

One Lexicon, Two Structures: So What Gives?

Nabil Gader, Sandrine Ollinger, Alain Polguère

▶ **To cite this version:**

Nabil Gader, Sandrine Ollinger, Alain Polguère. One Lexicon, Two Structures: So What Gives?. Seventh Global Wordnet Conference (GWC2014), Jan 2014, Tartu, Estonia. pp.163-171. hal-00937187

HAL Id: hal-00937187
https://hal.science/hal-00937187

Submitted on 28 Jan 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

One Lexicon, Two Structures: So What Gives?

Nabil Gader
MVS Publishing Solutions
Sainte-Marguerite, F-88100, France
nabil.gader@mvs.fr

Sandrine Ollinger
CNRS, ATILF, UMR 7118
Nancy, F-54063, France
sandrine.ollinger@atilf.fr

Alain Polguère
Université de Lorraine, ATILF, UMR 7118
Nancy, F-54063, France
alain.polguere@univ-lorraine.fr

Abstract

We present a reinterpretation of lexical information embedded in the English WordNet in an alternate type of structure called *lexical system*. First, we characterize lexical systems as graphs of lexical units (word senses) connected mainly by Meaning-Text lexical function relations, then introduce a hand-built lexical system: the French Lexical Network or fr-LN, a lexical resource that implements a new lexicography of virtual dictionaries. We later explain how a corresponding en-LN has been generated from the English WordNet. Finally, we propose a topological contrastive analysis of the two graphs showing that both structures can be characterized as being Hierarchical Small World Networks.

1 Introduction

1.1 Context: the French Lexical Network

The RELIEF project (Lux-Pogodalla and Polguère, 2011) is the first stage of a long-term lexicographic enterprise that aims at developing a broad-coverage French lexical resource: the *French Lexical Network*, hereafter fr-LN. This resource possesses two main characteristics.

Firstly, it is the product of actual lexicographic work but does not involve the writing of dictionary articles. Rather, textual dictionary-like descriptions can be automatically generated from linguistic information contained in the fr-LN, which can thus be considered as having embedded in it *virtual dictionaries*. For comparable approaches to the design of lexical resources, see for instance Atkins (1996) and Spohr (2012).

Secondly, it possesses a very specific type of graph structure called *lexical system*, conceptualized in Polguère (2009). While WordNets are before of all **graphs of synsets**, lexical systems are **graphs of specific word senses**—i.e. *lexical units*, in our terminology—, connected by a rich set of lexical relations based on Meaning-Text *lexical functions* (Mel'čuk, 1996; Mel'čuk, 2006). For instance, below is a typical synset relation taken from WordNet:

`{army#1,regular army#1,ground forces#1}` $\xrightarrow{\text{member meronym}}$ `{corps#1, army corps#1}`

whereas only lexical function relations holding between specific word senses such as:

ARMY **1** $\xrightarrow{\text{Sing}}$ CORPS **1**

exist in a lexical system.[1]

In addition, each piece of information in a lexical system (mainly, lexical nodes and lexical function arcs connecting nodes) is supplied with a *trust value*, that is a measure of the validity of lexical information. For instance, information directly entered by lexicographers receive high or, even, maximal trust values, while information automatically generated by analogy-based algorithms should receive a low trust value. This allows for the implementation of "fuzzy" reasoning on lexical information.

At the time of writing, the fr-LN's wordlist contains 14,311 vocable entries—the term *vocable* designates a (potentially) polysemic word—, and 20,791 *lexical units*—actual word senses. Complete statistical data on the fr-LN are provided in

[1] **Sing** is the singulative lexical function.

section 3, including data on lexical function relations that weave the lexical network. Notice that these relations are not the only lexical connections encoded in the fr-LN. Each idiom, i.e. phrasal lexical unit, is connected to the lexemes it formally contains. For instance, the noun POMME DE TERRE 'potato' is connected to the corresponding lexemes POMME 'apple', DE 'of' and TERRE 'soil', via the description of its internal syntactic structure. Additionally, we have just started to encode *copolysemy links*: i.e. metonymy, metaphor, etc. links that connect senses belonging to the same vocable and form its polysemic structure.

1.2 Going English

The goal of this paper is to present an experiment that we have conducted in order to automatically generate an *English Lexical Network*, hereafter en-LN, from the English WordNet. Such task presents some similarity with previous attempts at compiling WordNet into specific data structures—see for instance Graves & Gutierrez (2005) and Huang & Zhou (2007). However, in our case, we "transmute" WordNet data into an informational content that is fundamentally different in nature.

