Skillex: a graph-based lexical score for measuring the semantic efficiency of used verbs by human subjects describing actions

Bruno Gaume[*] — Karine Duvignau[*] — Emmanuel Navarro[**]
Yann Desalle[*] — Hintat Cheung[***] — Shu-Kai Hsieh[****]
Pierre Magistry[*****] — Laurent Prévot[******]

[*] *Cognition, Langues, Langage, Ergonomie, Université de Toulouse*
[**] *Kodex.Lab*
[***] *The Hong Kong Institute of Education, China*
[****] *National Taiwan University, Taiwan*
[*****] *Analyse linguistique profonde à grande échelle, France*
[******] *Parole et langage, Université Aix-Marseille*
gaume@univ-tlse2.fr, duvignau@univ-tlse2.fr, navarro@kodexlab.com, yann.desalle@gmail.com, hintat@ied.edu.hk, shukaihsieh@ntu.edu.tw, pmagistry@gmail.com, laurent.prevot@lpl-aix.fr

ABSTRACT. *Dictionaries are sociocultural objects that can be used as underlying structures of cognitive science models. We first show that the lexical networks constructed from dictionaries, despite a surface disagreement at links level, share a common topological structure. We assume that this deep structure reflects the semantic organisation of the lexicon shared by the members of a linguistic community. We propose a model based on the exploration of this specific structure to analyse and compare the semantic efficiency of [Children/Adults] productions in action labelling tasks. We define a generic score of semantic efficiency,* SKILLEX. *Assigned to participants of the* APPROX *protocol, this score enables us to accurately classify them into Children and Adults categories.*

RÉSUMÉ. *Les dictionnaires sont des objets socioculturels qui peuvent être utilisés comme structures sous-jacentes pour la modélisation en sciences cognitives. Nous montrons d'abord que les réseaux lexicaux construits à partir de dictionnaires, malgré un désaccord de surface au niveau des liens, partagent une structure topologique commune. En supposant que cette structure profonde reflète l'organisation sémantique du lexique partagée par les membres d'une communauté linguistique, nous proposons un modèle basé sur l'exploration de cette structure spécifique pour analyser et comparer l'efficacité sémantique des productions [Enfants/Adultes] dans une tâche d'étiquetage d'action. Nous définissons un score générique de l'efficacité sémantique,* SKILLEX. *Assigné aux participants du protocole* APPROX, *ce score nous permet de les classer avec précision dans les catégories enfants et adultes.*

KEYWORDS: *dictionary, networks, lexicon, Approx, Skillex.*

MOTS-CLÉS : *dictionnaire, réseaux, lexique, Approx, Skillex.*

TAL. Volume 55 – n° 3/2014, pages 97 à 121

Dictionaries are sociocultural objects. We take advantage of their structural features for defining *Skillex*, a graph-based lexical score for measuring the semantic efficiency of used verbs by human subjects describing specific actions. Assigned to participants of the *Approx* protocol, this measure enables us to accurately classify them into Children and Adults categories.

We first show in section 1 that the lexical networks constructed from resources of various origins (two resources built by lexicographers: the *Robert dictionary* (Robert and Rey, 1985) and the *Larousse dictionary* (Guilbert *et al.*, 1971-1978); one resource built by crowd sourcing: the *Jeux De Mots* (Lafourcade, 2007)), despite a surface disagreement at links level, share a common topological structure. Assuming that this structure reflects the semantic organisation of the lexicon shared by the members of a linguistic community, we propose then in section 2 a model based on the exploration of this specific structure to analyse and compare the semantic efficiency of language productions of Children versus Adults. Section 3 contains concluding remarks and presents our future works.

1. Structure of synonymy networks

Lexical resources can be modeled as graphs $G = (V, E)$ where a set of n vertices V encodes lexical entities (lemmas or word senses, syntactic frames...) and a set of m edges $E \subseteq \mathbf{P}_2^V$ encodes a lexical relation between these entities. A burning issue regarding these lexical networks is their apparent significant disagreement: for exemple, in two standard synonymy graphs on the same language, $G_1 = (V, E_1)$ and $G_2 = (V, E_2)$ a great proportion of pairs $\{x, y\} \in \mathbf{P}_2^V$ are linked in G_1 ($\{x, y\} \in E_1$) but not in G_2 ($\{x, y\} \notin E_2$): x and y are synonyms for G_1 but not for G_2. Such a large amount of disagreement is not compatible with the assumption that synonymy reflects a somewhat common understanding of the semantic structure of the lexicon of a given language community. To resolve this apparent contradiction, one must look at lexical networks from a broader perspective. Fig. 1 illustrates a toy example of generalised edge disagreement between two graphs, despite a similar structure. The structural similarity is only visible by looking at each of the graphs as a whole, as opposed to comparing their edges one by one. The two graphs G_1 and G_2 do not have any edge in common but still look very similar, because they both draw two dense zones that encompass the same vertices: {1, 2, 3, 4, 11, 12, 13} and {4, 5, 6, 7, 8, 9, 10}.

The dense zones found in the toy example of Fig. 1 are actually a widespread feature of *terrain networks* [1], as most of them are Hierarchical Small World (HSW) networks that share four similar properties (Watts and Strogatz, 1998; Albert and Barabasi, 2002; Newman, 2003; Gaume *et al.*, 2010). They exhibit:

$\mathbf{p_1}$: A low density (not many edges);

1. Terrain networks are graphs that model real data gathered by field work, for example in sociology, linguistics or biology. They contrast with artificial graphs (deterministic or random).

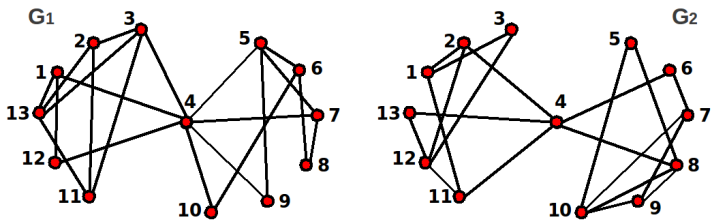


Figure 1. *Contradiction between local variability and overall similarity.*

$\mathbf{p_2}$: Short paths (the average number of edges L on the shortest path between two vertices is low);

$\mathbf{p_3}$: A high clustering rate $C = \frac{3 \times number\ of\ triangles}{number\ of\ connected\ triplets}$ (locally densely connected subgraphs can be found whereas the whole graph is globally sparse in edges);

$\mathbf{p_4}$: The distribution of their degrees can be approximated by a power law.

We show in section 1.1 that the lexical networks studied in this paper also have HSW properties. So, as Fig. 1 suggests, an apparent disagreement between lexical networks at links level does not necessarily imply structural incompatibility of their data. By choosing an appropriate level of representation, it should be possible to reconcile the information they convey into a global agreement on the synonymy they model. In section 1.2, we describe lexical networks with an original method based on random walks. Instead of characterising pairs of vertices according only to whether they are connected or not, we measure their structural proximity by evaluating the relative probability of reaching one vertex from the other via a short random walk. This proximity between vertex is the basis on which we can measure the structural quality of the surface divergence between two lexical networks because it outlines the similar dense zones of the graphs.

1.1. *Compare* $\mathbf{G_1} = (\mathbf{V}, \mathbf{E_1})$ *and* $\mathbf{G_2} = (\mathbf{V}, \mathbf{E_2})$ *by comparing the sets* $\mathbf{E_1}$ *and* $\mathbf{E_2}$ *as «links bags» without structures*

We illustrate our purpose in this section with the comparison of two lexical resources, both built by lexicographers at quite the same time:

– $\mathbf{Rob} = (\mathbf{V_{Rob}}, \mathbf{E_{Rob}})$: The *Robert* (Robert and Rey, 1985) *dictionary* was digitalised during an IBM / ATILF Research unit partnership [2]. The electronic resource lists the synonyms of the various senses of lemmas. The vertices of the lexical graph Rob that was built from it are the lemmas, not tagged by their various senses. The pair $\{x, y\} \in E_{Rob}$ if and only if one of the senses of x is considered synonymous

2. `http://www.atilf.fr/spip.php?rubrique18`.

with one sense of y by the lexicographic team of *Robert*. For example the polysemous verb *causer* is both synonymous with *parler (speak)* and *engendrer (cause)*.

– $\mathbf{Lar} = (\mathbf{V_{Lar}}, \mathbf{E_{Lar}})$: The lexical graph Lar was built from the *Larousse dictionary* (Guilbert *et al.*, 1971-1978) like the Rob was.

We synthesise the characteristic of a graph regarding its HSW properties by a set of figures called the *pedigree* of a graph. Table 1 provides the pedigrees of Rob and Lar and shows that they are all typical HSW (Motter *et al.*, 2002; De Jesus Holanda *et al.*, 2004; Gaume, 2004).

