first symptoms of Alzheimer's, also for several language families (Chen et al, 2006, Tonietto et al. 2006).

3. Confluences in hierarchical small world networks

We would now like to present Prox (http://Prox.irit.fr), an algorithm which calculates, on a hierarchical small world type graph, the structural confluences between vertices, which here are words, and which, as we will see in section 4, makes it possible to quantify the lexical semantic groupings for a given language. The important idea is to **calculate the confluence between two vertices from the graph as a whole**. This means that what is taken into account is not only the immediate neighbors of two vertices for the calculation of their confluence, but the whole graph as well. It is by applying this analysis method to dictionaries that we bring to light the structure of their graphs and "capture" their topological-semantic properties, among which one finds the proxemy which organizes the hyperonomy, the intra-domain cohyponymy, and the inter-domain cohyponymy within a continuum by quantifying the semantic groupings of lexical units.

3.1 Proxemy for confluence calculation

Notation:
If U is a line vector with dimension n, we will note $[U]_i$: the ith value of U;
If M is a n$\times$m matrix, then we will note for any i,k such that $1 \leq i \leq n$ and $1 \leq k \leq m$:

$[M]_{i\,k}$: the variable situated at the intersection of the ith line and the kth column of M;
$[M]_{i\,\bullet}$: the ith line vector of M;
$[M]_{\bullet\,k}$: the kth column vector of M.

Assume that we have a connected, symmetrical and reflexive graph, G=(V,E) with n=|V| vertices and m=|E| edges, and that on this graph a particle may at any time $t \in \mathbb{N}$ move around from vertex to vertex in a random fashion:

At instant t the particle is on a vertex $r \in V$.
When the particle is on a vertex $u \in V$ at instant t, it can only reach, at instant t+1, the vertices $s \in V$ such that $\{u,s\} \in E$ (meaning one of the neighbors of the vertex u). The particle moves from vertex to vertex at each instant by using the graph edges. Furthermore, we suppose that for every vertex $u \in V$, each of the edges incident to u is equiprobable.

Let $\hat{A}$ be the transition matrix at one step in the Markov chain corresponding to the random walk around the graph. This means that at each step, the probability of a transition from the

vertex $r\in V$ to the vertex $s\in V$ is equal to $[\hat{A}]_{r\,s}=[A]_{r\,s}/d(r)$ (where A is the adjacency matrix[18] of the graph G and d(r) the incidence degree[19] of the vertex r).
If the initial law of the Markov chain is given by the line vector P (which means that $[P]_r$ is the probability that the particle be on the vertex r at instant t=0) then $[P\hat{A}^t]_s$ is the probability that the particle be on the vertex s at instant t.

Let $F\subseteq V$ be a nonempty set of k vertices. Let us note P^F the vector of dimension n such that: $[P^F]_r=1/|F|$ if $r\in F$, and $[P^F]_r=0$ if $r\notin F$. If the initial law of the Markov chain is given by thevector P^F, then this corresponds to a random walk, beginning on one of the vertices of F, all equiprobable. Then $[(P^F)\hat{A}^t]_s$ is the probability that the particle be on vertex s at instant t when the particle begins the random walk equiprobably on one of the vertices of F at t=0. One notes that $[(P^{\{r\}})\hat{A}^t]_s=[\hat{A}]_{r\,s}$ which is therefore the probability that the particle be on vertex s at instant t when the particle begins the random walk on vertex r at instant t=0.

One proves[20] that if G=(V,E) is a connected and reflexive graph, then:

$$\forall r,s,u\in V,\ \lim_{t\to\infty}[\hat{A}^t]_{r\,s} = \lim_{t\to\infty}[\hat{A}^t]_{u\,s} = \frac{d(s)}{\sum_{x\in V}\{d(x)\}} \qquad (1)$$

This means that the probability for a particle, after a sufficiently long time t to be on vertex s does not depend on the initial vertex r or u, but only on vertex s, and is equal to $d(s)/\sum_{x\in V}\{d(x)\}$. However, two types of topological configurations can oppose the two vertices r and u in their relationships with vertex s.

> **Configuration 1)** the vertices r and s can be linked by a large number of short paths (r and s are strongly linked, there is strong *confluence* of paths from r to s);
> **Configuration 2)** the vertices u and s can be linked by only a few short paths (u and s are weakly linked: no *confluence* from u to s).

If formula (1) expresses that the probability for a particle, after a sufficiently long time t to be on vertex s does not depend on the initial vertex r or u, on the contrary, the dynamics towards this limit highly depends on the initial vertex and the type of confluence it entertains with vertex s. This means that the sequences $([\hat{A}^t]_{r\,s})_{0\leq t}$ and $([\hat{A}^t]_{u\,s})_{0\leq t}$ are not identical even though they converge towards the same limit $d(s)/\sum_{x\in V}\{d(x)\}$. Indeed, the trajectory dynamics of the particle in its random walk is entirely ruled by the topological structure of the graph: after t steps, every vertex s at a distance of t edges or fewer from the initial vertex can be reached. The probability of reaching vertex s at the tth step depends on the number of paths between the initial vertex and vertex s, on their lengths and on the structure of the graph around the intermediary vertices along the paths (the more paths there are, the shorter the paths, and the weaker the degree of the intermediary vertices, then the probability of reaching s from the initial vertex at the tth step is higher when t remains small). Thus there is a stronger confluence of the paths from vertices r towards s than from u towards s, whereas for a random walk with

[18] The adjacency matrix A of a Graph with n vertices G=(V,E) is the squared matrix nxn such that for every $r,s\in V$, $[A]r,s=1$ if $(r,s)\in E$ and $[A]r,s=0$ if $(r,s)\notin E$.

[19] Since we hypothesized that the graph is reflexive, then for every vertex $r\in V$, its incidence degree $d(r)\neq 0$ (in effect, reflexivity implies that every vertex is its own neighbor, which means that for every vertex $r\in V$ then $\{r,r\}\in E$: whence $d(r)\geq 1$).

[20] This is a consequence of the Perron Froebenius theorem (Bermann and Plemons 1994) because when the graph G=(V,E) is reflexive and strongly connected, the transition matrix $\hat{A}$ of the Markov chain associated to the random walk on graph G is then ergodic (Gaume 2004) (here the strong connectivity and reflexivity are necessary to prove the ergodicity).

a sufficiently short t length, one finds $[\hat{A}^t]_{r\,s} > [\hat{A}^t]_{u\,s}$. At the beginning of its random walk from the initial vertex, the particle has a higher probability of passing by those vertices with which the initial vertex entertains a high confluence relationship. For example, in DicoSynVerbe, the vertices DÉPIAUTER (to skin) and RÊVASSER (to daydream) have the same number of neighbors (d(DÉPIAUTER)=d(RÊVASSER)), and therefore, following (1),

$$\lim_{t\to\infty} [\hat{A}^t]_{\text{DESHABILLER DEPIAUTER}} = \lim_{t\to\infty} [\hat{A}^t]_{\text{DESHABILLER REVASSER}} = 6.3\text{x}10^{-5}.$$
$$\lim_{t\to\infty} [\hat{A}^t]_{\text{UNDRESS SKIN}} = \lim_{t\to\infty} [\hat{A}^t]_{\text{UNDRESS DAYDREAM}} = 6.3x10^{-5}$$

One can see however in Fig. 7 that the two sequences $([\hat{A}^t]_{\text{DÉSHABILLER DÉPIAUTER}})_{0\leq t}$ and $([\hat{A}^t]_{\text{DÉSHABILLER RÊVASSER}})_{0\leq t}$, are very different for a small t, which shows that the confluence from UNDRESS towards SKIN is stronger than that from UNDRESS towards DAYDREAM.