One consequence is that information embedded in WordNet that is "deeper" (more conceptual) than strict linguistic knowledge is lost. This loss of information is compensated by a very important gain: a data structure that allows us to perform lexicographic work on the English lexicon using exactly the same advanced lexicographic tools we are using in our fr-LN project (Gader et al., 2012). In other words, we can perform a lexicographic "graph weaving" activity on both French and English networks (cf. section 4).

The remainder of the paper is organized as follows. Section 2 describes how the task of compiling the English WordNet into an en-LN has been performed. Section 3 presents a contrastive topological analysis of the graph structure of both networks. Section 4 concludes on the practical interest of our experiment.

2 From WordNets to lexical systems

2.1 General characterization of the task

The extraction of an English lexical system out of WordNet's data is a process of bridging the gap between two non-equivalent information structures. The structure of lexical systems has been introduced in section 1.1. The structure of WordNet is well-known (Kamps, 2002) and a presentation in the present context would be overkill. It is however useful to summarize the main formal differences that exist between our source and target structures, i.e. to recapitulate our "one lexicon, two structures" problematics: see Table 1 below.

English WordNet	**en-LN**
Synsets as structural units of description	**Lexical units** as structural units of description
Global partition based on parts of speech (N, V, Adj, Adv)	No part of speech partition
Top-down hierarchical organization	Multidimensional organization
Chiefly based on the hyper-/hyponymy relation between **synsets**	Based on a set of lexical function relations between **lexical units**

Table 1: One lexicon, two structures.

Computationally, our source dataset was the ANSI Prolog version of Princeton WordNet 3.0.

This Prolog version of WordNet is made up of 21 files, each containing a Prolog database that is a set of Prolog "fact" clauses for a given predicate. For instance, the `wn_s.pl` file contains 212,558 clauses for the `s/6` Prolog predicate (the 6-place `s`(ense) predicate), each clause encoding the description of one WordNet sense. The structure of the `s/6` predicate is described as follows in the `prologdb.5.pdf` documentation file:

s(*synset_id*,*w_num*,*'word'*,
ss_type,*sense_number*,*tag_count*).

For example, the following Prolog clause:

```
s(107544351,4,
      'infatuation',n,2,0).
```

asserts that there exists a WordNet nominal sense `infatuation#2`, that is the fourth sense in the synset whose id is `107544351` and that was not semantically tagged in WordNet's Semantic Concordances (Miller et al., 1993).

Out of the 21 Prolog files, 18 have been identified as containing information that could indeed be translated into lexical system data.[2] Such data belong to three main categories: (i) lexical entities (mainly, lexical units and vocables), (ii) individual properties of lexical units (parts of speech, semantic gloses, etc.) and (iii) lexical function relations between lexical units.

[2]The three unused files are: `wn_cls.pl` (class relations between synsets), `wn_sa.pl` (rather heterogeneous relations between verbal or adjectival senses) and `wn_vgp.pl` (similarity relations between verbal synsets).

Next section explains how this information has been generated from WordNet's Prolog files.

2.2 Generation of lexical data

For lack of space, we cannot account for all aspects of the compilation process. We focus on the insertion of pieces of information into the en-LN that are central to the characterization of this database as a lexical system.

2.2.1 Lexical entities

As shown earlier in Table 1 (section 2.1, above), there are no lexical entities corresponding to synsets in a lexical system. The nodes of such lexical networks are mainly lexical units, i.e. words taken in a well-specified meaning.

Our first task was to compile the en-LN's wordlist, i.e. the set of all its lexical units, grouped under poly- or monosemic vocables. In order to do so, we implemented the three following operations, using information from the `wn_s.pl` sense file (presented in 2.1 above).

Operation 1 We had to perform a preliminary clean-up of Prolog data, as we found a significant number (5,580) of duplicated clauses in the `s/6` predicate database.[3]

Operation 2 We then created one vocable (new entry in the en-LN wordlist) for each distinct pair:

⟨ word form, synset grammatical type ⟩.

If there were two vocables with identical form but different synset grammatical types, we added the appropriate subscript to vocable names. For instance, from the two pairs:

⟨ `'package'`, `n` ⟩ and ⟨ `'package'`, `v` ⟩,

we generated two distinct vocables: PACKAGE$_{N}$ and PACKAGE$_{V}$.

Operation 3 For each sense in the `s/6` Prolog database, we created one lexical unit and connected it to the corresponding vocable—based on the ⟨ word form, synset grammatical type ⟩ pair found in the Prolog clause for the WordNet sense.