Lexical Graphs	**n**	**m**	$\langle \mathbf{k} \rangle$	**C**	$\mathbf{L_{lcc}}$	λ ($\mathbf{r^2}$)
Lar	22,066	73,091	6.62	0.19	6.36	-2.43 (0.90)
Rob	38,147	99,998	5.24	0.12	6.37	-2.43 (0.94)

Table 1. *Pedigrees of lexical graphs: n and m are the number of vertices and edges, $\langle k \rangle$ is the mean degree of the vertices, C is the clustering coefficient of the graph, L_{lcc} is the average shortest path between any two nodes of the largest connected component (subgraph in which there exist at least one path between any two nodes), λ is the coefficient of the best fitting power law of the degree distribution and r^2 is the correlation coefficient of the fit, measuring how well the data is modelled by the power law.*

Given two graphs $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$, we measure the similarity of lexical coverage of G_1 and G_2 by the *Jaccard* index: $J(G_1, G_2) = \frac{|V_1 \cap V_2|}{|V_1 \cup V_2|}$. We then have $J(Rob, Lar) = 0.49$. These two graphs therefore have a wide enough common lexical coverage for the comparison between synonymy judgments they model to be carried out on this common lexical coverage. So, to measure the agreement between edges of G_1 and G_2, one first reduces the two graphs to their common vertices: $G'_1 = \big(V' = (V_1 \cap V_2), E'_1 = E_1 \cap (V' \times V')\big)$ and $G'_2 = \big(V' = (V_1 \cap V_2), E'_2 = E_2 \cap (V' \times V')\big)$.

For each vertex pair $\{a, b\} \in (V' \times V')$, four cases are possible:

– $\{a, b\} \in E'_1 \bigcap E'_2$: agreement on pair $\{a, b\}$, $\{a, b\}$ is synonymous for G'_1 and for G'_2;

– $\{a, b\} \in \overline{E'_1} \bigcap \overline{E'_2}$: agreement on pair $\{a, b\}$, $\{a, b\}$ is neither synonymous for G'_1 nor for G'_2;

– $\{a, b\} \in E'_1 \bigcap \overline{E'_2}$: disagreement on pair $\{a, b\}$, $\{a, b\}$ is synonymous for G'_1 but not for G'_2;

– $\{a, b\} \in \overline{E'_1} \bigcap E'_2$: disagreement on pair $\{a, b\}$, $\{a, b\}$ is synonymous for G'_2 but not for G'_1.

A long tradition of graph comparison research consists in assessing if two graphs are isomorphic. Two graphs $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ are isomorphic if there exists a bijective function f: $V_1 \mapsto V_2$ such that, for any two vertices $\{u, v\}$ of

G_1: $\{u, v\} \in E_1 \Leftrightarrow \big\{f(u), f(v)\big\} \in E_2$. So the research in this tradition consists in looking for such an isomorphism. In the case that this paper proposes to study, nodes are labeled and can only be put in correspondence across graphs if they have the same label: the isomorphism is given, it is the identity function. Assessing whether two graphs $G_1 = (V, E_1)$ and $G_2 = (V, E_2)$ are isomorphic would then only involve verifying that $E_1 = E_2$.

Such a similarity is very basic: if no single edge differs, then the two graphs are similar, else they are different. In order to soften the isomorphism approach, and to provide a continuous, quantitative measure of how different two graphs are, several approaches were proposed (see a review in, for example, (Gao *et al.*, 2010)). These methods are inspired by the edit distance of character strings proposed by (Levenshtein, 1966). The graph edit distance is defined, between two graphs $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$, as the cheapest series of operations to make G_1 isomorphic to G_2. Operations are usually the simplest set of insertion, deletion and substitution of nodes and edges. This set is subject to be extended according to the data graphs attempt at modeling. For example, in the case of image segmentation, (Ambauen *et al.*, 2003) introduce operations such as node splitting and merging.

In the scope of this paper, since $V_1 = V_2 = V$, the only possible operations will be the deletion or insertion of edges. If the cost of editing an edge is 1, then the edit distance between G_1 and G_2 is: $ED = |E_1 \cap \overline{E_2}| + |E_2 \cap \overline{E_1}|$. Note that $ED \in [0, |E_1| + |E_2|]$. This dissimilarity measure does not take into account the number of edges of G_1 and G_2. Having to edit 10 edges to make two graphs of 15 edges isomorphic is not the same as having to edit 10 edges to make two graphs of 15,000 edges isomorphic. This edit distance has to be normalised:

$$GED(G_1, G_2) = \frac{|E_1 \cap \overline{E_2}| + |E_2 \cap \overline{E_1}|}{|E_1| + |E_2|} \quad [1]$$

Now, $GED(G_1, G_2) \in [0, 1]$. This measurement on Lar'/Rob' is: $GED(Lar', Rob') = 0.47$. This shows that Lar' and Rob' are dissimilar : *Larousse* and *Robert* dictionaries have only a weak agreement on which pairs of lemmas are synonymous. This can be explained by the fact that the projection of the gradual notion of near synonymy onto binary synonymy judgements leaves ample room for interpretation, even if the judges are expert lexicographers as for the *Larousse* and *Robert* standard dictionaries. In fact, independently built resources that describe the same linguistic reality often show a weak agreement even when based on human judgements under the same protocol (Murray and Green, 2004).

1.2. *Compare* $\mathbf{G_1} = (\mathbf{V}, \mathbf{E_1})$ *and* $\mathbf{G_2} = (\mathbf{V}, \mathbf{E_2})$ *by comparing the structure generated by* $\mathbf{E_1}$ *on* V *to the structure generated by* $\mathbf{E_2}$ *on* V

GED is a quantitative measure which compares graph edge-by-edge. It does not take their global structure into account, although it is very specific because they have HSW properties. The presence or absence of an edge between a pair is a judgement

on its synonymous nature that can be confirmed or contradicted by the topological structure of the graph around it. In this section we describe an alternative quantitative approach based on random walks that enables us to enrich edge information on pairs with this confirmation measure.

If $G = (V, E)$ is a reflexive [3] and undirected graph, let us define $d_G(u) = |\{v \in V / \{u, v\} \in E\}|$ the degree of vertex u in graph G, and let us imagine a walker wandering on the graph G: at a time $t \in \mathbb{N}$, the walker is on one vertex $u \in V$; at time $t + 1$, the walker can reach any neighbouring vertex of u, with uniform probability. This process is called a simple random walk (Bollobas, 2002). It can be defined by a Markov chain on V with a $n \times n$ transition matrix $[G]$:

$$[G] = (g_{u,v})_{u,v \in V} \text{ with } g_{u,v} = \begin{cases} \dfrac{1}{d_G(u)} \text{ if } \{u, v\} \in E, \\ 0 \text{ else.} \end{cases}$$

Since G is reflexive, each vertex has at least one neighbour (itself) thus $[G]$ is well defined. Furthermore, by construction, $[G]$ is a stochastic matrix: $\forall u \in V, \sum_{v \in V} g_{u,v} = 1$. The probability $P^t_G(u \rightsquigarrow v)$ of a walker starting on vertex u to reach a vertex v after t steps is:

$$P^t_G(u \rightsquigarrow v) = ([G]^t)_{u,v} \quad [2]$$

One can then prove (Gaume, 2004), with the Perron-Frobenius theorem (Stewart, 1994), that if G is connected (i.e. there is always at least one path between any two vertices), reflexive and undirected, then $\forall u, v \in V$:

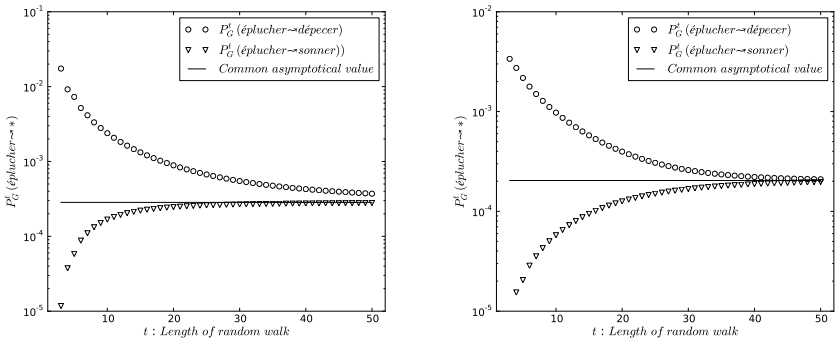
$$\lim_{t \to \infty} P^t_G(u \rightsquigarrow v) = \lim_{t \to \infty} ([G]^t)_{u,v} = \frac{d_G(v)}{\sum_{x \in V} d_G(x)} = \pi_G(v) \quad [3]$$

It means that when t tends to infinity, the probability of being on a vertex v at time t does not depend on the starting vertex but only on the degree of v. In the following we will refer to this limit as $\pi_G(v)$. The dynamics of the convergence of random walks towards the limit (Eq. [3]) is heavily dependent on the starting node. Indeed, the trajectory of the random walker is completely governed by the topology of the graph in the vicinity of the starting node: after t steps, any vertex v located at a distance of t links or less can be reached. The probability of this event depends on the number of paths between u and v, and on the structure of the graph around the intermediary vertices along those paths. The more interconnections between the vertices, the higher the probability of reaching v from u. For example, if we take $G_1 = Rob$ and $G_2 = Lar$, and choose the three vertices $u =$ *éplucher (peel)*, $r =$ *dépecer (tear apart)* and $s =$ *sonner (ring)*, where:

3. I.e. each vertex is connected to itself. If such self-loops do not exist in the data, they may be generally added without loss of information.

- u and r are judged synonymous in *Rob*: $\{u, r\} \in E_1$;
- u and r are judged not synonymous in *Lar*: $\{u, r\} \notin E_2$;
- r and s have the same number of synonyms in G_1: $d_{G_1}(r) = d_{G_1}(s) = d_1$;
- r and s have the same number of synonyms in G_2: $d_{G_2}(r) = d_{G_2}(s) = d_2$.

