Graph representation of synonymy and translation resources for cross-linguistic modelisation of meaning*

Benoît Gaillard[a], Yannick Chudy[a], Pierre Magistry[b,c], Shu-Kai Hsieh[c], Emmanuel Navarro[d]

[a] CLLE-ERSS, University of Toulouse II - Le Mirail, Toulouse, France
benoit.gaillard@univ-tlse2.fr, yannick.chudy@gmail.com
[b] Alpage, INRIA Paris-Rocquencourt & Université Paris 7, Paris, France.
magistry@gate.sinica.edu.tw
[c] Graduate Institute of Linguistics, National Taiwan University, Taipei, Taiwan
shukai@ntu.edu.tw
[d] IRIT, University of Toulouse III, Toulouse, France
navarro@irit.fr

Abstract In this paper we describe the data that will be used to compare the semantic structures that emerge from synonymy in French and in Mandarin. We aim at studying these semantic structures at both a global, lexicographic level, using lexicons, synonymy and translation dictionaries and at a more localised, experimental level, using data collected in parallel psycholinguistic experiments in French and Mandarin. After presenting our research project, the data we need to carry it out and the available resources, we analyse several linguistic issues arising from the structural differences between the French and Mandarin lexicons. We then explain the construction of the synonymy and translation networks from the available resources and detail specific choices that will enable us to produce meaningful experimental results based on this prepared data. Two kinds of networks are built: lexicographic networks and smaller movie-based networks extracted from experimental recordings. We conclude by describing how we intend to use this data.

Keywords: Synonymy, graphs, lexicon, translation, concept similarity

1 Introduction

This work describes the resources built to be used in a project that aims to compare, or contrast, synonymy structures across languages. We suppose that we can trace concepts of a community of speakers by detecting salient patterns in the networks constituted by all the words of the language, linked by synonymy. There is a long standing debate on the question whether there exist universal cognitive concepts (Pinker, 1994) whether they necessarily translate into language, or on the contrary whether the absence of universal concepts in the variety of languages in the world calls for an entirely different account of human cognition (Evans and Levinson, 2009). Evans and Levinson also argue that understanding the diversity of languages is fundamental to a realistic account of human cognition by psycholinguistics. We attempt, in this project, to move one step further than the debate over linguistic universals, and to create a measure of similarity between lexico-semantic concepts across languages. In this approach, the notion of a common concept, and a fortiori of universal concepts is broadened, because, more than asserting whether concepts are common to several languages, we aim to measure their similarity which can cover the whole range of possible values.

* The research presented in this paper is supported by the ANR-NSC (France-Taiwan) bilateral project M3 (Modeling and Measurement of Meaning).

What is a concept emerging from the paradigmatic structure of the lexicon of a language remains to be defined. In order to model this , paradigmatic structure we study synonymy graphs, that are networks of words connected by an edge if they are synonyms. This approach to studying lexicons has proved to be fruitfull for example in (Gaume, 2008; Gaume *et al.*, 2008). The necessary lexical knowledge to build such networks is usually extracted from electronic dictionaries of synonyms, thesauruses, lexical databases such as WordNets or even participatory resources such as Wiktionary (Navarro *et al.*, 2009; Sajous *et al.*, 2010). These networks, being *real word complex networks*, have Hierarchical Small World (HSW) characteristics (Gaume, 2008). One of these characteristics is a strong clustering: there are salient sets of vertices (clusters or communities) that are significantly more connected to each other than to the rest of the network. Formally defining and detecting such clusters is out of scope of this paper but is currently the subject of intense research (Fortunato, 2010). We assume that these clusters denote concepts, as they highlight semantic associations and sets of words with a strong confluence as defined in (Gaume *et al.*, 2008).

The cross-linguistic study of semantic associations that we will carry out in subsequent stages of this research project is based on lexical networks. The subject of this paper is the construction of these networks from available lexicographic resources: we describe the construction of a French synonymy graph, a Mandarin synonymy graph and a Mandarin-French translation graph.

So far we have introduced the long term goal of our research in cross-linguistic lexico-semantic similarity analysis at the global level of lexicons of whole languages. However as seen in (Evans and Levinson, 2009) the in-depth cross-linguistic analysis of some specific semantic concepts is interesting to study language acquisition. Numerous experimental studies of lexico-semantic typology have been carried out on small subsets of the lexicon (Vanhove, 2008; Koptjevskaja-Tamm, 2008; Levinson *et al.*, 2003). We will also focus on a subset of the lexicon that was collected with the movie experiment: children and adults of French and Mandarin language were asked to describe the same 17 movies. Based on the sets of verbs produced in this experiment, we built the French an Mandarin synonymy subgraphs. Synonymy relations were extracted from entire synonymy graphs. We also built Mandarin-French translation graphs, both as a restrictions of the entire translation graph and as graphs based on more exhaustive, human annotated, translation data. This was feasible because the size of the considered data was reasonable. We will subsequently refer to these (sub)graphs as *movie graphs*.

The next section is devoted to describing the lexicographic and experimental resources on which the construction of the graphs is based. Section 3 describes specific linguistic issues we had to address to be able to compare Mandarin and French lexicons. Section 4 describes the construction and filtering of the graphs that model the lexicons of Mandarin and French. Section 5 describes the construction and filtering of graphs that model a sub-part of the lexicons. Section 6 evaluates the characteristics of the resulting complex networks. We detail in Section 7 how we plan to use these resources and eventually (Section 8) propose some remarks on the quality of these resources and whether it would be possible to improve it.