- If only one lexical unit was attached to a given vocable, its WordNet sense number[4] was ignored—e.g., we generated the BACKGAMMON lexical unit in the corresponding monosemic vocable.
- If several lexical units were attached to a vocable, each one received the number of the corresponding WordNet sense—e.g., we generated two lexical units, GEEK **1** and GEEK **2**, in the GEEK polysemic vocable.

The process of lexical entity generation resulted in a huge fully disconnected graph (a cloud of nodes without connecting arcs) comprising 206,976 lexical units—nodes in the graph— associated to 156,584 vocables,[5] which gives a polysemy rate of around 1.322.

To conclude on the topic of the generation of lexical entities, it is important to recall that not all WordNet senses are indeed lexical units. There is a very significant quantity of phrasal entities[6] in WordNet's synsets, and only a small proportion of those phrases are actual idioms, i.e. lexical units (Osherson and Fellbaum, 2010). The automatic processing of WordNet data cannot separate true idioms from compositional phrases, and a manual post-processing of the en-LN will be necessary in order to validate the en-LN wordlist.

Important remark Our data structure allows us to specify a probability—understood as a measure of trust value—for each piece of lexicographic information entered into the en-LN (cf. properties of lexical systems, section 1.1 above). We have decided that information that is automatically generated will receive a 0.5 probably. This is true for the validity of vocables and lexical units, but also for lexical links and individual properties of lexical units that we have computed from WordNet. This strategy boils down to considering the current en-LN as being a "hypothesized lexical database."

2.2.2 Individual properties of lexical units

Five different types of individual properties have been assigned to lexical units in the en-LN: so-called WordNet "sense keys," parts of speech, syntactic features, semantic glosses and syntactic government patterns (subcategorization frames).

WordNet sense keys We found it essential to encode in the en-LN the correspondence between lexical units and WordNet senses, using WordNet

[3] We actually discovered other errors in the Prolog files (mainly, but not only duplicates) that we had to circumvent in order to avoid the generation of inconsistent data in the resulting en-LN. The list of errors can be provided on request.

[4] WordNet sense number is necessarily `1` in such cases.

[5] Cf. section 1.1 above: vocables are considered as more abstract lexical entities and are not counted as actual nodes of the lexical graph.

[6] Phrasal senses are called *collocations* in WordNet terminology. This is a different notion from that of collocation understood as semi-phraseological expression—e.g. support verb constructions such as *take a nap* (Benson, 1989).

IDs called *sense keys*. These IDs were extracted from the `wn_sk.pl` Prolog file and encoded as *WordNet source* features in the Grammatical Characteristics zone of the en-LN lexicographic articles. For instance, the lexeme INFATUATION **2** has received the value 'infatuation%1:12:02::' as WordNet source feature.

Semantic glosses In WordNet, semantic glosses are associated to synsets and not to individual senses. ⟨ Synset, gloss ⟩ pairs were extracted from the `wn_g.pl` file and the en-LN article of each member of a given synset received the same gloss attribute. Computationally, glosses are simply stored as strings of characters in the Definition lexicographic zone, more precisely in its Comments section.

Parts of speech (POS) WordNet 'synset types' have been retrieved from the `wn_s.pl` Prolog file and encoded as Part of speech features in the Grammatical Characteristics zone. The correspondence between WordNet synset type codes—SType—and en-LN's parts of speech—POS—is given in Table 2 below.

SType	POS
`v`	'verb'
`n`	'proper noun' if name starts with a capital letter, 'common noun' otherwise (of course, a very approximate rule of thumb)
`a` and `s`	'adjective'—we used only one part of speech for adjectives as we consider that WordNet's class of satellite adjectives (`s` type) pertains to WordNet internal organization rather than to the identification of grammatical behavior
r	'adverb'

Table 2: en-LN interpretation of synset types.

Syntactic features Features corresponding to information on syntactic behavior of adjectives (syntactic role and linear positioning) were retrieved from the `wn_syntax.pl` Prolog file, where they are associated to individual senses. Table 3 below describes how this information has been encoded as features in the Grammatical Characteristics zone of the en-LN.

Syntactic government patterns We retrieved associations between synsets and WordNet's syntactic frame codes in the `wn_syntax.pl` Prolog file. The definitions of syntactic frames themselves where taken from WordNet's documentation (`wninput.5.pdf` file). Then, for each sense member of a given verbal synset, we entered the associated frame description into the Government Pattern zone (Comments section) of the corresponding lexical unit.