Then Equation [3] states that $(P^t_{G_1}(u \rightsquigarrow r))_{1 \leq t}$ and $(P^t_{G_1}(u \rightsquigarrow s))_{1 \leq t}$ converge to the same limit: $\pi_{G_1}(r) = \pi_{G_1}(s) = \frac{d_1}{\sum_{x \in V_1} d_{G_1}(x)}$ as do $(P^t_{G_2}(u \rightsquigarrow r))_{1 \leq t}$ and $(P^t_{G_2}(u \rightsquigarrow s))_{1 \leq t}$: $\pi_{G_2}(r) = \pi_{G_2}(s) = \frac{d_2}{\sum_{x \in V_2} d_{G_2}(x)}$. However the two series do not converge with the same dynamics. At the beginning of the walk, when t is small, one can expect that $P^t_{G_1}(u \rightsquigarrow r) > P^t_{G_1}(u \rightsquigarrow s)$ and $P^t_{G_2}(u \rightsquigarrow r) > P^t_{G_2}(u \rightsquigarrow s)$ because *éplucher (peel)* is semantically closer to *dépecer (tear apart)* than to *sonner (ring)*. Indeed the number of short paths between *éplucher (peel)* and *dépecer (tear apart)* is much greater than those between *éplucher (peel)* and *sonner (ring)*.



(a) $G_1 = Rob$ (b) $G_2 = Lar$

Figure 2. *Different convergence dynamics of $P^t_G(u \rightsquigarrow v)$ to its limit for three cases of u, v relation: (1) u and v are synonyms like* éplucher (peel) *and* dépecer (tear apart) *in* Rob*; (2) u and v are not synonyms and are semantically distant like* éplucher (peel) *and* sonner (ring) *in* Rob *and in* Lar*; (3) u and v are not synonyms but semantically close like* dépecer (tear apart) *and* éplucher (peel) *in* Lar.

Figure 2(a) shows the values of $P^t_{G_1}(u \rightsquigarrow r)$ and $P^t_{G_1}(u \rightsquigarrow s)$ versus t, and compares them to their common limit. Figure 2(b) shows the values of $P^t_{G_2}(u \rightsquigarrow r)$ and $P^t_{G_2}(u \rightsquigarrow s)$ versus t, and compares them to their common limit. These figures confirm our intuition that, since *éplucher (peel)* and *dépecer (tear apart)* are semantically close, $P^t_{G_1}(u \rightsquigarrow r)$ and $P^t_{G_2}(u \rightsquigarrow r)$ decrease to their limit, even if, like in G_2, r and s are not synonymous.

The limit $\pi_G(v)$ does not actually provide any information about the proximity of u and v in the graph, but on the opposite, it masks it with the importance of v in the

graph. Therefore we define the t-confluence $CONF^t_G(u, v)$ between two vertices u, v on a graph G as follows:

$$CONF^t_G(u, v) = \frac{P^t_G(u \rightsquigarrow v)}{P^t_G(u \rightsquigarrow v) + \pi_G(v)} \quad [4]$$

$CONF^t_G$ actually defines an infinity of symmetrical vertex closeness measures, one for each random walk length t. For clarity reasons, for the rest of the paper we choose one closeness measure within this infinity: we set t with the following arguments:

– **if t is too large :** $\forall u_1, v_1, u_2, v_2 \in V,\ CONF^t_G(u_1, v_1) \approx CONF^t_G(u_2, v_2) \approx 0.5$. This would hinder the distinction between pairs of vertices in a same higher density zone from pairs in lower density zones;

– **if t is too small :** For any pair $\{u, v\}$ for which the shortest path length in G' is greater than t, $P^t_{G'}(u \rightsquigarrow v) = 0$, thus $CONF^t_G(u, v) = 0$. This does not indicate if the pair $\{u, v\}$ is in an higher or a lower density zone of G.

So, in the rest of this paper, t is set to $t = 5$ and we define $CONF_G = CONF^5_G$. We consider as "close" each pair of vertices $\{u, v\}$ having a confluence $CONF_G(u, v)$ greater than 0.5. In other words, u and v are close if the probability of reaching v from u after a 5 step random walk is greater than the probability to be on v after an infinite walk.

We have seen that the lexical networks are HSW, so they exhibit the properties p_2 (short paths) and p_3 (high clustering spleen). With a classic distance as *the shortest path between two vertices*, all vertices would be close to each other in a lexical network (because of the p_2 property). On the contrary, $CONF_G$ allows us to identify the vertices of a same cluster of G (because of the p_3 property):

– if u and v are in a same higher density zone of G, $P^5_G(u \rightsquigarrow v) > \pi_G(v)$ and thus $CONF_G(u, v) > 0.5$;

– if u and v are not particularly close or distant in G, $P^5_G(u \rightsquigarrow v) = \pi_G(v)$ and thus $CONF_G(u, v) \approx 0.5$;

– if u and v are not in a same higher density zone in G , $P^5_G(u \rightsquigarrow v) < \pi_G(v)$ and thus $CONF_G(u, v) < 0.5$.

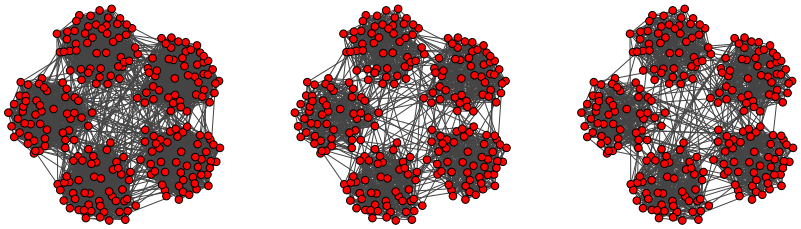
1.3. *A controlled experimental setup with artificial graphs*

We artificially build two types of pairs of graphs to compare :

– **two graphs with 5 higher density zones :** We first build a graph $G_a = (V, E_a)$ where V is the union of $k = 5$ groups of $n = 50$ vertices, and edges are drawn randomly between vertices of the graph with two different probabilities. They are drawn with a probability $p_1 = 0.5$ between two vertices of the same group, and $p_2 =$

0.01 between vertices of two different groups. We then build a new graph $G_b = (V, E_b)$ by randomly choosing half of the edges of G_a, and a new graph $G_c = (V, E_c)$ such that $E_c = E_a \setminus E_b$. These 3 graphs are plotted in Fig. 3. Although G_b and G_c do not have any edge in common, ($E_b \cap E_c = \varnothing$), G_b and G_c exhibit 5 identical, local, higher density zones;

– **two random graphs :** We first build a random graph $G_a^R = (V, E_a^R)$ such $|E_a^R| = |E_a|$. We then build a new graph $G_b^R = (V, E_b^R)$ by randomly choosing half of the edges of G_a^R, and a new graph $G_c^R = (V, E_c^R)$ such that $E_c^R = E_a^R \setminus E_b^R$. Neither G_b^R nor G_c^R have higher density zones.



(a) $G_a = (V, E_a)$ (b) $G_b = (V, E_b)$ (c) $G_c = (V, E_c)$

Figure 3. *Artificial graphs with* 5 *identical, local, higher density zones.*

– Since $E_b \cap E_c = \varnothing$, $E_b \cap \overline{E_c} = E_b$ and $E_c \cap \overline{E_b} = E_c$. Therefore, $GED(G_b, G_c) = \frac{|E_b \cap \overline{E_c}| + |E_c \cap \overline{E_b}|}{|E_b| + |E_c|} = \frac{|E_b| + |E_c|}{|E_b| + |E_c|} = 1$. This would mean that these two graphs are completely dissimilar, which is true in the sense that they have no edges in common, however it is clearly wrong with respect to the topological "organisation" they share. Indeed two vertices that are in the same relatively higher-density zone in the first graph will also be in the same relatively higher density zone in the other graph.

– Since $E_b^R \cap E_c^R = \varnothing$, $E_b^R \cap \overline{E_c^R} = E_b^R$ and $E_c^R \cap \overline{E_b^R} = E_c^R$. Therefore, $GED(G_b^R, G_c^R) = \frac{|E_b^R \cap \overline{E_c^R}| + |E_c^R \cap \overline{E_b^R}|}{|E_b^R| + |E_c^R|} = \frac{|E_b^R| + |E_c^R|}{|E_b^R| + |E_c^R|} = 1$.

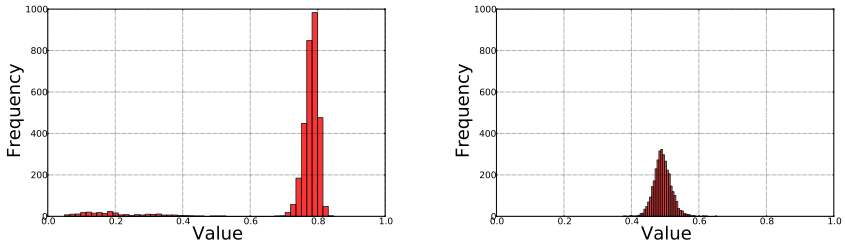
All quantitative measurements like GED, based only on the counting of the number of disagreements, have the drawback of only comparing graphs as "bag of edges", thus being insensitive to the topological contexts. But if we compare the distributions of the confluence of conflicting edges in G_b vs G_c, in figure 4(a) on the one hand, and in G_b^R vs G_c^R, in figure 4(b) on the other hand , the difference is striking.