2 Available resources

2.1 Lexicographic resources for synonymy

Our resource for French synonymy is Dicosyn[1], a compilation of synonymy relations extracted from seven dictionaries (Bailly, Benac, Du Chazaud, Guizot, Lafaye, Larousse and Robert): Dicosyn is a network of synonymy relations as they appear in these dictionaries. There is an edge $r \to s$ if r and s have the same part of speech and a synonymy relation $r \to s$ exists in at least

[1] Dicosyn has been first produced at ATILF (Analyse et Traitement Informatique de la Langue Française), before being corrected at CRISCO laboratory (`http://elsap1.unicaen.fr/dicosyn.html`).

one dictionary. Dicosyn provides three graphs: one for nouns (*DicoSyn.Noun*), one for verbs (*DicoSyn.Verb*) and one for adjectives (*DicoSyn.Adjective*).

The resource we used to build the Mandarin synonym graph is the *CilinCWN.verb*, a graph of verbs extracted from *CilinCWN*, a fusion of the Chinese Wordnet (CWN) and the Chinese thesaurus *TongYiCi CiLin* (Cilin). The Chinese Wordnet is a lexical resource modelled on Princeton WordNet, with many novel linguistic considerations for Chinese. It was proposed and launched by (Huang *et al.*, 2004), and, up to February 2010, it contained 10,533 lemmas with 30,898 senses, and 41,169 lexical semantic relations in total. Among them there are 28,815 synonyms. 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 70000 lexical items under 12 broad semantic classes marked from A to L. These broad classes are further divided into 94 subclasses, and 1,428 heads. But in our experiment, classes of A-E and L were removed, for they refer to non-verbal entities like human, physical object, time and space, features, etc. In order to compare with French graphs, data have been preprocessed to comply to the format of *DicoSyn.Verb* , described in the paragraph above.

2.2 Lexicographic resources for translation

Translation resources of quality for Mandarin are not easily built and few are available. To be consistent with the methodology we used to build our synonymy graphs, we chose to rely on handmade dictionaries that are less noisy than corpus-based automatically aligned lists of translation pairs. However we had to face the issue that translation dictionaries typically not only give a list of translation words but may also give some explanations in a gloss. Since vertices of our graphs are lexemes, we only considered parts of the entries that consist in list of words. We used the StarDict[2] dictionary which is freely available in a digital format and contains for each Mandarin word a list of French words that can be unambiguously isolated.

2.3 The movie experiment

Lexical data in the form of lists of verbs in French and Mandarin have been produced based on the movie experiment. The movie experiment consisted in asking adults and children to describe the action performed in a movie (Magistry *et al.*, 2009). Each verb they used was systematically registered in the list of verbs proposed for that particular movie. The same 17 movies were shown to French and Mandarin speaking children and adults. This data has been used to build the synonymy and translation movie graphs.

For each movie, all the proposed French verbs were paired with all the proposed Mandarin verbs for this movie. Three Taiwanese students of French have indicated whether these pairs could be translations of each others or not. In case of a French verb being translation of only one part of the Mandarin verb, that part was also noted by the students. For two movies (in which somebody tears off a newspaper, and in which somebody breaks a glass) the verb pairs have been annotated by the three students and verified by a Mandarin native speakers working in tandem with a French native speaker (Laurent Prévot and Tsyr-Huei Chiang), two members of the team working on the Franco-taiwanese ANR-NSC M3 joint project "Models and Measurement of meaning". The word pairs extracted from the files of the fifteen other movies have been annotated by only one student and have not been verified. However, the verification and cross checking performed on the two special movies enables us to relatively trust the quality of the unverified translations. This experimental resource provides us with 133 French verbs and 399 Mandarin verbs linked by 1160 translations annotated as possible translations. The construction and exploitation of this translation resource is described in details in (Prévot *et al.*, 2010).

[2] http://StarDict.sourceforge.net/

3 Linguistic issues arising from structural discrepancies between Mandarin and French lexicalization

3.1 Comparing action-result constructed Mandarin verbs to plain French verbs

The cross-linguistic study of semantic associations between French and Mandarin will focus on verbs because the experimental resource is verb-oriented. Mandarin is well-known for its Serial Verbs Constructions (SVC) and Resultative-Verb Compounds (RVC) that have both been deeply studied in the Mandarin linguistics literature (Thompson, 1973). Since we tried to stay at a lexical level, we did not address SVC that are clearly syntactic constructions of multiple separated lexemes. However the RVC were an important issue that we had to face, because a large part of our data consists of RVC (around 50%). For instance, the movies used in the experimental part show a woman performing some action on an object that results in a change of state of the object. The native Mandarin speakers naturally produced a lot of Type III RVCs (as classified by (Chang, 2003). Type III RVCs are compound verbs of two parts (we call $V1$ and $V2$) where $V1$ depicts an action that has been performed by an agent (typically in subject position) on a patient (typically in object position) and $V2$ depicts the change of state of the patient. See (1) for an example where $V1(da)$ means *to hit* and $V2(po)$ means *to break.*

(1) ta da-po le yi_ke qiqiu
she hit-break Perf. one balloon

A more detailed account of the issue of RVCs in the case of our project has been given in (Magistry *et al.*, 2009). The main point is that RVCs are semantically compositional and productive, and it has been shown that even very young children can produce novel resultative constructions (Erbaugh, 1992). Therefore not all $V1-V2$ compounds can be found in dictionaries which makes them difficult to align with French data.