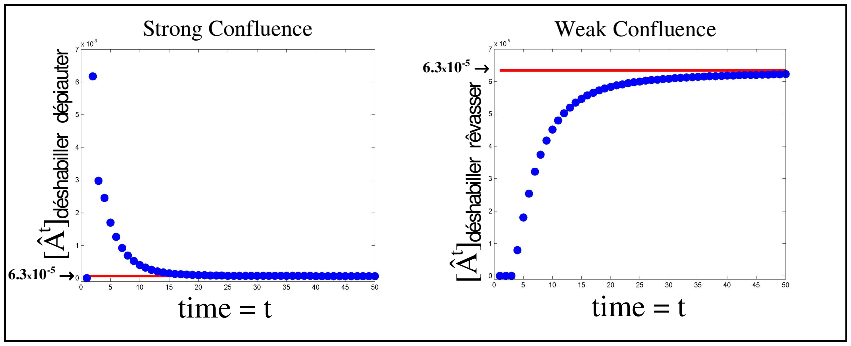


Fig. 7.a : $([\hat{A}^t]_{\text{DÉSHABILLER DÉPIAUTER}})_{0\leq t}$ et **7.b** : $([\hat{A}^t]_{\text{DÉSHABILLER RÊVASSER}})_{0\leq t}$ in DicoSynVerbe

Since L, the average length of the shortest paths, is small in a hierarchical small world, we know that two vertices are generally linked by at least one relatively short path. Thus we will choose t between L and 2L in order to generally reach almost all vertices from any given initial vertex, without however attaining the stationary probability of the Markov chain when t becomes too large.

3.2.Prox for disambiguating homonymy in dictionaries

In order to better perceive how Prox works, we will give here, as an example, a simple application for disambiguating homonyms in dictionaries.

In section 2.2 we did not mention a problem which is nonetheless fundamental in automatic language processing: disambiguation (Ide et al. 1990, Victorri et al. 1996).

For example, in *LE GRAND ROBERT* French dictionary, there are two distinct entries for the verb CAUSER:

CAUSER_1 « être la cause de. - Amener, apporter, attirer, déclencher, entraîner, faire, motiver, occasionner, produire, provoquer, susciter. *Causer un dommage. Causer du scandale. L'orage a causé de graves dommages aux récoltes…* »

"be the cause of. - Convey, bring, attract, set off, cause, do, motivate, occasion, produce, induce, provoke. *Cause damage. Cause a scandal. The storm caused heavy damage to the harvest…*"

CAUSER_2 « S'entretenir familièrement avec qqn. – Parler, converser, confabuler (vx), deviser, discuter. *Nous causons ensemble. Causer avec qqn…* »

"Have an informal conversation with someone. - Talk, converse, confabulate (archaic), devise, discuss. *We're chatting together. Chat with someone…*"

Thus, even if a French speaker naturally knows that in the definition of BAVARDER (gab):

BAVARDER « Parler beaucoup, longtemps ou parler ensemble de choses superficielles. - Parler; babiller, bavasser (fam.), cailleter, caqueter, causer, discourir, discuter, jaboter, jacasser, jaser, jaspiner (argot), lantiponner (vx), papoter, potiner. *Bavarder avec qqn …* »

"Talk a lot, for a long time or converse on superficial matters. - Speak; babble, blather on (colloquial), cackle, chat, discourse, discuss, gab, gabber, chatter, gossip. *Gab with someone…*"

the verb CAUSER refers to CAUSER_2 (chat), our procedure for constructing graphs (see section 2.2) cannot on its own disambiguate them. Thus, the procedure consist in creating a fictitious vertex CAUSER (which is not a dictionary entry since one only finds CAUSER_1 (cause) and CAUSER_2 (chat)) and to then add two vertices {CAUSER, CAUSER_1} and {CAUSER, CAUSER_2}. When CAUSER is found in the definition of a word such as BAVARDER (gab), then the vertex {BAVARDER, CAUSER} is added.

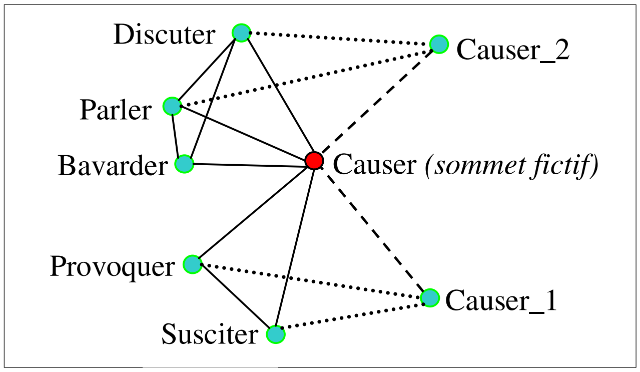


Fig. 8. CAUSER fictitious vertex

bavarder: 'gab'; *parler*: 'speak'; *discuter*: 'discuss'; *causer 1*: 'cause'; *causer 2*: 'chat'; *provoquer*: 'induce'; *susciter*: 'provoke'.

In Fig. 8 there are of course many edges and vertices that have been left out of our schema for reasons of legibility. The dotted edges {DISCUTER, CAUSER_2}, {PARLER, CAUSER_2} (discuss, chat), (speak, chat) are due to the fact that DISCUTER (discuss) and PARLER (speak) are in the definition of CAUSER_2 (chat), just as the edges {PROVOQUER, CAUSER_1} (induce, cause) and {SUSCITER, CAUSER_1} (provoke, cause), are in the definition of CAUSER_1 (cause).

We then apply Prox to the graph to obtain a matrix $[\hat{A}^t]$ as defined above (section 3.1).

$[\hat{A}^3]$	BAVARDER	PARLER	DISCUTER	CAUSER	CAUSER_1	CAUSER_2	PROVOQUER	SUSCITER
BAVARDER	0.325	0.165	0.165	0.189	0.025	0.075	0.025	0.025
PARLER	0.124	0.353	0.154	0.174	0.023	0.124	0.023	0.023
DISCUTER	0.124	0.154	0.353	0.174	0.023	0.124	0.023	0.023
CAUSER	0.081	0.099	0.099	0.379	0.086	0.081	0.086	0.086
CAUSER_1	0.025	0.030	0.030	0.201	0.351	0.025	0.166	0.166
CAUSER_2	0.075	0.165	0.165	0.189	0.025	0.325	0.025	0.025
PROVOQUER	0.025	0.030	0.030	0.201	0.166	0.025	0.351	0.166
SUSCITER	0.025	0.030	0.030	0.201	0.166	0.025	0.166	0.351

Table 2. for t=3

bavarder: 'gab'; *parler*: 'speak'; *discuter*: 'discuss'; *causer 1*: 'cause'; *causer 2*: 'chat'; *provoquer*: 'induce'; *susciter*: 'provoke'.

In Table 2, one observes that:
$[\hat{A}^3]_{\text{BAVARDER,CAUSER\_1(GAB,CAUSE)}}=0.025<[\hat{A}^3]_{\text{BAVARDER,CAUSER\_2(GAB,CHAT)}}=0.075$, which is as expected since the confluence of BAVARDER (gab) towards CAUSER_2 (chat) is stronger than from BAVARDER (gab) towards CAUSER_1 (cause), which is what makes it possible to disambiguate the two: assuming that a verb has k homonyms, there will therefore be vertices S, S_1, S_2, … S_k in the graph where S is the fictitious vertex. If there is an edge {R,S}, it will therefore be replaced by the edge $\{R,S_i\}$ where S_i is such that $[\hat{A}^3]_{R,S_i}=MAX_{0<z\leq k}\{[\hat{A}^3]_{R,S_z}\}$. One then deletes all the fictitious vertices from the graph in order to obtain a disambiguated graph as in Figure 9:

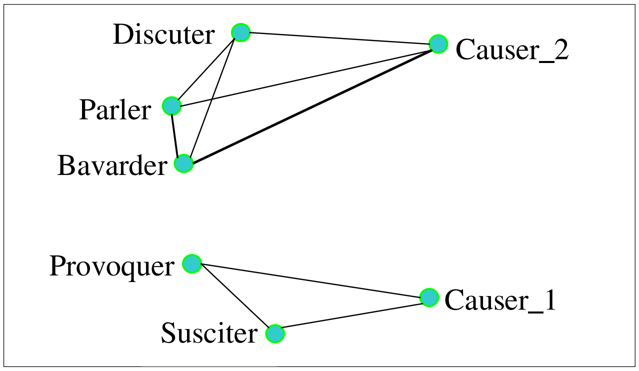


Fig. 9. Disambiguated graphe

discuter: 'discuss'; *parler*: 'speak'; *bavarder*: 'gab'; *causer 2*: 'chat'; *provoquer*: 'induce'; *susciter*: 'provoke': *causer 1*: 'cause'.