Therefore we define $\mu(G_1, G_2)$ a measure of conflicting edges in G_1 vs G_2:

$$\mu(G_1, G_2) = \frac{\left(\sum_{\{u,v\} \in (E_2 \cap \overline{E_1})} CONF_{G_1}(\{u, v\}) \; + \sum_{\{u,v\} \in (E_1 \cap \overline{E_2})} CONF_{G_2}(\{u, v\}) \right)}{|E_1 \cap \overline{E_2}| + |E_2 \cap \overline{E_1}|}$$

Although $GED(G_b, G_c) = GED(G_b^R, G_c^R) = 1$, with μ, we can now see the difference : On 50 realisations $\mu(G_b, G_c) = 0.74$ (with standard deviation $std < 0.005$)

while $\mu(G_b^R, G_c^R) = 0.49$ ($std < 0.005$). Quantitatively the difference between G_b/G_c and G_b^R/G_c^R is the same but it structuraly differs.



(a) G_b/G_c (b) G_b^R/G_c^R

Figure 4. *Histogram of the set* $\{CONF_{G_c}(\{u,v\})$ *such that* $\{u,v\} \in (E_b \cap \overline{E_c})\} \bigcup \{CONF_{G_b}(\{u,v\})$ *such that* $\{u,v\} \in (E_c \cap \overline{E_b})\}$, *compared to the histogram of the set* $\{CONF_{G_c^R}(\{u,v\})$ *such that* $\{u,v\} \in (E_b^R \cap \overline{E_c^R})\} \bigcup \{CONF_{G_b^R}(\{u,v\})$ *such that* $\{u,v\} \in (E_c^R \cap \overline{E_b^R})\}$.

1.4. *Applications on lexical graphs*

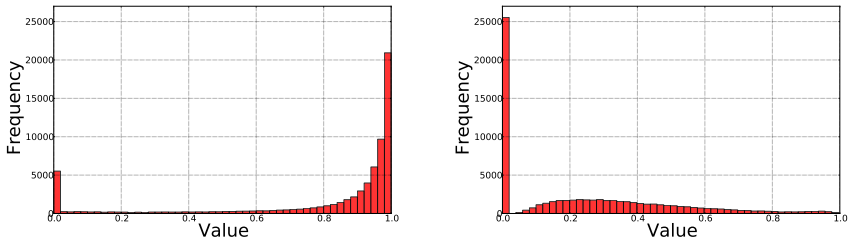
We begin by looking at the distribution of the confluence of conflicting edges in $Lar' = (V', E_{Lar'})$ vs $Rob' = (V', E_{Rob'})$. We compare it to this same distribution on pairs of equivalent *random* graphs $Lar'^R = (V', E^R_{Lar'})$ and $Rob'^R = (V', E^R_{Rob'})$ built such that: $|E^R_{Lar'} \cap E^R_{Rob'}| = |E_{Lar'} \cap E_{Rob'}|$, $|E^R_{Lar'} \cap \overline{E^R_{Rob'}}| = |E_{Lar'} \cap \overline{E_{Rob'}}|$, $|\overline{E^R_{Lar'}} \cap E^R_{Rob'}| = |\overline{E_{Lar'}} \cap E_{Rob'}|$.

By construction we have $GED(Lar', Rob') = GED(Lar'^R, Rob'^R)$, but if we compare the distributions of the confluence of conflicting edges in Lar' vs Rob', on the one hand in Figure 5(a), and in Lar'^R vs Rob'^R, on the other hand in Figure 5(b), the difference is striking.

Quantitatively, the difference between Lar'/Rob' and Lar'^R/Rob'^R is the same: $GED(Lar', Rob') = GED(Lar'^R, Rob'^R) = 0.47$, but it differs structurally $\mu(Lar', Rob') = 0.80$ while $\mu(Lar'^R, Rob'^R) = 0.24$ (with 50 realisations: $std < 0.005$). There's the same number of disagreements, but those disagreements are structurally weak between Lar' and Rob', while they are structurally stronger between Lar'^R and Rob'^R. This is what allows us to see the Figure 5 and that's what measures μ.

We now compare a set of lexical networks of various origins, resources built by lexicographers and by crowd sourcing:

– $\mathbf{Rob} = (\mathbf{V_{Rob}}, \mathbf{E_{Rob}})$ and $\mathbf{Lar} = (\mathbf{V_{Lar}}, \mathbf{E_{Lar}})$: see section 1.1;



(a) Lar'/Rob' (b) Lar'^R/Rob'^R

Figure 5. *Histogram of the set* $\{CONF_{Rob'}(\{u,v\})$ *such that* $\{u,v\} \in (E_{Lar'} \cap \overline{E_{Rob'}})\} \bigcup \{CONF_{Lar'}(\{u,v\})$ *such that* $\{u,v\} \in (E_{Rob'} \cap \overline{E_{Lar'}})\}$*, compared to the histogram of the set* $\{CONF_{Rob'^R}(\{u,v\})$ *such that* $\{u,v\} \in (E_{Lar'^R} \cap \overline{E_{Rob'^R}})\} \bigcup \{CONF_{Lar'^R}(\{u,v\})$ *such that* $\{u,v\} \in (E_{Rob'^R} \cap \overline{E_{Lar'^R}})\}$

– $\mathbf{Jdm} = (\mathbf{V_{Jdm}}, \mathbf{E_{Jdm}})$: The *Jeux De Mots* resource [4] is built from a form of crowd sourcing, using a game described in (Lafourcade, 2007). Players must find as many words as possible that are associated to a term presented to the screen, according to a rule provided by the game. The aim is to find as many semantic associations as possible amongst what other players have found, but that the concurrent player has not found. Several rules can be proposed, including the request for listing as many synonyms or quasi-synonyms as possible. The collected results in January 2014 build a graph of words linked by typed semantic relations (according to the rules) that is freely accessible. We work here on the sub-graph of synonymy relations [5].

Each of these resources is split by parts of speech (Nouns, Verbs, Adjectives) resulting in three different graphs, designated, for example for the *Robert dictionary*, as follows: (ex: $Rob \Rightarrow Rob_N = (V_{Rob_N}, E_{Rob_N}), Rob_V = (V_{Rob_V}, E_{Rob_V}), Rob_A = (V_{Rob_A}, E_{Rob_A})$). Table 2 provides the pedigrees of these graphs and shows that they are all typical HSW. In table 3 we compare 3 pairs of graphs by parts of speech.

Between the graphs Lar, Rob, and Jdm, the surface measurement GED is always between 0.45 and 0.51, indicating a low agreement at links level compared irrespective of their structural contexts. However the structural measure μ is always greater than or equal to 0.70, which means that, despite the substantial proportion of local disagreements, these graphs have a similar deep structure.

4. `http://www.lirmm.fr/jeuxdemots/jdm-accueil.php`.
5. This synonymy graph can be viewed here : `http://autourdumot.fr/fr.V.peler`.

Lexical Graphs		n	m	⟨k⟩	C	L_{lcc}	λ (r^2)
Lar	**Adjectives**	5,510	21,147	7.68	0.21	4.92	-2.06 (0.88)
	Nouns	12,159	31,601	5.20	0.20	6.10	-2.39 (0.88)
	Verbs	5,377	22,042	8.20	0.17	4.61	-1.94 (0.88)
Rob	**Adjectives**	7,693	20,011	5.20	0.14	5.26	-2.05 (0.94)
	Nouns	24,570	55,418	4.51	0.11	6.08	-2.34 (0.94)
	Verbs	7,357	26,567	7.22	0.12	4.59	-2.01 (0.93)
Jdm	**Adjectives**	9,859	30,087	6.10	0.16	5.44	-2.24 (0.90)
	Nouns	29,213	56,381	3.86	0.14	6.48	-2.66 (0.93)
	Verbs	7,658	22,260	5.81	0.14	5.06	-2.08 (0.89)

Table 2. *Pedigrees of lexical graphs (we refer to the legend of table 1 for the description of the columns).*

	Rob/Lar	**Jdm/Lar**	**Jdm/Rob**
	GED (μ) (μ^R)	**GED** (μ) (μ^R)	**GED** (μ) (μ^R)
A	**0.45** (**0.76**) (0.34)	**0.47** (**0.71**) (0.38)	**0.51** (**0.70**) (0.29)
N	**0.48** (**0.70**) (0.20)	**0.48** (**0.70**) (0.19)	**0.47** (**0.70**) (0.13)
V	**0.48** (**0.73**) (0.40)	**0.46** (**0.70**) (0.39)	**0.47** (**0.70**) (0.37)

Table 3. *To compare two lexical graphs G_1/G_2, one first reduces the two graphs to their common vertices: $G'_1 = \big(V' = (V_1 \cap V_2), E'_1 = E_1 \cap (V' \times V')\big)$ and $G'_2 = \big(V' = (V_1 \cap V_2), E'_2 = E_2 \cap (V' \times V')\big)$. Then, we build their equivalent random graphs G'^R_1 and G'^R_2 and compute: $GED = GED(G'_1, G'2)$, $(\mu) = \mu(G'_1, G'2)$ and $(\mu^R) = \mu(G'^R_1, G'^R_2)$. Each value (μ^R) on each of equivalent random graphs, is the average with 50 realisations $\mu(G'^R_1, G'^R_2)$ (all standard deviations $std < 0.005$).*

2. Skillex

We saw in the previous section 1 that synonyms dictionaries are sociocultural objects which share a common deep structure, i.e., the same local densities. We make the assumption that this deep structure reflects the shared semantic organisation of the lexicon by members of a same linguistic community. In this section, we will exploit this assumption to define an analytical tool for diagnosing properties of a mental lexical network based on language production. Instead of using psycholinguistics criteria that must be analysed by a human hand (specificity, conventionality, imageability...) to evaluate how efficient a verb is to label an action, we automatically compute the semantic efficiency of verbs by mapping the semantics of labelled actions onto current thesauruses and by exploring the specific structure of these thesauruses as complex

networks. Thus, we propose a model to compute a generic score of semantic efficiency, called *Skillex*, which combines two other efficiency measurements. Applied to participants of *Approx*, a protocol for collecting action labellings, *Skillex* accurately categorizes these participants into the two [Young Children/Adults] age groups.