French do not have this kind of productive action-result construction. French verbs are much more lexicalized and one verb can depict the action only (ex: "*scier*", *to saw*), the result only (ex: "*casser*", *to break*) or both. The semantics of French verbs also present complex interactions with flexional morphology and possible syntactic constructions enabling diathesis alternation that can not be covered here. Note that it is possible to syntactically build action-result constructions in French (ex: "*couper en deux*", *to cut into two pieces*) but our experimental data shows that this possibility is seldomly used by French native speakers whereas it constitutes a large part of the data produced by Mandarin native speakers.

Since our graphs are mostly based on available lexical resources, Mandarin RVCs are present in our graphs just as they are in dictionaries, following other lexicographers choices. Some typical combinations and highly frequent RVCs may enter the dictionaries (and may progressively loose the compositionality property), those are included in our graphs. The more compositional RVCs used in a productive fashion are more likely to be treated as two distinct dictionary entries ($V1$ and $V2$) and will be present as two vertices in our graphs. For experimental graphs, we also allowed ourselves to manually include experimentally recorded compounds into our graphs.

3.2 Homographs and homonyms in French and Mandarin

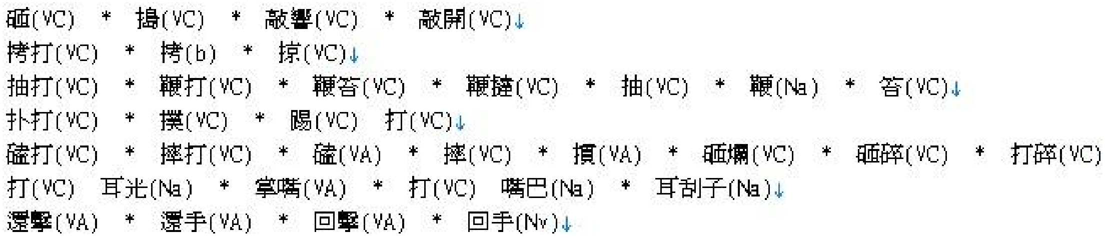
Another issue comes from the ideographic property of Chinese writing. Lexicographically speaking, the prevailing groups of words in Chinese are *homographs*, which are usually defined as a group of words that share the same spelling (i.e., ideographic form), regardless of their pronunciation and meaning. If they are pronounced the same, they are also called *homophones* (and *homonyms*). For instance, *da3* *ren2* ("*to beat*") and *yi1-da3* ("*a dozen*"); If they are pronounced differently, they are also called *heteronyms*, like *huei4-yi4* ("*meeting*") and *kuai4-ji4* (accounting). Conversely, in dictionaries, one can find many *polysemes*. A polyseme is a lemma that has several distinct meanings. The difference between homonyms and polysemes is thus subtle, that is, how they are treated as several different meanings (polysemes) within a single dictionary lemma, or treated as several separate lemmas. In Chinese Wordnet, the decision is made based on Chinese lexicological tradition, however, in this experiment, we take the translations of the various homographs as the same lemma: In order to build graphs from resources containing homographs, a simple approach consists in modelling all homographs as polysemes, that is, grouping all homographs as one single lemma, that will produce a single vertex. This single vertex then has for a label the common lemma of this set of homographs and for neighbours the unions of the synonyms of all these homographs. We call this simplifying approach the *flattening approach*.

4 Construction of synonymy and translation graphs from lexicographic resources

In the following we present synonymy graphs for French and Mandarin. These graphs are reflexive[3] and symmetric[4]. We also present translation graphs which are also symmetric but not reflexive.

4.1 Synonymy graphs

From CWN, we chose words containing verb senses and synonymy relations (here the variant and near-synonym relations were included), i.e., verbs without synonymy relation were discarded. From Cilin, words whose length were less than 4 were chosen, and then tagged with POS by Academia Sinica CKIP tool[5], the Chinese segmentation and tagging system. Figure 1 shows some of the results. Note that we only used CKIP to tag words and then choose the ones whose POS was "verb", and not to separate words. Since there are some phrases in Cilin, CKIP's segmentator would have been able to segment, from phrases, words of length equal to or greater than 4, but these words are not what was needed in this study, we disregarded them.



砸(VC) * 搗(VC) * 敲響(VC) * 敲開(VC)↓
拷打(VC) * 拷(b) * 掠(VC)↓
抽打(VC) * 鞭打(VC) * 鞭笞(VC) * 鞭撻(VC) * 抽(VC) * 鞭(Na) * 笞(VC)↓
扑打(VC) * 撲(VC) * 踢(VC) 打(VC)↓
磕打(VC) * 摔打(VC) * 磕(VA) * 摔(VC) * 摜(VA) * 砸爛(VC) * 砸碎(VC) * 打碎(VC)
打(VC) 耳光(Na) * 掌嘴(VA) * 打(VC) 嘴巴(Na) * 耳刮子(Na)↓
還擊(VA) * 還手(VA) * 回擊(VA) * 回手(Nv)↓

Figure 1: Example of Cilin lexemes tagged with the CKIP tool. Lexemes followed by the tag (VC) are verbs.