One then reapplies Prox, but to the disambiguated graph. Illustration: a list of the 100 vertices showing the strongest confluences with the verb ÉCORCER (to bark) (from the highest ranking: strong confluence with ÉCORCER (to bark) – to the lowest ranking: weakest confluence with ÉCORCER (to bark) –) calculated by Prox for t=3 on DicoSynVerbe.

1 →**ECORCER (to bark)**, 2 →DÉPOUILLER (strip), 3 →PELER (peel), 4 →TONDRE (mow, shear), 5 ÔTER (remove), 6 ÉPLUCHER (peel, pare), 7 RASER (shave), 8 DÉMUNIR (divest), 9 →DÉCORTIQUER (decorticate), 10 ÉGORGER (slit the throat of), 11 ÉCORCHER (skin), 12 ÉCALER (husk), 13 VOLER (steal), 14 TAILLER (prune), 15 RÂPER (grate), 16 PLUMER (pluck), 17 GRATTER (scrape), 18 ENLEVER (remove), 19 DÉSOSSER (bone), 20 DÉPOSSÉDER (dispossess), 21 COUPER (cut), 22 BRETAUDER (shear sloppily), 23 →INCISER (incise), 24 →GEMMER (tap), 25 →DÉMASCLER (remove first layer of cork), 26 →BAGUER (ring), 27 ÉVINCER (evict), 28 ÉTRILLER (curry), 29 ÉTRANGLER (strangle), 30 ÉPURER (purify), 31 ÉMONDER (blanch), 32 ÉCAILLER (scale), 33 ÉBRANCHER (prune, lop), 34 ÉBOURRER (remove tangles), 35 ÉBARBER (clip, trim), 36 TAMISER (sift), 37 TAILLADER (slash), 38 SPOLIER (despoil), 39 SEVRER (sever), 40 SCRUTER (scrutinize), 41 SCARIFIER (scar), 42 SALER (salt), 43 SAIGNER (bleed), 44 S'ÉPOILER (pluck one's self), 45 RÉVOQUER (revoke), 46 RUINER (ruin), 47 RETOURNER (turn over), 48 RETIRER (withdraw), 49 RANÇONNER (ransom), 50 RAISONNER (reason), 51 QUITTER (leave), 52 PRIVER (deprive), 53 PILLER (loot), 54 PERDRE (lose), 55 OUVRIR (open), 56 NETTOYER (clean), 57 MONDER (hull), 58 MARQUER (brand), 59 LIRE (read), 60 ISOLER (isolate), 61 GRUGER (swindle), 62 FUSILLER (shoot), 63 FRUSTRER (frustrate), 64 FOUILLER (search), 65 FILOUTER (cheat), 66 FAUFILER (tack, baste), 67 FAUCHER (reap), 68 EXPROPRIER (expropriate), 69 EXAMINER (examine), 70 ESTAMPER (stamp), 71 ESCROQUER (swindle), 72 ENTAMER (open, broach), 73 ENTAILLER (nick), 74 EFFEUILLER (thin out leaves), 75 DÉVÊTIR (undress), 76 DÉVELOPPER (develop), 77 DÉVASTER (devastate), 78 DÉVALISER (burglarize), 79 DÉTRÔNER (dethrone), 80 DÉTROUSSER (rob), 81 DÉSHÉRITER (disinherit), 82 DÉSHABILLER (undress), 83 DÉSENVELOPPER (remove the envelope of), 84 DÉSENCOMBRER (disencumber), 85 DÉSAVANTAGER (disadvantage), 86 DÉROBER (steal), 87 DÉPOURVOIR (render destitute), 88 DÉPIAUTER (skin), 89 DÉPECER (skin), 90 DÉNUER (deprive), 91 DÉNUDER (denude), 92 DÉNANTIR (deprive), 93 DÉMONÉTISER (demonetize), 94 DÉGARNIR (empty), 95 DÉGAGER (clear), 96 DÉFEUILLER (thin the leaves of), 97 DÉFAIRE (undo), 98 DÉCÉRÉBRER (decerebrate), 99 DÉCOURONNER (depose), 100 DÉCHAUSSER (expose/remove shoes), …

Fig. 10. Proxemy of ÉCORCER (to bark) from DicoSynVerbe at t=3

In DicoSynVerbe the vertex ÉCORCER (bark) has 8 synonyms: {BAGUER, DÉCORTIQUER, DÉMASCLER, DÉPOUILLER, GEMMER, INCISER, PELER, TONDRE} (ring, decorticate, remove first layer of cork, strip, tap, incise, peel, mow/shear). In Figure 10, the number preceding each verb gives its rank according to its proxemy with ÉCORCER (to bark) and the neighbors of ÉCORCER (to bark) are preceded by an arrow→. One sees that after ÉCORCER (to bark) itself, DÉPOUILLER (to strip) which appears at the top of the list (being the one that entertains the strongest confluence with ÉCORCER (to bark) according to Prox) is a hyperonym of the verb ÉCORCER (to bark). The proxemy calculated by the Prox algorithm thus organizes, within a continuum, the notions of intra-domain cohyponymy (through the vertices which are the most 'Prox') and of inter-domain cohyponymy (through the vertices which are a little less 'Prox'), (Duvignau and Gaume 2004b). The introduction of the notion of *proxemy* makes it possible to highlight the meaning shift that takes place between a word in a quasi-synonymous relation (intra-domain cohyponyms) towards a word in a metaphorical relation (inter-domain cohyponyms) the more the proxemy to the reference term diminishes.

4. Confluence and semantic associations

The polysemy of lexical units is a universal phenomenon in all natural languages which is difficult to grasp from a cognitive point of view (how relevant meanings are accessed), in the domain of automatic language processing (how to disambiguate in cotexts), and in semantics (how the different meanings of a given term are organized on the level of the linguistic system). This last point leads to the question of the possible existence of universals of semantic groupings (also called semantic parallelisms, semantic derivation or semantic associations). For example, in her work "From Etymology to Pragmatics. Metaphorical and

Cultural Aspects of Semantic Structure", Eve Sweetser (1991: 21) brought to light the strong links between the lexicon of physical perception and that of knowledge in Indo-European languages: *"Deep and pervasive metaphorical connections link our vocabulary of physical perception and our vocabulary of intellect and knowledge".*

In French, for example (and English), this lexical link between physical perceptions and knowledge is common practice. To illustrate, below are six text extracts from the World Wide Web where the verbs SENTIR (feel), ENTENDRE (hear), VOIR (see) can easily be replaced in their contexts by the verbs COMPRENDRE (understand) or SAVOIR (know) all the while keeping the main meaning of each of the sentences.

- http://www.modia.org/etapes-vie/jeunes/teamim.html : *« -faire les pauses en conséquence lors de la lecture, -**sentir** ce que devient le sens de la phrase avec ces pauses diverses, -réfléchir au sens que cela donne à la phrase, »*
  *("-pause accordingly while reading, -**feel** what the meaning of the sentence becomes with the different pauses,- reflect upon the meaning this gives to the sentence,")*
- http://www.leseditionsdeminuit.fr/titres/2002/nepastoucher.htm : *« … des textes capables d'extirper et faire **sentir** le sens profond du temps que nous vivons. »*
  *("... texts capable of extracting and making one **feel** the deep meaning of the times we live in.")*
- http://globetrotter.berkeley.edu/people/Karekezi/karekezi-con.f2.html : *« Je voulais **voir** ce que ça veut dire. Je voulais **voir** ce qu'une femme rwandaise, juriste, pouvait apporter à Clémentine et aux autres. Parce qu'elle n'était pas une exception. Je voulais **voir**. »*
  *(I wanted to **see** what it meant. I wanted to **see** what a Rwandese woman, a jurist, could give Clémentine and the others. Because she wasn't an exception. I wanted to **see**.")*
- http://forum.decroissance.info/viewtopic.php?t=882& : *« Radicaliser son propos en proposant le pire n'a qu'une finalité rhétorique pour faire **voir** le sens du capitalisme. »*
  *("To harden one's discourse by proposing the worst has only rhetorical finality to make people **see** the meaning of capitalism.")*
- http://www.stopsuicide.ch/5/marches/texte%207.pdf : *« Puissions-nous **entendre** ce que l'Autre si près de nous ne peut pas dire. »*
  *("May we **hear** what the Other, so close to us, cannot say.")*
- http://www.theatre-odeon.fr/fichiers/t_downloads/file_70_dp_10.pdf#search=%22%22entendre%20le%20sens%22%20le%20petit%20prince%22 : *« en nous faisant **entendre** le sens de certaines paroles … »*
  *("By making us **hear** the meaning of certain words...")*

These semantic groupings in French between physical perception and knowledge are also measurable in French language dictionary graphs. Figure 11 below illustrates the list of the 100 vertices with the strongest confluence relationships with the verb SAVOIR (know) (from the highest ranking: the strongest confluence with SAVOIR (know) – to the lowest ranking: the weakest confluence with SAVOIR (know) –) calculated by Prox at t=3 on DicoSynVerbe.