State-of-the-art research about lexical networks and lexical acquisition process are briefly reviewed in section 2.1. Then in section 2.2 we detail our model and its evaluation in section 2.3, on the basis of the *Approx* protocol.

2.1. *Lexical networks and lexical acquisition process*

Firstly, recent research (Bowerman, 2005; Duvignau *et al.*, 2012; Gaume *et al.*, 2008) has discovered two salient patterns in the verb productions of young children:

(a) verbs that, although semantically close to the expected conventional verb, don't match the labelled action on at least one of their semantic components;

(b) verbs that expect generic categories on their semantic components: many objects fit in such categories.

On the other hand, several research projects have demonstrated a relation between the structure of lexical networks and the lexical acquisition process. According to (Steyvers and Tenenbaum, 2005), in lexical networks built from the Roget's thesaurus, WordNet and the USF word association norms (Nelson *et al.*, 2004), vertex degrees are correlated with:

– the age of acquisition (AoA) of English words (Morrison *et al.*, 1997);

– the frequency of occurrence of such words in English, itself correlated with their AoA (Kučera *et al.*, 1967).

These findings are confirmed by a study of (De Deyne and Storms, 2008a), for the Dutch language, on the basis of the graph extracted from the Dutch Word Association norms (De Deyne and Storms, 2008b). The study also shows that both the clustering coefficient in the word's neighbourhood (distance 2) and its betweenness centrality (measure of the centrality of a vertex in a graph) are correlated to its AoA.

2.2. *Model*

2.2.1. *Theoritical motivations of the model*

Our model is motivated by the parallel between (a) experimental results on semantic acquisition of action verbs and (b) our hypotheses on HSW properties of synonymy networks (Duvignau *et al.*, 2004; Gaume *et al.*, 2008):

– **1.a** verbs produced by adults are more specific than those produced by children;

– **1.b** specific verbs' degrees are low ($\mathbf{p_4}$);

– **2.a** action verbs produced by children are less appropriate to the labelled actions than those produced by adults;

– **2.b** in synonymy networks, verbs are brought closer if their meanings are closely related ($\mathbf{p_3}$).

Rank	$\mathbf{Rob_V}$	$\mathbf{Lar_V}$	$\mathbf{Jdm_V}$
1	peler* (peel)	décortiquer* (peel/shell)	peler* (peel)
2	s'époiler* (shave)	éplucher* (peel/pare)	décortiquer* (peel/shell)
3	desquamer* (skin)	peler* (peel)	exfolier* (exfoliate)
4	écorcer* (put the bark off)	écaler* (shell)	épiler* (shave)
5	dépouiller* (skin/strip)	écorcer* (put the bark off)	éplucher* (peel/pare)
6	éplucher* (peel/pare)	écosser (shell/pod)	écorcer* (put the bark off)
7	écorcher* (skin)	dépouiller (skin/strip)	tondre* (mow)
8	enlever* (remove)	plumer* (pluck)	écorcher* (skin)
9	décortiquer (peel/shell)	monder (blanch)	dépouiller (skin/strip)
10	démascler (≈ pull the bark off)	scruter (scrutinise)	dépiler (remove hair from)
11	gemmer (≈ pull the gum off)	raisonner (reason)	cailler* (curdle)
12	baguer (ring)	tamiser (sieve/sift)	raser (shave)
13	inciser (lance)	émonder (prune/blanch)	écaler (shell)
14	désosser (bone)	épinceter (pull the buds off)	analyser (analyse)
15	dépiauter (skin)	nettoyer (clean)	voler (steal)
16	couper (cut)	disséquer (dissect)	désosser (bone)
17	voler (steal)	retourner (turn over, around ...)	écosser (shell/pod)
18	examiner (examine)	époutir (pull impurity off a cloth)	inciser (lance)
19	plumer (pluck)	épointiller (pull the dirt off a sheet)	érafler (scrape/graze)
20	épouiller (delouse)	analyser (analyse)	égratigner (scratch/graze)

Table 4. *The 20 closest verbs to peler (peel) in Rob_V, in Lar_V and in Jdm_V (with $t = 5$) (* for neighbours).*

This model is based on two measures: (1) the degree of a verb in a synonymy network and (2) a verb's proximity to a lexico-semantic zone of a synonymy network that is detailed in the next section.

2.2.2. *Prox*

Let us define a lexico-semantic zone of the graph $G = (V, E)$ by a probability distribution Δ on V, its vertex set (more details on such a definition are given hereafter in section 2.3.2). We then define the proximity of a verb $v \in V$ to a lexico-semantic zone defined by a probability distribution Δ by:

$$prox_G(v, \Delta) = \frac{(\Delta[G]^5)_v}{\max_{r \in V}(\Delta[G]^5)_r} \quad [5]$$

For example, table 4 provides the list of the 20 closest French verbs to *peler* (*to peel*) in Rob_V, in Lar_V and in Jdm_V ($\Delta = \delta_{peler}$, the certainty to be located on *peler*).

2.2.3. *Efficiency of a verb*

Let $G = (V, E)$ be a verb synonymy graph, $v \in V$ a verb and Δ_a the probability distribution on V that delimits the meaning of an action a. We define $s(v, \Delta_a)$ the *efficiency* of verb v in relation to Δ_a by:

$$s(v, \Delta_a) = \frac{prox_G(v, \Delta_a)}{d_G(v)} \quad [6]$$

Our model is based on the hypothesis that adults produce verbs that have a better efficiency in relation to Δ_a than the efficiency of verbs produced by children to label the same action. The measures $prox_G(v, \Delta_a)$ and $d_G(v)$ both play a meaningful part in the efficiency in relation to the Δ_a score:

– $prox_G(v, \Delta_a)$: the greater the proximity of verb v to Δ_a, the more semantically appropriate the verb v is, to describe a;

– $d_G(v)$: the smaller the degree of verb v, the more specific the verb v.

2.2.4. *Four scores*

This section details how our model attributes four scores of lexical performance to each individual, given a language L, a graph $G_L = (V_L, E_L)$, and a set of actions $A = \{a_1, \cdots, a_i, \cdots a_n\}$. Let $\Delta^L = \{\Delta^L_{a_1} \cdots, \Delta^L_{a_i} \cdots, \Delta^L_{a_n}\}$ be the lexico-semantic zones that correspond, in G_L, to the actions of A. Let x be an individual who produced a set of verbs $W_{a_i,x}$ to label action a_i. For each verb set $W_{a_i,x}$ such that $W_{a_i,x} \cap V_L \neq \varnothing$, the following figures are computed:

– $D(W_{a_i,x})$ is the mean [6] of the set $\{d_G(v) \mid v \in W_{a_i,x} \cap V_L\}$

– $P(W_{a_i,x})$ is the mean of the set $\{prox_G(v, \Delta^L_{a_i}) \mid v \in W_{a_i,x} \cap V_L\}$

– $S(W_{a_i,x})$ is the mean of the set $\{s(v, \Delta^L_{a_i}) \mid v \in W_{a_i,x} \cap V_L\}$.

These three figures are the basis on which we compute the four scores of each participant x for the action category defined by A:

– **Productiveness score** $N_A(x)$**:** mean of $\{|W_{a,x}| \mid a \in A\}$

– **Degree score** $D_A(x)$**:** mean of $\{D(W_{a,x}) \mid a \in A$ and $W_a \cap V_L \neq \varnothing\}$

– **Prox score** $P_A(x)$**:** mean of $\{P(W_{a,x}) \mid a \in A$ and $W_a \cap V_L \neq \varnothing\}$

– **Skillex score** $S_A(x)$**:** mean [7] of $\{S(W_{a,x}) \mid a \in A\}$.

2.3. *Evaluation*

2.3.1. *Approx protocol*

The *Approx* protocol (Méligne *et al.*, 2011; Duvignau *et al.*, 2012) permits to gather verbal production labeling an action. The *Approx* protocol is, on average, completed by a participant in 20 minutes, and enables us to compute a lexical performance score for each participant.

2.3.1.1. Material and participants

The material (illustrated in figure 6) consists in sixteen 30-second action-films without speech, that show acts of deterioration/separation of objects. In each film a woman alters an object with the help of her hands or with an instrument, explicitly showing an initial state and a final state.

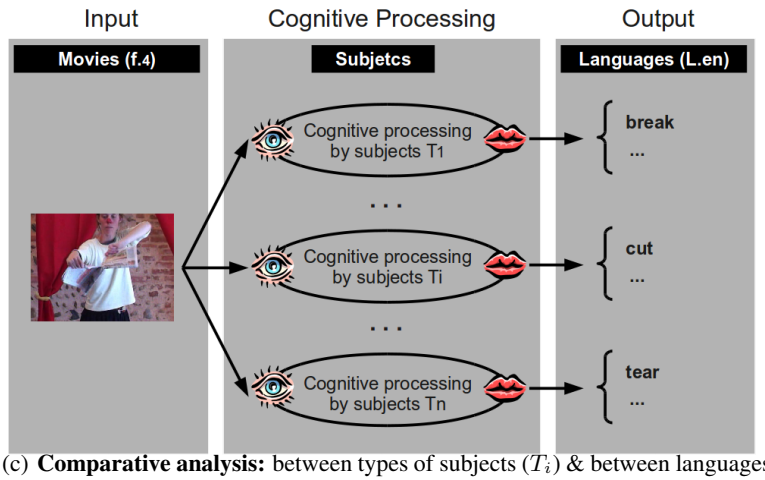
6. Mean is for arithmetic mean.