Then we compute $w_{i,j}$, the *weight* between every two verb pair (v_i, v_j), by counting their co-occurrences in synsets: the weight means how many senses of (v_i and v_j) are synonyms of each other (or variant or near-synonym) in CWN (and in Cilin). The weight is assigned to 1 when $i = j$ to make the graph reflexive, and assigned to 0 if there are none of their senses have a synonymy relation. Note that a word without synonymy relation will have no weight with others and will be a lonely node without any edge in the graph. Words of this kind are thus discarded. In total, there are 3,499 verbs and 16,815 verb relations (weight is not 0) from CWN, and 12,145 verbs and 100,657 relations from Cilin. By merging them (sum the weight from CWN and Cilin), a Chinese synonymy graph with 13,439 verbs and 111,985 relations is produced. We then simplified the weight by setting every weight greater than zero to one. Figure 2 shows an example of the synonymy links of *kan4* (*see*).

[3] A graph is reflexive if there is a self-loop on each vertex

[4] When referring to the number of edges we refer to the total number of *directed* edges i.e. twice the number of *undirected* relations plus the number of self-loops

[5] http://rocling.iis.sinica.edu.tw/CKIP/engversion/index.htm

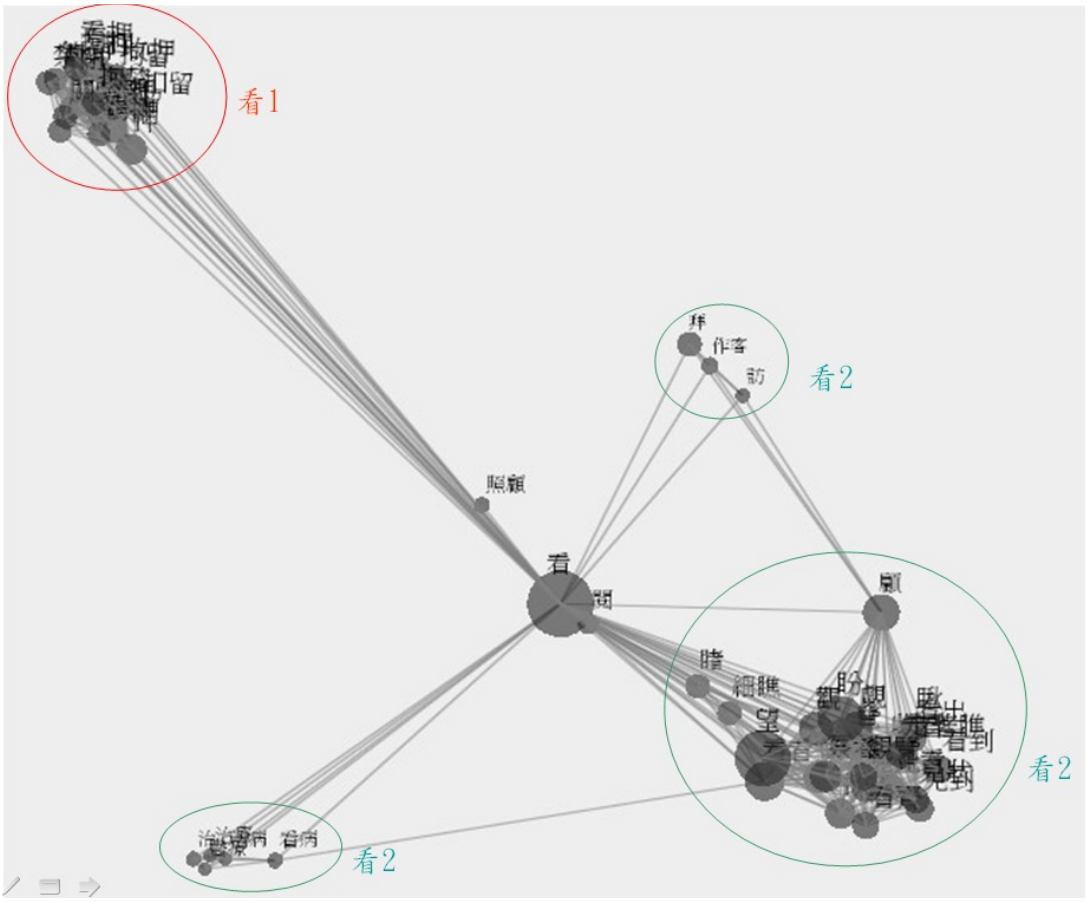


Figure 2: Example of synonymy links around one lexeme (*kan4*) in the *CilinCWN* graph.

Since the CilinCWN was updated several times, several versions of the Mandarin synonymy graph have been created. Between the version we built in the beginning of July and the most recent version, the density of arcs has increased whereas the number of vertices has decreased by about 1000 verbs. The most recent version, called *CilinCWN_20100823* is of better quality as the significantly better results of some experiments based on this resource show in (Desalle *et al.*, 2010). In the July version of the Cilin-based graph, homographs were considered to be different verbs. In order to maintain consistency in our experiments we therefore had to *flatten* it as explained in section 3.2. Three graphs based on the Cilin and Chinese Word Net resource were built:

- *CilinCWN_20100703*, in the beginning of July 2010, has 14307 vertices and 109562 edges
- *CilinCWN_flat_20100703* has 14042 vertices and 109333 edges
- *CilinCWN_20100823*, in the end of August 2010, has 13439 vertices and 111985 edges

The French synonymy graph is simply the *DicoSyn.Verb* resource described in 2.1. It has *9147* vertices and *111993* edges, the Mandarin synonymy graph (*CilinCWN_20100823*) has 13439 vertices and 111985 edges. In order to compare clusters across these two graphs, it would be better to have comparable structures in terms of size and edge density. We will discuss in the conclusion how we could filter their edges and vertices to adjust their structures.