1 →CONNAÎTRE (know), 2 →**SAVOIR** (**know**), 3 →ÊTRE INFORMÉ DE (be informed of), 4 →ÊTRE AU COURANT(be aware of), 5 →POUVOIR (be able to), 6 →ÊTRE AVERTI (be informed), 7 →ÊTRE AU FAIT (be aware of), **8 →VOIR** (see), 9 →APPRENDRE (learn), 10 →COMPRENDRE (understand), 11 →IMAGINER (imagine), 12 →POSSÉDER (possess), 13 →S'ATTENDRE (expect), 14 →PRENDRE GARDE (be attentive to), 15 PENSER (think), 16 APERCEVOIR (perceive), 17 JUGER (judge), 18 CONCEVOIR (conceive), 19 CROIRE (believe), 20 PÉNÉTRER (penetrate), 21 CONSIDÉRER (consider), 22 ÊTRE APTE (be apt), **23 SENTIR** (feel), 24 PRENDRE (take), **25 ENTENDRE** (hear), 26 PERCEVOIR (perceive), 27 DEVINER (guess), 28 ÊTRE EN MESURE DE (be able to), 29 ÊTRE CAPABLE DE (be capable of), 30 APPRÉCIER (appreciate), 31 S'APERCEVOIR (notice), 32 ÊTRE AUTORISÉ À (have permission to), 33 SE FIGURER (figure), 34 ENTREVOIR (get a glimpse of), 35 ÊTRE EXPERT (be expert at), 36 S'OCCUPER (take care of), 37 DISCERNER (discern), 38 ESTIMER (estimate), 39 EMBRASSER (embrace), 40 CONSTATER (see, notice), 41 ÉPROUVER (feel), 42 APPRÉHENDER (grasp), 43 ÊTRE CALÉ (be good at), 44 ÊTRE À MÊME DE (be able to), 45 ÊTRE EN PASSE DE (be in the process of), 46 PRÉVOIR (foresee), 47 COMPTER (count), 48 ÊTRE SAVANT (be knowledgeable), 49 ÊTRE COMPÉTENT (be competent), 50 ATTENDRE (wait), 51 PRESSENTIR (foresee), 52 RECONNAÎTRE (recognize), 53 PRATIQUER (practice), 54 ÊTRE FERRÉ (be good at), 55 DÉCOUVRIR (discover), 56 ESPÉRER (hope), 57 EXPÉRIMENTER (experiment), 58 AVOIR LA PRATIQUE (be practiced at), 59 AVOIR L'USAGE (have the use of), 60 ASSAVOIR (make known), 61 SUBIR (undergo), 62 S'IMAGINER (imagine), 63 AVOIR CONNAISSANCE (have knowledge of), 64 SE PRÉOCCUPER (worry about), 65 RESSENTIR (feel), 66 SE REPRÉSENTER (imagine), 67 REMARQUER (notice), 68 AVOIR LA CAPACITÉ (be able to), 69 TROUVER (find), 70 SAISIR (seize), 71 ENDURER (endure), 72 ÊTRE À PORTÉE DE (be able to), 73 AVOIR LE DROIT (have the right to), 74 AVOIR LA PERMISSION (have permission to), 75 SUPPORTER (bear), 76 AVOIR (have), 77 ÊTRE TAILLÉ POUR (be made for), 78 ÊTRE EN SITUATION DE (be in a situation to), 79 AVOIR LA POSSIBILITÉ DE (have the possibility of), 80 REGARDER (look at), 81 PRÉSUMER (presume), 82 SONGER (wonder), 83 SE SOUVENIR (remember), 84 SE DOUTER (expect), 85 AVOIR LE CHOIX (have the choice), 86 AVOIR LA LATITUDE (have the latitude), 87 TENIR DE (take after), 88 ESCOMPTER (count on), 89 NOTER (note), 90 SUPPOSER (suppose), 91 ÊTRE EN ÉTAT DE (be in a state to), 92 SOUPÇONNER (suspect), 93 CHERCHER (look for), 94 VOIR VENIR (see something coming), 95 AVOIR SOIN (be careful of), 96 ÊTRE SUSCEPTIBLE DE (be susceptible of), 97 S'ÉVERTUER (persevere), 98 FAIRE ATTENTION (pay attention), 99 CONJECTURER (conjecture), 100 EXAMINER (examine), …

Fig. 11. Proxemy of SAVOIR (know) from DicoSynVerbe at t=3

In DicoSynVerbe the vertex SAVOIR (know) has 13 synonyms, the neighbors of SAVOIR are preceded by an arrow→ and the number that precedes each verb is its rank according to its proxemy with SAVOIR (know).
In Figure 11, if a verb Y_1 is ranked k^{th}, and another verb Y_2 is ranked $k+1^{th}$, it is because $[\hat{A}^3]_{\text{SAVOIR } Y_1} \geq [\hat{A}^3]_{\text{SAVOIR } Y_2}$, meaning that when the particle begins its random walk along the edges of the DicoSynVerbe graph at instant t=0 on the vertex SAVOIR (know), the probability that a particle be at instant t=3 on the vertex Y_1 is greater than or equal to the probability that it be on vertex Y_2 at instant t=3 (meaning that the confluence from SAVOIR (know) towards Y_1 is greater than or equal to the confluence from SAVOIR (know) towards Y_2).

One may note that the verbs VOIR (see), SENTIR (feel) and ENTENDRE (hear) are ranked respectively 8th, 23rd and 25th, which is very high considering that the DicoSynVerbe has 9043 verbs (these three verbs are paradigmatically ranked Top_3_per_1000 for the verb SAVOIR (know)). This tells us that in DicoSynVerbe there is a strong confluence from SAVOIR (know) towards VOIR (see), SENTIR (feel) and ENTENDRE (hear), even despite the fact that SENTIR (feel) and ENTENDRE (hear) are not directly connected to SAVOIR (know).

If we now consider the matrix $\hat{A}^3$ as the 9043x9043 matrix of the coordinates of the 9043 line vectors $([\hat{A}^3]_{x\bullet})_{x\in V}$ in $\mathbb{R}^{9043}$, this perspective allows us to embed the graph G=(V,E) into $\mathbb{R}^{9043}$, where a given vertex r∈ V has as coordinates in $\mathbb{R}^{9043}$the line vector $[\hat{A}^3]_{r\bullet}$.

The idea is that two vertices r and s with the coordinates $[\hat{A}^3]_{r\bullet}$ and $[\hat{A}^3]_{s\bullet}$ in $\mathbb{R}^{9043}$, will be all that much closer in $\mathbb{R}^{9043}$ if their relationships to the graph as a whole are similar.

If one then projects the matrix $\hat{A}^3$ in $\mathbb{R}^3$ by the technique of Principal Component Analysis (PCA) and if one sees what happens around the vertex SAVOIR (know), we obtain the form illustrated in Figure 12, where one well perceives[21] that the verbs VOIR (see), SENTIR (feel) and ENTENDRE (hear) are very close to the verb SAVOIR (know) because numerous very short paths link the verb SAVOIR (know) to these three verbs (the entire French lexicon is available at http://Prox.irit.fr).