7. When $W_{a,x} \cap V_L = \varnothing$ we assign $S(W_{a,x}) = 0$.

(a) **Material :** 16 movies of DETERIORATION-SEPARATION

1- *"What did the woman do?"*
2- *"What the woman did, tell it differently, with other words"*

(b) **Tasks :** 1-Naming & 2-Rewording



(c) **Comparative analysis:** between types of subjects (T_i) & between languages (L_j).

Figure 6. *The* APPROX *protocol.*

The *Approx* protocol is without cultural bias (Cheung *et al.*, 2010), the films include sounds but no speech, the *Approx* protocol can therefore be used with different people of different languages, different ages, with or without pathology. We focus here on 2 groups of French native speakers [8]:

– C_F : 74 French young children (2-5 years old)

– A_F : 76 French young adults (18-40 years old)

2.3.1.2. Procedure

The films are shown in random order to a participant. After each film, the experimenter asks [9] the participant what the woman did. Between each action film, a distractor is shown to avoid perseveration effects. Results of participants who do not watch all 16 films are not taken into account. Lexical action labels are extracted from the elicited responses, and lemmatised. Compound labels are split according to their components :

– simple verb + complement (e.g. to break into pieces → to break + into pieces)

– simple verb + simple verb (e.g. to make broken → to make + to break)

2.3.2. *From action-stimuli to lexico-semantic zones*

For French language F with the synonymy graph $Rob_V = (V_{Rob_V}, E_{Rob_V})$ and the 2 groups of participants: C_F young children and A_F young adults, a lexico-semantic zone $\Delta_a^{Rob_V}$ is the distribution of probability on V_{Rob_V} that denotes, as objectively as possible, a action-stimulus a of the protocol. To define this distribution, a mixed [10] Pop_F sample of participants is gathered by randomly choosing 25 participants from C_F and 25 from A_F. For each verb v of V_{Rob_V} and each action a is attributed the frequency $freq_a^F(v)$ with which it was used by participants of Pop_F to label action a. The probability distribution $\Delta_a^{Rob_V}$, on V_{Rob_V}, then defines a's lexico-semantic zone in Rob_V:

$$\forall v \in V_{Rob_V}, (\Delta_a^{Rob_V})_v = \frac{freq_a^F(v)}{\sum_{s \in V_{Rob_V},} freq_a^F(s)} \quad [7]$$

For French language F with the 2 groups of participants C_F and A_F, and for each synonymy graph, $Lar_V = (V_{Lar_V}, E_{Lar_V})$ and $Jdm_V = (V_{Jdm_V}, E_{Jdm_V})$, in the same way we define respectively for each action a : $(\Delta_a^{Lar_V})$, $(\Delta_a^{Jdm_V})$.

2.3.3. *Tasks*

The two tasks are detailed using the French synonymy graph $Rob_V = (V_{Rob_V}, E_{Rob_V})$. The exact same procedures are done with the two other French synonymy graphs $Lar_V = (V_{Lar_V}, E_{Lar_V})$, $Jdm_V = (V_{Jdm_V}, E_{Jdm_V})$.

8. Participants do not have any cognitive impairment.

9. We use here only the first task : *"What did the woman do?"*

10. So that lexico-semantic zones do not induce a bias towards the adult or child age group.

2.3.3.1. Task 1: Computing participant's scores

We refer to the 16 action-stimuli of the protocol as $A = \{a_1, \cdots, a_i, \cdots, a_{16}\}$, and to their corresponding lexico-semantic zones on V_{Rob_V} as $\Delta^{Rob_V} = \{\Delta_{a_1}^{Rob_V} \cdots, \Delta_{a_i}^{Rob_V} \cdots, \Delta_{a_{16}}^{Rob_V}\}$. Three scores $D_A^{Rob_V}(x)$, $P_A^{Rob_V}(x)$ and $S_A^{Rob_V}(x)$ are computed for each native French speaker participant to the *Approx* protocol on the action category "deterioration/separation of objects" denoted by A.

In order to evaluate our model on the basis of this task, we compare young children's scores to scores of adult participants: a significant difference would mean that such scores accurately discriminate the two age groups [Children/Adults].

2.3.3.2. Task 2 : Automatic [Children/Adults] age group categorisation

It consists in measuring the accuracy of the automatic categorisation of the two age groups C_F and A_F, on the basis of the three scores computed in task 1. With each of the 3 scores, we use the k-means algorithm ($k = 2$) (Hartigan and Wong, 1979) to separate the set of participants into two categories. When considering the $Degree\ score$, the category with the greatest centroid is assigned to the *young children* category, the other to the *adults* category. Conversely, when considering the $Prox\ score$ or the $Skillex\ score$, the category with the greatest centroid is assigned to the *adults* category, the other to the *young children* category.

The accuracy of the automatic categorisation is measured by the agreement rate between the expected categories (C_F and A_F) and the score-computed categories.

2.3.4. *Results*

2.3.4.1. Task 1 results

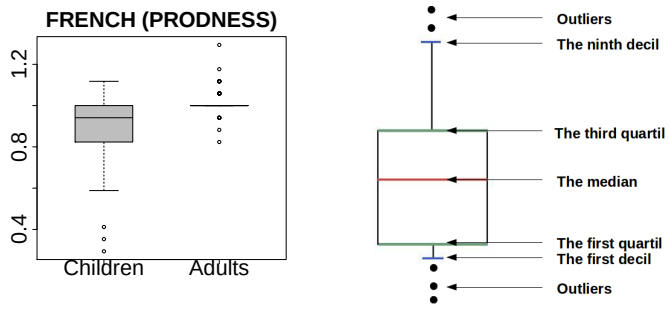
We used an ANOVA to measure how significant the difference between young children's and adults' PROX scores is, and a non-parametric Man-Whitney-Wilcoxon test to measure how significant the differences were between the $Productiveness$, $Degree$ and *Skillex* scores of young children and adults [11]. Results are shown in figures 7, and 8:

– **Productiveness score:** shows a significant difference between the *productiveness scores* of children and adults ($W(150) = 4788$, $p < 0.001$);

– **Degree score:** shows a significant difference between the *degree scores* of children and adults, (FR-ROB: $W(150) = 5252$, $p < 0.001$; FR-LAR: $W(150) = 4857$, $p < 0.001$; FR-JDM: $W(150) = 5176$, $p < 0.001$);

– **Prox score:** shows a significant difference between the *Prox scores* of children and adults, (FR-ROB: $W(150) = 22.46$, $p < 0.001$; FR-LAR: $W(150) = 63.19$, $p < 0.001$; FR-JDM: $W(150) = 26.91$, $p < 0.001$);

11. Since, according to the Shapiro-Wilk test, the distribution of the $Productiveness$, $Degree$ and *Skillex* scores are not normal distributions, ANOVA was not applicable.



(a) Box-and-whisker diagrams (b) Graphical representation

Figure 7. *$Productiveness$ scores of the [Children/Adults] age groups.*

– **Skillex score:** shows a significant difference between the *Skillex scores* of children and adults, (FR-ROB: $W(150) = 5531$, $p < 0.001$; FR-LAR: $W(150) = 5467$, $p < 0.001$; FR-JDM: $W(150) = 5551$, $p < 0.001$).

The $Productiveness$, $Degree$, $Prox$ and $Skillex\ scores$ highlight a significant difference between the verb productions of young children and of adults, upon a task that consists in labelling actions that show deteriorations or separations of objects, with the three graphs Rob_V, Lar_V and Jdm_V.

2.3.4.2. Task 2 results

Task 2 aims to confirm that task 1 results are significant and consistent enough to enable automatic categorisation of adults and children. These automatic categorisation is shown in figures 9(a), 9(b), 9(c). It is evaluated by the rate of agreement between automatically computed categories and expected categories, which is measured with the Precision and κ, the Kappa of Cohen (Cohen, 1960):

– precision is the observed agreement probability p_o;

– the κ is defined as : $\kappa = \frac{p_o - p_e}{1 - p_e}$ in which p_e is the expected agreement probability knowing (a) the distribution of individuals on the Adult and Child categories that were built by the $2-mean$ algorithm and (b) the distribution of individuals on the expected C_F and A_F groups.