4.2 Translation graph

Using the StarDict bilingual dictionary, we built a graph reduced to the sets of vertices of *DicoSyn.Verb* and *CilinCWN_20100823*. Using the flattening approach described in section 3.2, we created an edge between two vertices in the graph if there exists at least one meaning of one lexeme that can be translated by one meaning of the other lexeme. We had to convert Simplified Chinese to Traditional Chinese. The conversion was semi-automatic, manually checked when needed. We obtained a translation bi-graph[6] called *frzh_StarDict_fCC_fDS* with 6096 vertices, of which 2254 French verbs and 3842 Mandarin verbs, connected by 14742 edges.

4.3 Vertex filtering of synonymy graphs

Some verbs are present in synonymy graphs but not in the translation graph (*frzh_StarDict_fCC_fDS*). We then had to reduce the two synonymy graphs (*CilinCWN_20100823* and *DicoSyn.Verb*) to the set of vertices of the translation graph. Since the translation graph was already filtered by the synonymy graphs, we obtained vertices that are in the intersections of the synonymy and translation resources. We obtained the 2 following subgraphs:

- *dycosyn_fSD*, a French synonymy graph with *2269* vertices and *39787* edges.
- *CilinCWN_20100823_fSD* a Mandarin synonymy graph with *3842* vertices and *23600* edges.

5 Construction of movie synonymy and translation graphs

5.1 Translation graphs

We compiled the movie translation resources described in 2.3. To build the annotation-based bi-graph of translation, every French and Mandarin verb found in the movie file became a vertex, and an edge was added between two verbs if they were annotated as a correct translation by at least one student. In addition to this experimentally built translation graph, we have built a translation graph based on the same experimental verb lists and on the StarDict translation bi-graph. Here, we added an edge if it was present in the entire translation graph *frzh_StarDict_fCC_fDS*.

Table 1: Number of vertices (n) and edges (m) of the translation graphs. *frzh_StarDict_fCC_DS* and *frzh_movies_annotated_fCC_fDS* are the entire translation graph extracted from the StarDict resource and the movie translation graph, reduced to the vertices of CilinCWN and DicoSyn.Verb. The following four graphs are based on the movie experiment's data. *frzh_movies_annotated* and *frzh_movies_annotated_fSD* are translations deemed valid by student annotations, whereas *frzh_movies_StarDict_fSD* is build from StarDict. Vertices of *frzh_movies_annotated_fSD* are filtered with StarDict vertices.

$name$	n	n_{fr}	n_{zh}	m
frzh_StarDict_fCC_fDS	6096	2254	3842	14742
frzh_movies_annotated	410	113	297	1160
frzh_movies_annotated_fCC_fDS	158	73	85	570
frzh_movies_annotated_fSD	109	54	55	160
frzh_movies_StarDict_fSD	109	54	55	69

In order to compare the StarDict-based translations with the annotated translations, we built two graphs with the same vertices: the intersection of the vertices of the movie-based translation graph *frzh_movies_annotated* and the vertices of the StarDict translation graph *frzh_StarDict_fCC_fDS*.

Only 25% of edges of *frzh_movies_annotated_fSD* could be found in *frzh_movies_StarDict_fSD* (*Recall*) whereas 59% of edges of *frzh_movies_StarDict_fSD* could be found in

[6] A bi-graph (or bipartite graph) is a graph with two distinct kinds of vertices, with edges that only link one kind to the other.

frzh_movies_annotated_fSD (*Precision*). These two measures, combined gave us a measure of agreement between the two graphs (*F-score*) equal to 0.17, which is quite low.

The recall of StarDict edges compared to annotation edges is small whereas their precision is acceptable. This means that, provided that the annotated translation are a reference, StarDict lacks many translations. It shows that the polysemic richness of verbs, very visible in various contexts such as the movies context, is not well captured by dictionaries that only list words out of context.

That the precision figure is not greater is surprising. Why would experts decide that two verbs can not be translated by each other when an official dictionary provides this translation ? This is partly due to the experimental protocol. Pairs were proposed to the annotators on the basis of the French and Mandarin verbs proposed for the same movie. Pairs that would link words across movies were not proposed to the annotators. Half of the StarDict translations missing in the annotated graph are these "cross movie" pairs that annotators did not get a chance to validate or dismiss.

5.2 Movie synonymy graphs

For the construction of the movie synonymy graph, we selected from *DicoSyn.Verb* the subgraph whose vertices are French verbs present in the movie files and translated (by *frzh_movies_annotated*) by a Mandarin verb that exists in the *CilinCWN_20100823* graph. We called this graph *fr_movies_fDS*. It contains $V = 64$ verbs and $E = 430$ edges. The small number of vertices (64 out of 133 French verbs in the movie files) can be explained by the fact that most movie descriptions were actually verbal expressions such as "*casser en mille morceaux*", when often only the verb "*casser*" could be found in *DicoSyn.Verb*. We notice here that the "*en mille morceaux*" phrase expresses the result of the action, quite similarly to what the $V2$ part of Mandarin verbs does. As we will stress in the conclusion of this paper, this issue might justify a French verbal lexicon based on the $V1 - V2$ construction of Mandarin verbs.