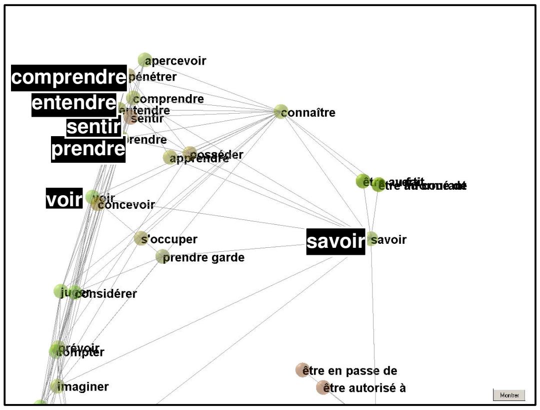


Fig. 12. Around SAVOIR (know) in DicoSynVerbe at t=3 (accessible at: *http://Prox.irit.fr*)

In the same manner, Figure 13 below illustrates the list of the 100 vertices with the strongest confluence relations with the verb COMPRENDRE (understand) (from the highest ranked: strong confluence with COMPRENDRE (understand) – to the lowest ranked: weakest confluence with COMPRENDRE (understand) –) calculated by Prox at t=3 for DicoSynVerbe. One notes that the verbs VOIR (see), SENTIR (feel) and ENTENDRE (hear) are ranked respectively 3rd, 12th and 19th, which is very high given the 9043 verbs present in DicoSynVerbe (these 3 verbs are in the paradigmatic Top_3_per_1000 of the verb COMPRENDRE (understand). If one looks at what takes place around the vertex COMPRENDRE (understand), we obtain the form illustrated in Figure 14 where one well perceives that the verbs VOIR (see), SENTIR (feel) and ENTENDRE (hear) are very close to the verb COMPRENDRE (understand) because numerous very short paths link COMPRENDRE (understand) to these three verbs.

[21] This principle of the perception of the topological-semantic structures is accessible at *http://Prox.irit.fr* and is formally described in Gaume (2006) and Gaume and Mathieu (2006) with several applications for cognitive psychology: language acquisition and pathologies (Duvignau et al. 2005a, Duvignau et al. 2004, Duvignau and Gaume 2004b) and the ergonomics of information access interfaces: dictionaries and the World Wide Web (Gaume and Duvignau 2004).

1 →**COMPRENDRE** (understand), 2 →CONNAITRE (know), **3 →VOIR** (see), 4 →DECOUVRIR (discover), 5 →SAISIR (grasp), 6 →PENETRER (penetrate), 7 →DEVINER (guess), 8 →PRENDRE (take), 9 →ATTEINDRE A (attain to), 10 →RENFERMER (enclose), 11 →ENFERMER (close in), **12 →SENTIR** (feel), 13 →DECHIFFRER (decipher), 14 →APERCEVOIR (glimpse), 15 →TROUVER (find), 16 →EMBRASSER (embrace), 17 →COMPTER (count), 18 →CONSISTER (consist), **19 →ENTENDRE** (hear), 20 →CONTENIR (contain), 21 →IMAGINER (imagine), 22 →REVELER (reveal), 23 →REPERER (notice), 24 →PERCER (pierce), 25 →REMARQUER (notice), 26 →CONCEVOIR (conceive), 27 →SE COMPOSER (compose one's self), 28 LIRE (read), 29 →COMPORTER (contain), 30 →ADMETTRE (admit), 31 →SAVOIR (know), 32 →ENGLOBER (surround), 33 →APPRENDRE (learn), 34 →INCLURE (include), 35 →ENCLORE (shut in), 36 PERCEVOIR (perceive), 37 →S'APERCEVOIR (notice), 38 →DECODER (decode), 39 →APPREHENDER (grasp), 40 →ENVELOPPER (envelope), 41 →MELANGER (mix), 42 →S'EXPLIQUER (explain to one's self), 43 →DEMELER (unravel), 44 →INCORPORER (incorporate), 45 →INTERPRETER (interpret), 46 →APPRECIER (appreciate), 47 →MELER (tangle), 48 →ENTRER (enter), 49 →INTEGRER (integrate), 50 PENSER (think), 51 JUGER (judge), 52 →ENTREVOIR (glimpse), 53 →SUIVRE (follow), 54 →IMPLIQUER (imply), 55 →PIGER (get), 56 →ASSIMILER (assimilate), 57 DISCERNER (discern), 58 →APPROUVER (approve), 59 INTRODUIRE (introduce), 60 →REALISER (realize), 61 DECRYPTER (decipher), 62 →SE RENDRE COMPTE (realize), 63 REUNIR (reunite), 64 →TRADUIRE (translate), 65 ENTOURER (surround), 66 →GROUPER (gather), 67 →FAIRE ENTRER (make enter), 68 DISTINGUER (distinguish), 69 TENIR (hold), 70 JOINDRE (join), 71 RECONNAITRE (recognize), 72 →SE REPRESENTER (represent to one's self), 73 FAIRE (do), 74 →MORDRE (bite), 75 PRESSENTIR (have a presentment), 76 ETRE FORME DE (be made up of), 77 ETRE CONSTITUE DE (be composed of), 78 DECELER (detect), 79 CROIRE (believe), 80 NOTER (note), 81 →SE METTRE A (begin), 82 CONSTATER (note), 83 MARQUER (mark), 84 SURPRENDRE (surprise), 85 FLAIRER (smell something out), 86 ASSOCIER (associate), 87 ENSERRER (clasp), 88 DEBROUILLER (make do), 89 EXPLIQUER (explain), 90 MONTRER (show), 91 UNIR (unite), 92 CONSIDERER (consider), 93 TOUCHER (touch), 94 EMPRISONNER (imprison), 95 PREVOIR (foresee), 96 REGARDER (look at), 97 EPROUVER (feel), 98 OBSERVER (observe), 99 ESTIMER (estimate), 100 ACCEPTER (accept), …

Fig. 13. Proxemy of COMPRENDRE (understand) at t=3 from DicoSynVerbe at t=3

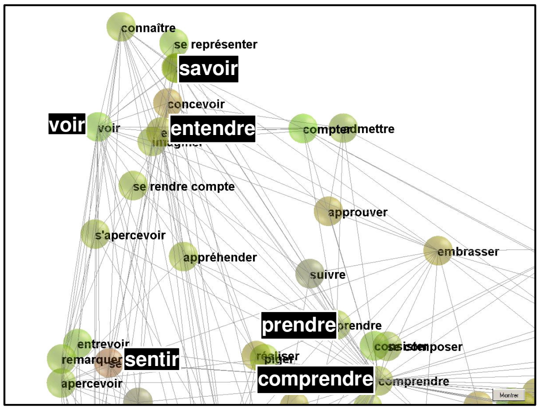


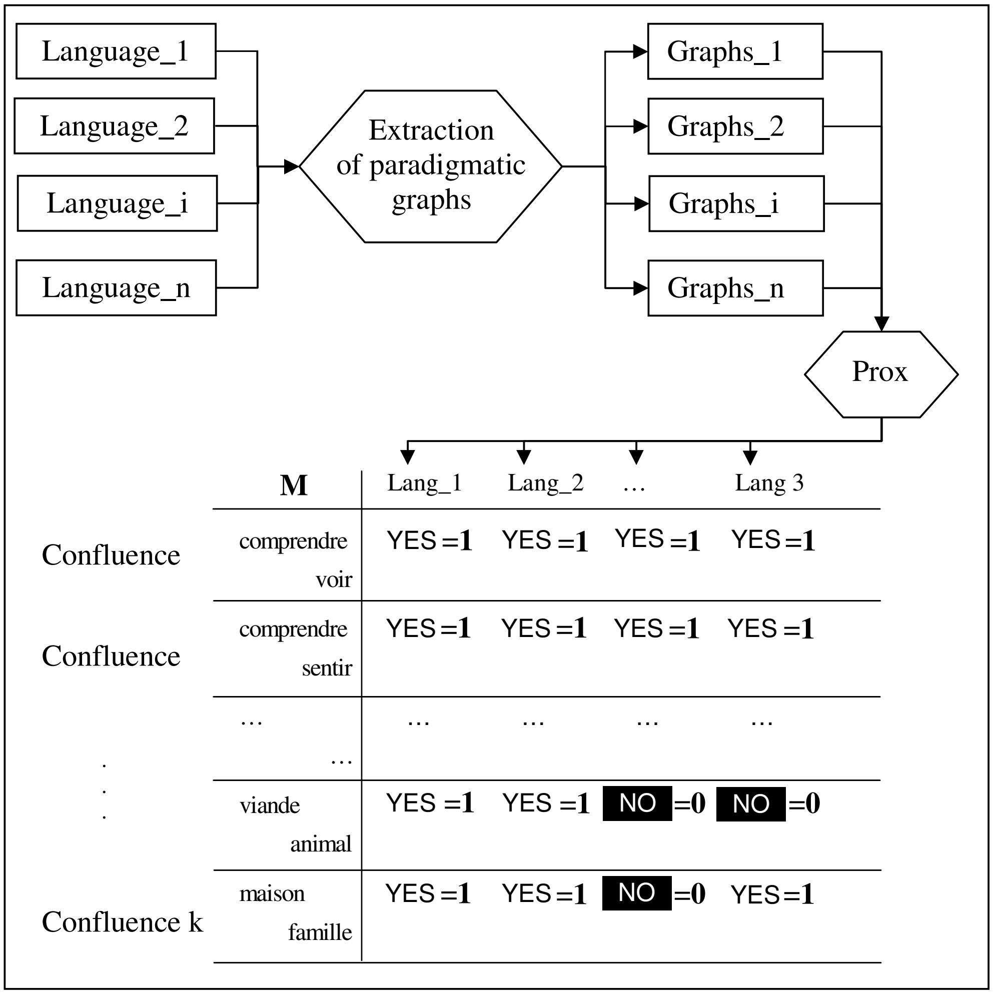
Fig. 14. Around comprendre (understand) in DicoSynVerbe at t=3