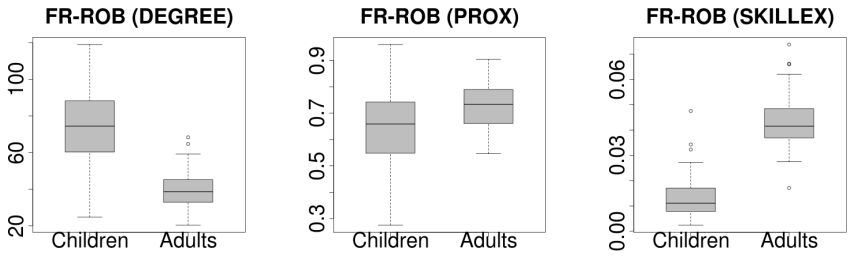
In table 5 we show that the $Skillex\ score$ categorises [Children/Adults] with a excellent agreement (according to the scale of (Landis and Koch, 1977) independently of the graph ($\kappa = .93$, $\kappa = .88$, $\kappa = .91$ respectively with Rob_V, Lar_V and Jdm_V)

2.4. *Skillex on Mandarin*

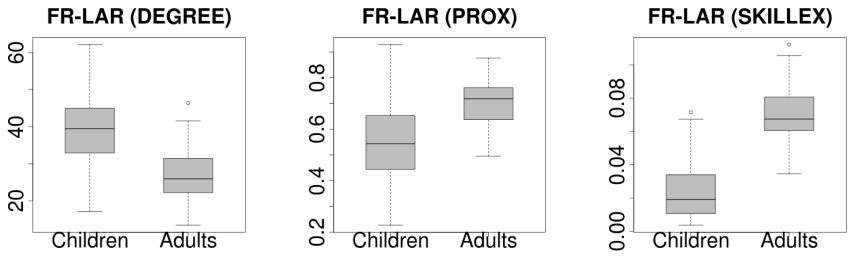
We apply the same procedure on the Mandarin with:

– C_M: 29 Mandarin young children (2-5 years old)

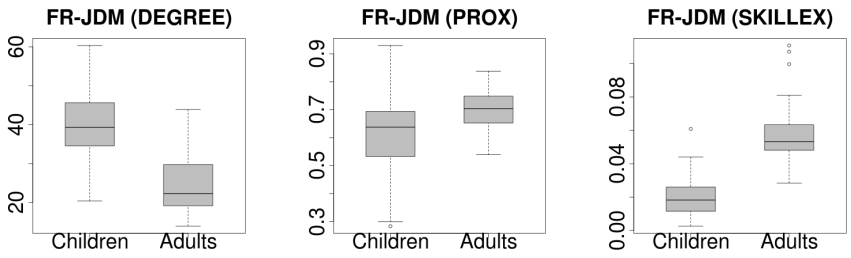
– A_M: 60 Mandarin young adults (18-30 years old)



(a) $Degree$ with Rob_V (b) $Prox$ with Rob_V (c) *Skillex* with Rob_V



(d) $Degree$ with Lar_V (e) $Prox$ with Lar_V (f) *Skillex* with Lar_V



(g) $Degree$ with Jdm_V (h) $Prox$ with Jdm_V (i) *Skillex* with Jdm_V

Figure 8. *Box-and-whisker diagrams: $Degree$, $Prox$, and $Skillex$ scores of the [Children/Adults] age groups with Rob_V, Lar_V and Jdm_V.*

– One Mandarin synonymy graph: $\mathbf{Ccw_V} = (\mathbf{V_{Ccw_V}}, \mathbf{E_{Ccw_V}})$: It is a graph of verbs extracted from CilinCWN: a fusion of Chinese WordNet (CWN) [12] and a

12. Chinese WordNet is a lexical resource modelled on Princeton WordNet, with many novel linguistic considerations for Chinese. It is proposed and launched by *Huang et al.* (Huang *et al.*, 2004), at the time of writing it contains 28,815 synonyms. It has been maintained and extended at National Taiwan University, and available at `http://lope.linguistics.ntu.edu.tw/cwn2`.

GRAPH	n=150	SCORE			
		$Degree$	$Prox$	*Skillex*	$Productiveness$
with Rob_V	Precision	.83	.74	**.97**	
	κ	.67	.48	**.93**	
with Lar_V	Precision	.78	.75	**.94**	Precision: .71
	κ	.56	.49	**.88**	κ: .42
with Jdm_V	Precision	.85	.75	**.95**	
	κ	.69	.50	**.91**	

Table 5. *2-means clustering results on French:* $Degree$*,* $Prox$*, Skillex and* $Productiveness$*.*

GRAPH	n=89	SCORE			
		$Degree$	$Prox$	*Skillex*	$Productiveness$
with Ccw_V	Precision	.64	.84	**.91**	.44
	κ	.17	.62	**.80**	.03

Table 6. *2-means clustering results on Mandarin :* $Degree$*,* $Prox$*, Skillex and* $Productiveness$*.*

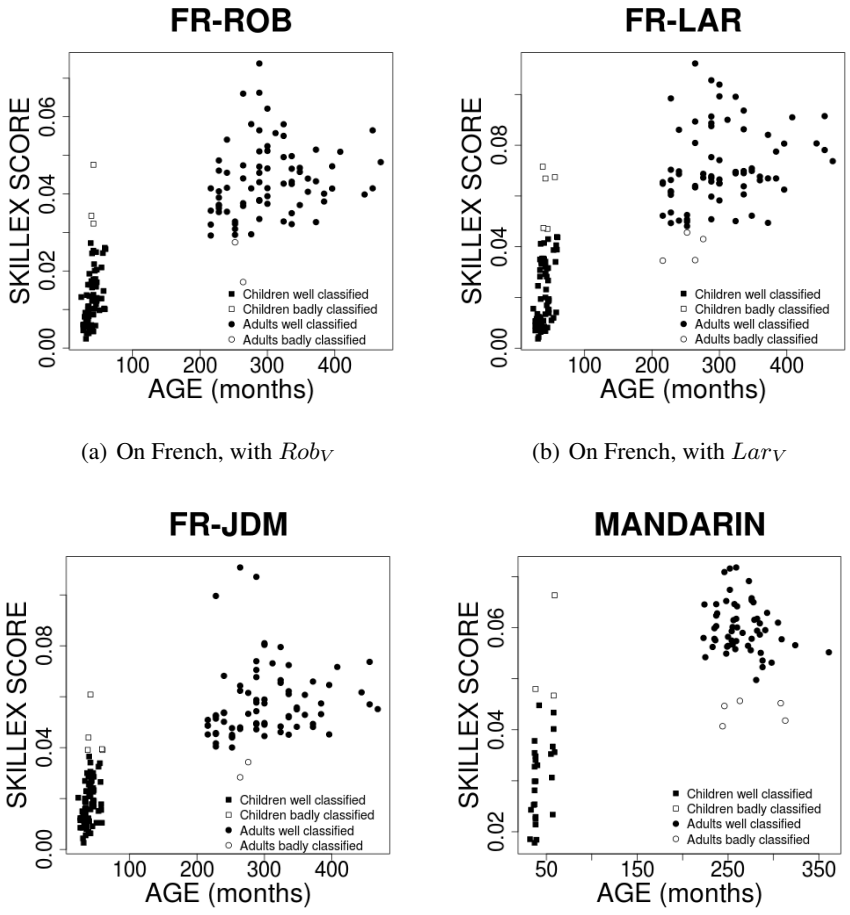
Chinese thesaurus TongYiCi CiLin (Cilin) [13]. Data was processed similarly to the way Rob_V, Lar_V and Jdm_V were built.

– Like for French, compound labels are split according to their components, moreover, when the produced verb by a participant is a Mandarin resultative compound verb (Li and Thompson, 1981), it is split according to: simple verb + result.

Table 6 and Figure 9(d) show the results on Mandarin. Table 5 and 6 suggest that the main component ($degree$ or $prox$) of the lack of efficiency in action labelling during lexical acquisition depends on the language to acquire: (a) whereas, in Mandarin, the $Prox\ score$ categorises [Children/Adults] with a substantial agreement (according to the scale of (Landis and Koch, 1977), $\kappa = .62$ with G_{Ccw}), this is less the case in French ($\kappa = .48$, $\kappa = .49$, $\kappa = .50$ respectively with Rob_V, Lar_V and Jdm_V); (b) whereas, in French, the $Degree\ score$ categorises [Children/Adults] with an substantial agreement ($\kappa = .67$, $\kappa = .56$, $\kappa = .69$ respectively with Rob_V, Lar_V and Jdm_V), this is not the case in Mandarin ($\kappa = .17$ with Ccw_V).

In fact, the $Skillex\ score$ is the only score able to highlight differences of semantic efficiency of action labelling between children and adults independently from the language. It is the only score that accurately categorises participants into the two [Children/Adults] age groups in both languages ($kappa \geqslant .80$: almost perfect agreement).

13. The Tongyici Cilin (Mei *et al.*, 1984) is a Chinese synonym dictionary known as a thesaurus in the tradition of Roget's Thesaurus in English. It contains about 70,000 lexical items.



(a) On French, with Rob_V

(b) On French, with Lar_V

(c) On French, with Jdm_V

(d) On Mandarin, with Ccw_V

Figure 9. *Automatic categorisation [Children/Adults] with 2-means clustering on Skillex scores with Rob_V, Lar_V, Jdm_V for French and G_{Ccw} for Mandarin.*

3. Conclusion and future works

In this paper, we first show that the lexical networks constructed from resources of various origins are Hierarchical Small World, and despite a surface disagreement at links level, share a common topological structure. We make the assumption that this deep structure reflects the shared semantic organisation of the lexicon by members of a same linguistic community. We then use of this deep structure as an artefact of the humans' representation of lexical knowledge for defining *Skillex*, a lexical score for measuring the semantic efficiency of used verbs by human subjects describing specific actions. Assigned to participants of the *Approx* protocol, this measure enables us to accurately classify them into Children and Adults categories for Mandarin and also for French. The *Approx* protocol is directly applicable to different languages

and participant samples (adults, L1 (first language), L2 (second language), children, participants with pathologies ...). Moreover, results do not significantly vary with the lexical resources on which *Skillex* score computations are based. The participant's *Skillex* score computation is therefore robust to resource variation.