We used the same method to build Mandarin movie synonymy graphs. We obtained two Mandarin synonymy graphs based on *CilinCWN_20100703* and *CilinCWN_20100823* . The first graph, called *zh_movies_CilinCWN_20100703* is made of $V = 83$ verbs and $E = 280$ edges and the second, called *zh_movies_CilinCWN_20100823* is made of $V = 85$ verbs and $E = 382$ edges. We have more vertices in this graph than in *fr_movies_fDS*, but 85 verbs out of 399 Mandarin verbs present in the movie files is still a small amount. This is due to the fact that verbs expressed by Mandarin children are often not fully correct, especially in their $V2$ part, and that they consequently do not belong to the CilinCWN lexicon. This results seems to point at the learning process of children, who would grasp the $V1$ senses first and then proceed on to refining their use of $V2$ parts over many years.

6 Hierarchical Small World properties of the resulting graphs

Most of lexical networks, as other real world complex networks, are Hierarchical Small Worlds (HSW) networks (Watts and Strogatz, 1998; Albert and Barabasi, 2002; Newman, 2003; Gaume *et al.*, 2010) sharing similar properties.

The four main properties of HSW are the following:

- **Edge sparsity:** HSW are sparse in edges, the number of edges is in the same order as the number of vertices.
- **Short paths:** in HSW, the average path length[7] L is short. There is generally at least one short path between any two vertices.
- **High clustering:** in HSW, the clustering coefficient (C) that expresses the probability that two distinct nodes adjacent to a given third one are adjacent, is an order of magnitude higher

[7] Average length of the shortest path between any two nodes.

than for Erdős-Rényi (random) graphs: this indicates that the graph is locally dense, although it is globally sparse.

- **Heavy-tailed degree distribution:** in a HSW graph few vertices account for a large number of neighbours whereas others only have a few connections. This degree distribution often fits a power-law distribution: the probability $P(k)$ that a given node has k neighbours decreases as a power-law: $P(k) \approx k^{-\lambda}$ (λ being a constant characteristic of the graph).

Table 2 sums-up the structural characteristics of several synonymy graphs presented in section 4 and 5. In this table, λ is the coefficient of the power-law that approximates the distribution of the nodes incidence degrees with a correlation coefficient r^2. When the values are computed on the *largest connected component* they are subscripted by $—_{lcc}$. Other notations are explained above.

Translation graphs are bi-graphs to which these measures do not apply in this straightforward manner. However, their numbers of edges and vertices for each language are synthesized in table 1.

Table 2: Pedigrees of the synonymy graphs. Clustering (C), average shortest path (L), power law coefficient (λ) and confidence coefficients (r^2) have been measured on the largest connected component (lcc). The first two graphs model French synonymy, the following four model Mandarin synonymy. *CilinCWN_20100823* is filtered with StarDict based translation graph to obtain *CilinCWN_20100823_fSD*. The last three graphs model synonymy as visible in the movie experiment, they are built from *DicoSyn.Verb* and the two versions of the *CilinCWN graph*.

$name$	n	n_{lcc}	m	m_{lcc}	L_{lcc}	C_{lcc}	λ_{lcc}	r^2_{lcc}
DicoSyn.Verb	9 147	8 993	111 993	111 659	4.20	0.14	-2.02	0.93
DicoSyn.Verb_fSD	2 269	2 221	39 787	39 735	3.31	0.17	-1.65	0.80
CilinCWN_20100703	14 307	8 113	109 562	90 137	5.91	0.64	-2.27	0.85
CilinCWN_flat_20100703	14 042	8 080	109 333	90 548	5.69	0.62	-2.32	0.86
CilinCWN_20100823	13 439	8 393	111 985	94 316	5.65	0.61	-1.79	0.61
CilinCWN_20100823_fSD	3 842	2 658	23 600	21 046	5.25	0.44	-2.17	0.88
fr_movies_fDS	64	62	430	428	2.80	0.33	-0.94	0.50
zh_movies_fCC-20100703	83	47	280	217	3.68	0.54	-1.26	0.69
zh_movies_fCC-20100823	85	70	382	363	3.50	0.32	-0.54	0.18

Except movies graph's that are too small to have reliable statistics, all graphs are HSW. Note that *CilinCWN_20100823*'s degree distribution does not fit well a power-law. However as one can see in Figure 3, it clearly is a heavy-tailed distribution.

All graphs of Mandarin have a higher clustering coefficient than French ones. It means that there are more triangles in these graphs than in French ones. This may be explained by a difference in the construction methodology of initial resources rather than by a difference in the synonymy structure of the two languages. Indeed WordNet synsets tend to create numerous triangles in the resulting synonymy graph, whereas such phenomena don't happen in DicoSyn.

The power law coefficient λ is almost the same in Mandarin and in French graph: close to -2 and slightly higher in Mandarin than in French graphs. L is clearly longer in Mandarin graphs. Nevertheless it's hard to draw some conclusion about these figures: do they stress real linguistic invariants and variations, or are discrepancies just caused by a difference of resources modelisation (as it seems to be for the clustering coefficient) ?