Since the works done by Viberg and then Sweetser, some studies have been carried out on the links between perception and knowledge in various languages (for example in Australian languages, Evans and Wilkins 2000) but the question remains open today as to the universality of these semantic links between physical perception and knowledge (see Vanhove, this volume).

One may ask the same questions about other semantic associations: VIANDE⇆ANIMAL (meat/animal), MAISON⇆FAMILLE (house/family), PORTE⇆BOUCHE (door/mouth), ENFANT⇆FRUIT (child/fruit), IMITER⇆VOLER (imitate/steal)…: are these associations symmetrical, are they universal, or, on the contrary, are they more limited geographically, genetically or culturally, and if so, which language families are they limited to? (see Boyeldieu, this volume).

5. A typology of languages based on co-confluence in paradigmatic graphs

We saw in section 3.1 that applying Prox to a hierarchical small world type graph makes it possible to quantify confluences between vertices. We then saw in section 4 above that when the graph is paradigmatic, then the notion of confluence allows the quantification of semantic associations of the type PERCEPTION⇆CONNAISSANCE (perception/knowledge) for a given language. Our hypothesis (H1): *the paradigmatic graphs of all natural languages are hierarchical small worlds*, gives rise to the possibility of a semi-automatic and systematic research on crosslinguistic semantic associations based on their paradigmatic graphs.
Figure 15 illustrates this method:



	M	Lang_1	Lang_2	…	Lang 3
Confluence	comprendre voir	YES =1	YES =1	YES =1	YES =1
Confluence	comprendre sentir	YES =1	YES =1	YES =1	YES =1
	… …	…	…	…	…
. . .	viande animal	YES =1	YES =1	NO =0	NO =0
Confluence k	maison famille	YES =1	YES =1	NO =0	YES =1

Fig. 15. Construction of the confluence Matrix through n languages

In Figure 15, one begins by choosing n languages that well represent language diversity (Altaic, Amerindian, Australian, Caucasian, Afro-Asiatic, Dravidian, Indo-European, Niger-Congo, Sino-Tibetan languages…). Then, for each of the n languages, one builds a/several paradigmatic graph(s). We have already begun the graph extraction process for several languages. We started with French for practical reasons: we had several directly operational digitized sources at our disposal: two standard dictionaries (the digitized Trésor de la Langue Française, the electronic Grand Robert), 7 digitized synonym dictionaries (Bailly, Benac, Du Chazaud, Guizot, Lafaye, Larousse and Robert) and several large electronic corpuses such as for example 10 years of the daily newspaper *Le Monde*. Using the database WordNet as well as the LDOCE dictionary we are currently building graphs for English, and are beginning to build a graph for Portuguese. We are planning on building graphs for Mandarin in the near future.
It is easier to build graphs for languages already having dictionaries and/or databases accessible on the World Wide Web, such as WordNet. There are however linguistic databases for other 'less digitized' languages, such as those used, internally for the present, by the researchers working on the project on semantic groupings within the CNRS Fédération Typologie et Universaux Linguistiques (http://www.typologie.cnrs.fr).

One may wonder, however, whether all of the existing data: dictionaries, databases… are relevant for our approach. Indeed, it is possible that some links be wrong in a dictionary or database, or that other links be missing for reflecting the exact reality of a language. Of course, it depends on the quality of the data in question, but experience has shown that the data established by linguists and/or lexicographers generally turns out to be relevant: the graphs extracted from the digitized Trésor de la Langue Française, the Grand Robert or the compilation of the seven synonym dictionaries mentioned above all agree in the confluences they show with Prox. Indeed, Prox is a robust method, which means that even if one changes several edges at random in a graph, it does not fundamentally change the results obtained. If an edge confluence exists in a graph's zone, the suppression or redirection of a few edges chosen at random in the graph does not strongly modify the confluence. This is an effect of the relativity of the confluences between themselves which is important for Prox, and therefore, unless one chooses the edges of the same confluence, the suppression or random redirection of edges will not profoundly affect the relativity of these confluences. It is in this matter that Prox is robust.

To illustrate the robustness of Prox, using DicoSynVerbe we built a graph DicoSynVerbe_10R by randomly redirecting 10% of the non reflexive edges. Thus we began by randomly removing, in an equiprobable manner, 10% of the non reflexive edges, then by randomly adding, in an equiprobable manner, the same amount of edges in order to obtain the DicoSynVerbe_10R graph.

Figure 16 below illustrates the list of 100 edges which entertain the strongest confluence relationships with the verb COMPRENDRE (understand) (from the highest ranked: strong confluence with COMPRENDRE (understand) – to the lowest ranked: the weakest confluence with COMPRENDRE (understand) –) calculated by Prox at t=3 on DicoSynVerbe_10R.

1 →**COMPRENDRE** (understand), 2 →DECOUVRIR (discover), 3 →DEVINER (guess), 4 →ENFERMER (close in), 5 →CONNAITRE (know), 6 →TROUVER (find), 7 →PRENDRE (take), **8 VOIR** (see), 9 →PENETRER (penetrate), 10 →SAISIR (grasp), 11 →SE COMPOSER (be made up of), 12 →APERCEVOIR (glimpse), **13 →ENTENDRE** (hear), 14 →RENFERMER (enclose), 15 →PERCER (pierce), 16 →REMARQUER (notice), 17 →CONSISTER (consist of), 18 →DECHIFFRER (decipher), 19 →COMPORTER (include), 20 →REPERER (spot), 21 →APPRENDRE (learn), 22 →ENGLOBER (enclose), 23 →COMPTER (count), 24 →ENCLORE (enclose), 25 →S'EXPLIQUER (become clear), 26 →AVOIR L'INTENTION (intend), 27 →MELANGER (mix), 28 →SE RENDRE COMPTE (realize), 29 →ENVELOPPER (envelop), 30 →MELER (tangle), 31 →ENTREVOIR (glimpse), 32 →SUIVRE (follow), **33 SENTIR** (feel), 34 →INTERPRETER (interpret), 35 →ADMETTRE (admit), 36 →SAVOIR (know), 37 →S'APERCEVOIR (realize), 38 →DEMELER (untangle), 39 →APPREHENDER (grasp), 40 →ENTRER (enter), 41 →VENIR A QUAI (dock), 42 →DECODER (decipher), 43 →INCORPORER (incorporate), 44 →PIGER (get), 45 →ASSIMILER (assimilate), 46 →INCLURE (include), 47 →REALISER (realize), 48 →SE CAVALER (run off), 49 →GROUPER (group together), 50 EMBRASSER (embrace), 51 →IMPLIQUER (imply), 52 →BAGARRER (fight), 53 →MORDRE (bite), 54 PERCEVOIR (perceive), 55 →SE METTRE A (begin), 56 DISCERNER (discern), 57 LIRE (read), 58 →TRADUIRE (translate), 59 DISTINGUER (distinguish), 60 JUGER (judge), 61 SURPRENDRE (surprise), 62 CONTENIR (contain), 63 REVELER (reveal), 64 REUNIR (reunite), 65 PRESSENTIR (foresee), 66 CONSTATER (note), 67 DECELER (detect), 68 VOULOIR (want), 69 S'AGENOUILLER (kneel), 70 NAITRE (be born), 71 ETRE FORME DE (be made up of), 72 ETRE CONSTITUE DE (be made up of), 73 DEBROUILLER (unravel), 74 NOTER (note), 75 FAIRE (do), 76 ASSOCIER (associate), 77 PASSER (pass), 78 TENIR (hold), 79 JOINDRE (join), 80 ENSERRER (ring), 81 IMAGINER (imagine), 82 DECRYPTER (decipher), 83 CONCEVOIR (conceive), 84 INTRODUIRE (introduce), 85 UNIR (unite), 86 FLAIRER (smell out), 87 TOUCHER (touch), 88 REGARDER (look at), 89 ENTOURER (surround), 90 APPARAITRE (appear), 91 EMPRISONNER (imprison), 92 COMBINER (combine), 93 CEINDRE (encircle), 94 AVOIR DANS L'IDEE (intend), 95 CACHER (hide), 96 CERNER (surround), 97 MARQUER (mark), 98 PENSER (think), 99 EPROUVER (feel), 100 S'AVISER (realize), …