We focused in this paper on the comparative study between children and adults on French and on Mandarin, but there is hope that *Skillex* can be successfully used in other contexts for investigating humans' representation of lexical knowledge. So, we intend to further this initial study into four directions: **(a)** validate the deep structure of synonym networks as an artefact of the humans' representation of lexical knowledge by using random walk-based measure to simulate elicited judgments of word similarity by humans (Hill *et al.*, 2014; Bruni *et al.*, 2014), **(b)** to extend the analysis to other languages, with the long term perspective of initiating a language typology of lexical acquisition dynamics (i.e. with multilingual ressources like Wiktionary [14]), **(c)** to extend the protocol to other action categories (for example verbs of movement) in order to compare semantic efficiency of humans' productions in action labelling tasks across action types, and **(d)** to extend the study to the analysis of pathologies:

• **Various stages of the Alzheimer's disease** (Joubert *et al.*, 2010). Building on works of Méligne *et al.* (2011), we formulate the two following hypotheses: On the basis of their *Approx* protocol verb production (first task only), participants can be attributed a $Skillex\ score$ that:

– **H1.1**: will accurately categorise participants into two [Moderate Alzheimer/Older without pathology] groups;

– **H1.2**: will NOT enable their accurate categorisation into two [Moderate Alzheimer/Child without pathology] groups;

• **Asperger's syndrome** (Atwood, 1998). Building on works of Maffre *et al.* (2012), we formulate the two following hypotheses: On the basis of their *Approx* protocol verb production (first task only), participants can be attributed a $Skillex\ score$ that:

– **H2.1**: will accurately categorise participants into two [Asperger Child/Child without pathology] groups;

– **H2.2**: will NOT enable their accurate categorisation into two [Asperger Child/Adult without pathology] groups.

4. References

Albert R., Barabasi A.-L., "Statistical Mechanics of Complex Networks", *Reviews of Modern Physics*, vol. 74, p. 74-47, 2002.

Ambauen R., Fischer S., Bunke H., "Graph Edit Distance with Node Splitting and Merging, and Its Application to Diatom Idenfication", *Graph Based Representations in Pattern Recognition, 4th IAPR International Workshop*, York, UK, p. 95-106, July, 2003.

Atwood T., *Asperger's Syndrome*, Jessica Kingsley Publishers, 1998.

Bollobas B., *Modern Graph Theory*, Springer-Verlag New York Inc., October, 2002.

14. `http://www.wiktionary.org/`.

Bowerman M., "Why Can't You 'Open' a Nut or 'Break' a Cooked Noodle? Learning Covert Object Categories in Action Word Meanings", *in* L. Gershkoff, D. Rakison (eds), *Building Object Categories in Action Word Meanings*, Lawrence Erlbaum Associates, Mahwah, NJ, p. 209-244, 2005.

Bruni E., Tran N. K., Baroni M., "Multimodal Distributional Semantics", *Journal of Artificial Intelligence Research (JAIR)*, vol. 49, p. 1-47, 2014.

Cheung H., Desalle Y., Duvignau K., Gaume B., Chang C.-H., Magistry P., "The Use of a Cultural Protocol for Quantifying Cultural Variations in Verbs Semantic between Chinese and French", *Proceedings of 24th Pacific Asia Conference on Language, Information and Computation: Workshop on Model and Measurement of Meaning (M3)*, Sendai, Japan, 2010.

Cohen J., "A Coefficient of Agreement for Nominal Scales", *Educ. Psychol. Meas.*, vol. 20, n° 1, p. 27-46, 1960.

De Deyne S., Storms G., "Word Associations: Network and Semantic Properties", *Behavior Research Methods*, vol. 40, n° 1, p. 213-231, 2008a.

De Deyne S., Storms G., "Word Associations: Norms for $1,424$ Dutch Words in a Continuous Task", *Behavior Research Methods*, vol. 40, n° 1, p. 198-205, 2008b.

De Jesus Holanda A., Pisa I. T., Kinouchi O., Martinez A. S., Ruiz E. E. S., "Thesaurus as a Complex Network", *Physica A: Statistical Mechanics and its Applications*, vol. 344, n° 3-4, p. 530-536, December, 2004.

Duvignau K., Gaume B., Nespoulous J.-L., "Proximité sémantique et stratégies palliatives chez le jeune enfant et l'aphasique", *in* J.-L. Nespoulous, J. Virbel (eds), *Revue Parole, numéro spécial Handicap langagier et recherches cognitives : apports mutuels*, vol. 31-32, Université de Mons-Hainaut, Belgique, p. 219-255, 2004.

Duvignau K., Tran T., Manchon M., "For a New Look at 'Lexical Errors': Evidences from Semantic Approximations with Verbs in Aphasia", *Journal of psycholinguistics Research*, vol. 42, n° 4, p. 339-347, 2012.

Gao X., Xiao B., Tao D., Li X., "A Survey of Graph Edit Distance", *Pattern Anal. Appl.*, vol. 13, n° 1, p. 113-129, 2010.

Gaume B., "Balades aléatoires dans les petits mondes lexicaux", *I3: Information Interaction Intelligence*, 2004.

Gaume B., Duvignau K., Prévot L., Desalle Y., "Toward a Cognitive Organization for Electronic Dictionaries, the Case for Semantic Proxemy", *Coling 2008: Proceedings of the Workshop on Cognitive Aspects of the Lexicon (COGALEX 2008)*, Manchester, p. 86-93, 2008.

Gaume B., Mathieu F., Navarro E., "Building Real-World Complex Networks by Wandering on Random Graphs", *I3: Information Interaction Intelligence*, vol. 10, n° 1, p. 73-91, 2010.

Guilbert L., Lagane R., Niobey G. (eds), *Le Grand Larousse de la langue française 1971-1978*, Larousse, 1971-1978.

Hartigan J., Wong M., "A K-Means Clustering Algorithm", *Applied Statistics*, vol. 28, p. 100-108, 1979.

Hill F., Reichart R., Korhonen A., "SimLex-999: Evaluating Semantic Models with (Genuine) Similarity Estimation", *CoRR*, 2014.

Huang C.-R., Chang R.-Y., Lee S.-B., "Sinica BOW (Bilingual Ontological WordNet): Integration of Bilingual WordNet and SUMO", *4th International Conference on Language Resources and Evaluation*, May, 2004.

Joubert S., Brambati S., Ansado J., Barbeau E., Felician O., Didic M., Lacombe J., Goldstein R., Chayer C., Kergoat M., "The Cognitive and Neural Expression of Semantic Memory Impairment in Mild Cognitive Impairment and Early Alzheimer's Disease", *Neuropsychologia*, vol. 48, n° 4, p. 978-988, 2010.

Kučera H., Francis W., Carroll J., Twaddell W., *Computational Analysis of Present-Day American English*, Brown University Press, Providence, RI, 1967.

Lafourcade M., "Making People Play for Lexical Acquisition with the JeuxDeMots prototype", *SNLP'07: 7th Int. Symposium on NLP*, Pattaya, Thailand, 12, 2007.

Landis J. R., Koch G. G., "The Measurement of Observer Agreement for Categorical Data", *Biometrics*, vol. 33, p. 159-174, 1977.

Levenshtein V., "Binary Codes Capable of Correcting Deletions, Insertions and Reversals", *Soviet Physics Doklady*, vol. 10, n° 8, p. 707-710, 1966.

Li C. N., Thompson S. A., *Mandarin Chinese: a Functional Reference Grammar*, University of California Press, Berkeley, 1981.

Maffre T., Bregeon C., Garrigou C., Barry I., Raynaud J. P., Duvignau K., "The Comprehension of Verb-Based Metaphors by Children with Pervasive Developmental Disorder (PDD): A Marker of Lexical Rigidity?", *Neuropsychiatrie de l'enfance et de l'adolescence*, vol. 60, n° 5, p. 198-218, 2012.

Mei J.-J., Zheng Y.-M., Gao Y.-Q., Yin H.-X., *TongYiCi Cilin*, Commercial Press, Shangai, 1984.

Méligne D., Fossard M., Belliard S., Moreaud O., Duvignau K., Démonet J., "Verb Production During Action Naming in Semantic Dementia", *Journal of Communication Disorders*, vol. 44, p. 379-391, 2011.

Morrison C., Chappell T., Ellis A., "Age of Acquisition Norms for a Large Set of Object Names and Their Relation to Adult Estimates and Other Variables", *The Quarterly Journal of Experimental Psychology Section A*, vol. 50, n° 3, p. 528-559, 1997.

Motter A. E., Moura A. P. S., Lai Y. C., Dasgupta P., "Topology of the Conceptual Network of Language", *Physical Review E*, vol. 65, p. 065102, 2002.

Murray G. C., Green R., "Lexical Knowledge and Human Disagreement on a WSD Task", *Computer Speech & Language*, vol. 18, n° 3, p. 209-222, 2004.

Nelson D., McEvoy C., Schreiber T., "The University of South Florida Free Association Rhyme, and Word Fragment Norms", *Behavior Research Methods, Instruments, & Computers*, vol. 36, n° 3, p. 402-407, 2004.

Newman M. E. J., "The Structure and Function of Complex Networks", *SIAM Review*, vol. 45, p. 167-256, 2003.

Robert P., Rey A. (eds), *Dictionnaire alphabétique et analogique de la langue française*, 2e éd., Le Robert, 1985.

Stewart G. W., Perron-Frobenius Theory: a New Proof of the Basics, Technical report, University of Maryland at College Park, College Park, MD, USA, 1994.

Steyvers M., Tenenbaum J. B., "The Large Scale Structure of Semantic Networks: Statistical Analyses and Model of Semantic Growth", *Cognitive Science*, vol. 29, p. 41-78, 2005.

Watts D. J., Strogatz S. H., "Collective Dynamics of Small-World Networks", *Nature*, vol. 393, p. 440-442, 1998.