7 Perspectives: cross-linguistic similarity of lexico-semantic concepts

Our longer term objective in this project is to evaluate the similarity of lexico-semantic concepts of Mandarin and French, based on the structure of the synonymy graphs that model their lexicon. But if evaluated the similarity between to synonymy graphs without translation information, we wouldn't know which French word to associate with which Mandarin word. This exercise exists

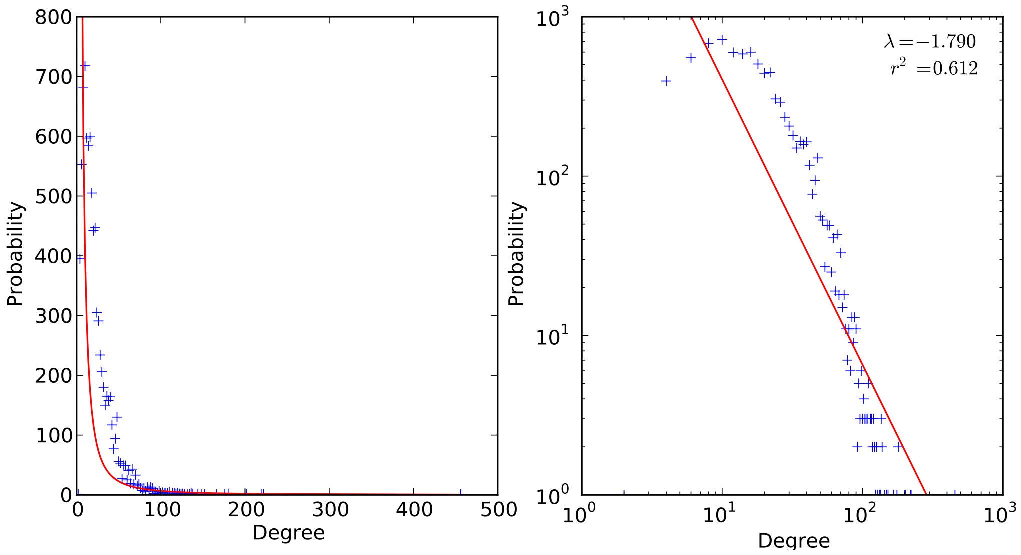


Figure 3: Degree distribution of the graph *CilinCWN_20100823*. Despite the coefficient correlation with the power-law fitting is relatively smaller than the other graph's coefficient, one can see that it's still an heavy tailed distribution.

as the search of graph isomorphisms (Barecke, 2009; Sorlin, 2006), but as we will show in a subsequent experimental article, is not relevant to lexico-semantic similarity analysis. To solve this issue we have to use translation links. For each Mandarin synonymy cluster, we will define the set of French words that are the *trace* of this cluster via translation. We will then compare them to the synonymy clusters of French words. In parallel we will perform the same operation on the Mandarin side. We will produce two lists of pairs of concepts: French concepts vs. traces of Mandarin concepts, and Mandarin concepts vs. traces of French concepts. Then, for each French or Mandarin concept, we can find the trace of a (Mandarin or French) concept that has the maximal similarity. This similarity value is the "trace value of a concept on a target language, via translation". Then, the mean trace value is a measure of the conceptual similarity between two languages[8].

These measures will first be applied to the entire verb lexicons of French and Mandarin. They will be based on the French to Mandarin translation bi-graph *frzh_StarDict_fCC_fDS*, the French synonymy graph, filtered with the translation graph vertices *Dicosyn_fSD* and the Mandarin synonymy graph, filtered with the translation graph vertices *CilinCWN_20100823_fSD*. The same set of measures we will be performed on the 3 following movie graphs: the translation graph, built with annotated movie pairs, *frzh_movies_annotated*, the French movie-based synonymy graph *fr_movies_fDS* and the Mandarin movie-based synonymy graph *zh_movies_fCC-20100823*.

8 Conclusion

Building resources to compare lexico-semantic concepts across languages is a very complex task, because from the very beginning issues on how to model lexicons are raised. Models not only have to be consistent across the various experiments but, more importantly in a cross-linguistic study, they have to be consistent across languages that do not necessarily have comparable lexicalizations.

One major issue we had to tackle was the decomposition of Mandarin verbs in two parts $V1$ and $V2$ as described in section 3.1. Most French verbs do merge in one single form the action

[8] Note that the mean trace value is not symmetric.

and the result description and therefore there is no $V1 - V2$ formalism for French verbs. For the movie graphs, we have modelled that a Mandarin verb is any $V1$, $V2$ or $V1 - V2$ that is the translation of a French verb. However we realised that many occurrences of French speakers did not belong as such to the lexicon formalised by *DicoSyn.Verbs*, despite being translatable by a Mandarin verb. Such expressions as "*déchirer en deux*" (*tear into two pieces*) are not part of the lexicon but do translate well into Mandarin. See section 5.2. Therefore, it seems that, instead of simplifying the $V1 - V2$ Mandarin structure into "*verbs*", it would be interesting to build a French lexicon that takes the $V1 - V2$ structure into account. Another issue was raised by homographs as seen in section 3.2. We decided to flatten these homographs out because this is what is usually done in French where homographs that are not also homophones hardly exist, and therefore most homographs are actually polysemes, not homonyms.

We also have noticed that experimental results are very sensitive to which version of the Cilin-extracted synonymy graph is used. The latest version had about 1000 verbs less and 2000 edges more than the earliest, which led to significantly different results in (Desalle *et al.*, 2010)'s work. We suspect therefore that the requirement for comparable density and number of vertices between the *StarDict* filtered French and Mandarin synonymy graphs is important. Some more work is thus necessary to meet this requirement. An idea would be to consider the frequencies of verbs in various corpuses in order to filter vertices, and to filter synonymy links according to their weights (see Section 4).

The sensitivity of experimental results to slight differences in the resources points to questioning whether modelling lexicons with simple, binary synonymy links is a reliable enough approach to study semantic associations. Some answers to it were introduced in (Gaume *et al.*, 2008) where authors propose to use *proxemy* and *confluence* as a way to abstract the semantic modelling of a lexicon from their fickle synonymy graph representation. Along this line we are currently investigating in (Gaillard *et al.*, 2010) how the notion of near-synonymy (Edmonds and Hirst, 2002) could be grasped by *proxemy* approaches.