Fig. 16. Proxemy of COMPRENDRE (understand) from DicoSynVerbe_10R at t=3

In DicoSynVerbe_10R the vertex COMPRENDRE (understand) has 52 neighbors, the neighbors of COMPRENDRE (understand) are preceded by an arrow → and the number that precedes each verb is its rank according to its proxemy to COMPRENDRE (understand) in DicoSynVerbe_10R.

One notes that in Figure 16 the verbs VOIR (see), SENTIR (feel) and ENTENDRE (hear) are ranked respectively 8th, 33rd and 13th which, as for Figure 13 remains very high considering the 9043 verbs present in DicoSynVerbe_10R (these three verbs are in the Top_4_per_1000 of the verb COMPRENDRE (understand)). This indicates that there subsist strong confluences in DicoSynVerbe_10R from COMPRENDRE (understand) towards VOIR (see), SENTIR (feel) and ENTENDRE (hear). The 10% of redirected edges having been chosen at random in the entire graph, this is the reason why if there is an over-dense edge zone in DicoSynVerbe, then this over-dense zone subsists in DicoSynVerbe_10R. To make a zone over-dense in edges disappear[22], one must not only choose the edges at random from the entire graph, but also choose them from the designated zone. In the same way, even if the most experienced lexicographers sometimes omit certain relations that one would linguistically be entitled to expect, or even to postulate other, less justifiable, relations, this sort of 'noise' thus created is nonetheless never concentrated in a particular zone but is spread out over all the data. And that is why the existing data: dictionaries, databases… are relevant for our approach with Prox, which is robust in the way described above.

[22] For example, if one were to randomly remove 10% of the trees planted on earth, then the forests (which is to say the zones relatively over-dense in trees) would still be forests (namely zones relatively over-dense in trees). To make a forest disappear one would have to not only randomly choose trees from the entire earth, but also choose them from the designated forest.

Once all the graphs are built from the existing data (data constructed by linguists and/or lexicographers, which, as we saw above, are generally relevant for our approach), we systematically inventory all the confluences which exceed a certain limit with Prox. This work is only partially automatic in that the results of the algorithms must of course be validated and adjusted by several native speakers for each of the languages studied. After validation, one obtains C which is the set of the k confluences detected among the set of our n languages. One may then build M the k×n matrix as illustrated in Figure 15 where the line i indexes the i^{th} confluence whereas the column j indexes the j^{th} language with:

$\forall i,\ 1 \leq i \leq k,\ \forall j,\ 1 \leq j \leq n\ ,\ [M]_{i,j} = 1$ if the i^{th} confluence is present in the graph of the j^{th} language and $[M]_{i,j} = 0$ otherwise.

The number i semantic association is then universal if and only if $\forall j,\ 1 \leq j \leq n,\ [M]_{i,j} = 1$.

Moreover, the n column vectors $([M]_{\bullet j})_{1 \leq j \leq n}$ identify each of the n languages studied according to their confluences. The set of these n vectors can then permit a classification of languages according to their semantic confluences and these classes can be compared to the classical typological models, notably to semantic maps (Haspelmath et al. 2001, Haspelmath 2003).

6. Conclusion

To organize a cartography of all natural languages according to their semantic associations by hand, would be a gigantic task. Having a robust method capable of capturing and measuring the confluences present in a paradigmatic network makes it possible to open the barriers which are (i) constructing the data and (ii) the systematic and quantitative inventory of the semantic associations present in the data, because:

1) As we saw in section 4, with Prox, one disposes of an automated tool for systematic searches and measurements of semantic associations (barrier ii);

2) As we saw in section 5, one can use existing data even if it shows certain weaknesses as compared to linguistic reality (barrier i);

However, this perspective is subordinate to our hypothesis:

> **(H1)** The paradigmatic graphs of all natural languages are hierarchical small worlds.

Indeed, on a random graph (which is not a hierarchical small world) Prox is less robust: for example to randomly redirect 10% of the non reflexive edges in a random graph can quite seriously modify the results. This is due to the fact that in a random graph, even if there are zones which are slightly denser than average, these zones are very fragile, and it is enough to remove a few edges in a zone for the results to be significantly different on the zone’s vertices. This means that in the case of a graph which is not a hierarchical small world, the omission or approximation of a few relations can imperil the exactness of the confluences measured. It would therefore be necessary that the data be without the slightest divergence from linguistic reality, and also that it be exhaustive, which is practically impossible, even for a sub-part of a language’s lexicon.

The first task is therefore to validate hypothesis (H1), or, if it is invalidated, one is faced with two classes of language:

(a) Languages whose paradigmatic graphs are hierarchical small worlds;
(b) Languages whose paradigmatic graphs are not hierarchical small worlds.

But, as we saw in section 2.3, several linguistic and psycholinguistic studies show the usefulness and efficacy of such structures for natural languages. That the structure in question be a hierarchical small world may be a sine qua non condition of the lexicon of a natural language, for its efficiency, transmission, evolution and because of human cognitive constraints.

This hypothesis has important consequences for linguistics and psycholinguistics, as well as for the theory of evolution. The proposition (A): *most large field graphs resemble each other in their hierarchical small world structures* and the hypothesis (H1): *the paradigmatic graphs of all natural languages are hierarchical small worlds* have as a consequence the proposition (B) *the paradigmatic structure of the lexicons of all natural languages resembles the structure of most of the world's objects*.

7 References

Adamic L. A., 1999. The small world Web. http://www.hpl.hp.com/shl/papers/smallworld/smallworld.pdf

Ancel L. W., Newman M. E. J., Martin M., Schrag S., 2001. *Applying Network Theory to Epidemics*, Control Measures for Outbreaks of "Mycoplasma pneumoniae", *SFI Working Paper*, n° 01-12-083, http://www.santafe.edu/sfi/publications/Working-Papers/01-12-083.ps.gz

Abello J., Pardalos P.M., Resende M.G.C., 1999. *On maximum cliques problems in very large graphs.* External memory algorithms, J. Abello and J. Vitter, Eds., DIMACS Series on Discrete Mathematics and Theoretical Computer Science, vol. 50, pp. 119-130, American Mathematical Society, http://www.research.att.com/~mgcr/doc/vlclq.ps.Z

Barabási A.-L., Albert R., Jeong H., and Bianconi G., 2000. *Power-Law Distribution of the World Wide Web*, Science 287 2115a (in Technical Comments) http://www.nd.edu/~networks/Papers/comments.pdf

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

Chen P., Pimenta M.-A., Duvignau K., Tonietto L., Gaume, B. 2006. Semantic approximations in the early verbal lexicon acquisition of Chinese: flexibility against error. Proceedings of the 7th Chinese Lexical Semantics Workshop (CLSW-7), May 22-23 2006, Taiwan (Forthcoming).

Duvignau K., Fossard M., Gaume B, Pimenta M.-A. 2005. From early lexical acquisition to the "disacquisition" of verbal lexicon: Verbal metaphor as semantic approximation. Proceedings of the II Conference on metaphor in language and thought, Universidade Federal Fluminense, 17-20 August 2005, Niteroi, Rio de Janeiro, Brazil (http://www2.lael.pucsp.br/~tony/metaphor/2005/)

[DG 04a] Duvignau K., Gaume B. 2004a. Linguistic, Psycholinguistic and Computational Approaches to the Lexicon: For Early Verb-Learning. A special issue on 'learning'. ESSCS Journal, Journal of the European Society for the study of cognitive systems. March, 6-2 (3) : 255-269.

Duvignau K., Gaume B. 2004b. First Words and Small Worlds: Flexibility and proximity in normal development Interdisciplinary conference "Architectures and Mechanisms for Language Processing", Aix en Provence, 16-18 septembre 2004, Aix en Provence.

Duvignau K., Gaume B., Kern S. 2005b. Semantic approximations intra-concept vs inter-concepts in early verbal lexicon: flexibility against error**.** Proceedings of ELA 2005, Emergence of language abilities: ontogeny and phylogeny, Lyon, December 8-10, 2005 (http://www.ddl.ish-lyon.cnrs.fr/ELA2005/)

Duvignau K., Gaume B., Nespoulous, J.-.L 2004. Proximité sémantique et stratégies palliatives chez le jeune enfant et l'aphasique, In Revue Parole, numéro spécial, J.-L. Nespoulous & J. Virbel (Coord.) : « Handicap langagier et recherches cognitives : apports mutuels », UMH, Belgique, Vol 31-32 : 219-255.

Duvignau K. 2002. 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.

Duvignau, K. 2003. Métaphore verbale et approximation. In Duvignau, K., Gaume, O. (eds) Regards croisés sur l'analogie. *Revue d'Intelligence Artificielle*, n° spécial, Vol 5/6. Hermès Lavoisier, Paris : 869-881.

Erdös & Renyi 1960. Erdös P. and Renyi A., *Publ. Math. Inst. Hung. Acad. Sci* 5, 17-61.

Evans, N. & Wilkins, D. 2000. In the mind's ear: The semantic extensions of perception verbs in Australian languages. *Language* 76/3, p. 546-592.

Fellbaum C., 1999. *La représentation des verbes dans le réseau sémantique WordNet.* In Langages, Sémantique lexicale et grammaticale, 136.

Ferrer R., Solé R. V. 2001. The small world of human language. Proceedings of The Royal Society of London. Series B, Biological Sciences, 268(1482):2261—2265, 2001 http://www.santafe.edu/sfi/publications/Working-Papers/01-03-016.pdf

Gaume B., 2004. Balades Aléatoires dans les Petits Mondes Lexicaux, In *I3 Information Interaction Intelligence* vol.4 - n°2, CEPADUES édition.

Gaume, B. 2006. Cartographier la forme du sens dans les petits mondes Lexicaux, *Proceedings of the 8th* Journées internationales d'analyse statistique des données textuelles, April 19, 20, 21 Besançon, JADT 2006. http://msh.univ-fcomte.fr/jadt2006/index.php?page=accueil

Gaume, B., Duvignau, K. 2004. Pour une ergonomie cognitive des dictionnaires électroniques, *Document Numérique*, numéro spécial : « Fouille de Textes et Organisation de Documents ». Lavoisier, Paris, (3) : 157-181.

Gaume B., Hathout N., Muller P. 2004. Word sense disambiguation using a dictionary for sens similarity measure *in acte COLING 2004,* The 20th International Conference on Computational Linguistics, COLING 2004 , Geneva.

Gaume B., Mathieu F., 2006. PageRank Induced Topology for Real-World Networks *(forthcoming).*

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

Goddard C. & Wierzbicka A. (eds.). 1994. *Semantic and Lexical Universals - Theory and Empirical Findings*. Amsterdam: John Benjamins.

Huberman B. A., and Adamic L.A., 1999. Growth dynamics of the world-wide web, *Nature* 401:131. http://xxx.lanl.gov/abs/cond-mat/9901071

Haspelmath, M*., et al.* (éds.). 2001. *Language typology and language universals. Vol. 1 et 2*. Berlin, New York: Mouton de Gruyter.

Haspelmath, M. 2003. The geometry of grammatical meaning: semantic maps and cross-linguistic comparison. In M. Tomasello (ed.) *The new psychology of language: cognitive and functional approaches to language structure*, vol. 2. Mahwah, NJ: Lawrence Erlbaum, 211-42.

Ide N., Véronis J., 1998. Introduction to the Special Issue on Word Sense Disambiguation : The State of the Art. Computational Linguistics 24(1), 1: .40.

Jeong H, Mason S.P., Barabasi A.-L. and Oltvai Z.N., 2001. *Lethality and centrality in protein networks*, *Nature* 411 41, http://www.nd.edu/~networks/cell/papers/protein.pdf

Karov Y., Edelman S., 1998. *Similarity-based Word Sense Disambiguation*. Computational Linguistics. 24(1), 41-59.

Koch P. 2001. Lexical typology from a cognitive and linguistic point of view. *Language Typology and Language Universals*. M. Haspelmath, E. König, W. Oesterreicher and W. Raible. Berlin, New York, Mouton de Gruyter. 2/2: 1142-1178.

Kochen M. (ed.), 1989. *The small world*, Ablex, Norwood, NJ.

Kleinberg J.M., Ravi K., Prabhakar R., Sridhar R., Andrews T. 1999. *The web as a graph: measurements, models, and methods*. In Proceedings of the Fifth International Conference on Computing and Combinatorics, Tokyo July 26-28, 1999 (COCOON'99). Berlin: Springer Verlag, 1-17.

Lakoff G. 1987 (1990). *Women, Fire, and Dangerous Things.* Chicago and London: University of Chicago Press.

Lakoff G. & Johnson M.. 1980 (2003). *Metaphors We Live By. With a new Afterword.* Chicago and London: University of Chicago Press.

Lebart L. Salem A. 1994. *Statistique textuelle*, Paris Dunod.

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

Newman M.E.J 2003a. *The structure and fonction of complex networks*, http://www.santafe.edu/~mark/recentpubs.html

Newman M.E.J. 2003b. Ego-centerer networks and the riple effect, *Social Networks* 25, 83-95, 2003

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

Ravasz E., Barabási A.L. 2003. *Hierarchical Organization in Complex Networks*. Phys. Rev. E 67, 026112, 2003 http://arxiv.org/abs/cond-mat?0206130

Redner 1998. *How Popular is Your Paper? An Empirical Study of the Citation Distribution*, Redner S., cond-mat/9804163, *European Physical Journal B*, 4, 131-134 (1998). http://cbd.bu.edu/members/sredner.html

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

Sweetser E. 1991. *From Etymology to Pragmatics. Metaphorical and Cultural Aspects of Semantic Structure*. Cambridge: Cambridge University Press.

Tonietto L., Pimenta M.-A., Duvignau K., Gaume B., Bosa C.A., 2006. Aquisição inicial do léxico verbal e aproximações semânticas em português. In, Psicologia: Reflexão e Crítica, Brazilia. (Forthcoming)

Victorri B., Fuchs C. 1996. La polysémie – Construction dynamique du sens, Paris, Hermès.

Viberg A., 1984. The verbs of perception: A typological study. in Butterworth, B., Comrie, B. and Dahl, Ö. (eds.), *Explanations for language universals*. Berlin: Mouton de Gruyter, p. 123-162.

Watts D.J. 1999. *Small Worlds: The Dynamics of Networks between Order and Randomness*, Princeton University Press.

Wierzbicka A. 1992. *Semantics, Culture, and Cognition. Universal Human Concepts in Culture-Specific Configurations*. New York - Oxford: Oxford University Press.

Wilkins D.P. 1996. Natural tendencies of semantic change and the search for cognates. in Durie, Mark and Ross, Malcolm (eds) *The comparative method reviewed*, 224-304, New York: Oxford University Press.

Watts D.J., Strogatz S.H., 1998. *Collective dynamics of 'small-world' networks*. **Nature** 393: 440-442, 1998 http://tam.cornell.edu/SS_nature_smallworld.pdf