References

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

Barecke, Thomas. 2009. *Isomorphisme inexact de graphes par optimisation évolutionnaire*. Ph.D. thesis, Université Pierre et Marie Curie - Paris VI, Octobre.

Chang, Jung-Hsing. 2003. Event structure and argument linking in chinese. *Language and Linguistics*, 4(2), 317–351.

Desalle, Yann, Bruno Gaume, Shukai Hsieh, and Hintat Cheung. 2010. Toward an automatic measurement of verbal lexicon acquisition: the case for a young chidren vs adult categorization in French and Mandarin. In *Proceedings of the Workshop on Models and Measurement of Meaning, at PACLIC conference*, Sendai, Japan.

Edmonds, Philip and Graeme Hirst. 2002. Near-synonymy and lexical choice. *Computational Linguistics*, 28(2), 105—144.

Erbaugh, M., 1992. *The acquisition of mandarin*, pp. 373–455. Lawrence Erlbaum, Mahwah.

Evans, Nicholas and Stephen C. Levinson. 2009. The myth of language universals: language diversity and its importance for cognitive science. *Behavioural Brain Science*, 32(5), 429–48.

Fortunato, S. 2010. Community detection in graphs. *Physics Reports*, 486(3-5).

Gaillard, Benoit, Emmanuel Navarro, and Bruno Gaume. 2010. From binary synonymy to near synonymy by optimal proxemy of lexical resources. In *Proceedings of Workshop on Computational Approaches to Synonymy, at the Symposium on Re-Thinking Synonymy*.

Gaume, Bruno. 2008. Mapping the forms of meaning in small worlds. *International Journal of Intelligent Systems*, 23(7), 848–862, May.

Gaume, Bruno, Karine Duvignau, and Martine Vanhove. 2008. Semantic associations and confluences in paradigmatic networks. In Martine Vanhove, ed., *From Polysemy to Semantic Change: Towards a Typology of Lexical Semantic Associations*, pp. 233–264. John Benjamins.

Gaume, Bruno, Fabien Mathieu, and Emmanuel Navarro. 2010. Building Real-World Complex Networks by Wandering on Random Graphs. *Information - Interaction - Intelligence*, 10(1).

Huang, Chu-Ren, Ru-Yng Chang, and Hsiang-Pin Lee. 2004. Sinica bow (bilingual ontological wordnet): Integration of bilingual wordnet and sumo. In *LREC*, Portugal.

Koptjevskaja-Tamm, Maria. 2008. Approaching lexical typology. In Martine Vanhove, ed., *From Polysemy to Semantic Change: Towards a Typology of Lexical Semantic Associations*, Studies in Language Companion. John Benjamins Publishing Co.

Levinson, Stephen, Sergio Meira, and the Language Cognition Group. 2003. Natural concepts in the spatial topological domain-adpositional meanings in crosslinguistic perspective: and exercise in semantic typology. *Language Linguistic Society of America*, 79(3), 485–516.

Magistry, Pierre, Laurent Prévot, Hintat Cheung, Chien yun Shiao, Yann Desalle, and Bruno Gaume. 2009. Using extra-linguistic material for mandarin-french verbal constructions comparison. In *PACLIC 23: the 23rd Pacific Asia Conference on Language, Information and Computation (PACLIC 2009)*, Hong Kong.

Mei, Jia-Ju, Yi ming Zheng, Yun-Qi Gao, and Hung-Xian Yin. 1984. *TongYiCi CiLin*. Commercial Press, Shanghai.

Navarro, Emmanuel, Franck Sajous, Bruno Gaume, Laurent Prévot, ShuKai Hsieh, Ivy Kuo, Pierre Magistry, and Chu-Ren Huang. 2009. Wiktionary and NLP: improving synonymy networks. In *Proceedings of the 2009 ACL-IJCNLP Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources*, pp. 19–27, Suntec, Singapore, August. ACL.

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

Pinker, Steven. 1994. *The language instinct*. W. Morrow and Co.

Prévot, Laurent, Chun-Han Chang, and Yann Desalle. 2010. Computational modelling of verb acquisition, from a monolingual to a bilingual study. In *Proceedings of Workshop on Models and Measurement of Meaning, at PACLIC conference*, Sendai, Japan.

Sajous, Franck, Emmanuel Navarro, Bruno Gaume, Laurent Prévot, and Yannick Chudy. 2010. Semi-automatic Endogenous Enrichment of Collaboratively Constructed Lexical Resources: Piggybacking onto Wiktionary. In H. Loftsson, E. Rögnvaldsson, and S. Helgadóttir, eds., *Proceedings of the 7th International Conference on NLP (IceTAL 2010)*, volume 6233 of *LNAI*, pp. 332–344, Reykjavik, Iceland. Springer-Verlag.

Sorlin, Sébastien. 2006. *Mesurer la similarité de graphes*. Ph.D. thesis, Univerté Claude Bernard Lyon 1, Novembre.

Thompson, S. A. 1973. Resultative verb compounds in mandarin chinese: A case for lexical rules. *Language*, 49, 361–379.

Vanhove, Martine. 2008. *From Polysemy to Semantic Change: Towards a Typology of Lexical Semantic Associations*. Studies in Language Companion. John Benjamins Publishing Co.

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