WordNet feature	en-LN gram. charac.
`a`	'attributive'
`p`	'predicative'
`ip`	'postposed'

Table 3: en-LN interpretation of syntactic features.

Now that the generation of lexical properties has been explained, let us move to the crucial topic of weaving lexical function relations, that give the en-LN its connected graph structure.

2.2.3 Lexical function relations

In total, 12 Meaning-Text lexical functions (Mel'čuk, 1996) have been used to encode lexical relations extracted from WordNet. They can be grouped into three different classes:

- 7 standard lexical functions: **Syn**$_\cap$, **Anti**$_\cap$, **Gener**, **Mult**, **Sing**, **A**$_2$ and **Caus**;
- 4 that have been "standardized" (Polguère, 2007) in the context of previous projects: **Cf**, **Hypo**, **Holo** and **Mero**;
- 1 non-standard: **Unspecified derivative**.

Table 4 below gives statistics on the distribution of lexical links pulled in the en-LN for each of those twelve lexical functions.

Number of links	Lexical function
315,984	**Syn**$_\cap$
145,880	**Gener**
145,880	**Hypo**
89,107	**Unspecified derivative**
59,981	**Mult**
59,981	**Sing**
50,746	**Cf**
35,663	**Mero**
33,684	**Holo**
7,979	**Anti**$_\cap$
1,250	**Caus**
73	**A**$_2$

Table 4: Lexical function links in the en-LN.

For lack of space, we focus below on the generation of only three lexical links, that are the most significant statistically: **Syn**$_\cap$, **Gener** and **Hypo**.

Extraction of Syn$_\cap$ relations The **Syn$_\cap$** lexical function stands for 'intersecting synonym'; its extraction from WordNet was done as follows:

If sense 's' belongs to synset S
  And $L_{'s'}$ is the lexicalization of 's' in the en-LN
Then the lexicalizations of all other senses belonging to S are targets of **Syn$_\cap$** links originating from $L_{'s'}$
  And the same principle applies recursively to all other senses of S.

This principle entails the "saturation" of all possible **Syn$_\cap$** links among all elements of all synsets in WordNets. And each application of this principle on a synset generates a saturated subgraph.

Figure 1 below shows the **Syn$_\cap$** saturated subgraph generated from synset (1).

(1) `{puppy love,calf love, crush#3-n,infatuation#2}`

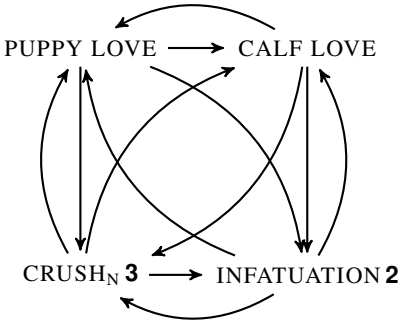


Figure 1: **Syn$_\cap$** saturated subgraph for synset (1)

We made the hypothesis that most synset members in WordNet are connected by the intersecting approximate synonymy relation **Syn$_\cap$**, rather than by exact synonymy **Syn**. We expect that our strategy will entail less manual corrections when the en-LN will be used for lexicographic purposes (section 4). Synset (1) is a clear illustration of the potential relevance of our hypothesis, as senses in (1) are indeed not **exact** synonyms.

Extraction of Gener $\sim$ Hypo relations **Gener** is a Meaning-Text standard lexical function and stands for 'generic term'. Though it is close to WordNet's hypernymy, it was not possible to systematically extract **Gener** relations from WordNet's hierarchical organization, for two reasons.

Firstly, the hypernymy relation holds between synsets, whereas **Gener** connects lexical units.

Secondly, **Gener** is more specific than WordNet's hypernymy. It holds between two lexical units in only two specific cases, illustrated below.

A. FRUIT is a **Gener** of BANANA because it is possible to say (2).

(2) ***bananas**, apples, oranges and other **fruits***

B. SUBSTANCE is a **Gener** of GAS because (3a) can be paraphrased as (3b) using GASEOUS, the adjectival counterpart of GAS.

(3) a. *gas*
  b. *gaseous substance*

Gener is thus before all a lexical, rather than conceptual or denotational relation. In the context of our lexical projects, **Gener** is paired with a symmetrical lexical function called **Hypo**, for 'hyponym'. Notice that this latter lexical function does not belong to the original set of Meaning-Text standard lexical functions.

Gener $\sim$ **Hypo** relations were mainly extracted from hypernym relations between synsets (`wn_hyp.pl` file) as follows:

If synset S_1 is a hypernym of synset S_2
  And S_1 is the hypernym of more than 15 synsets
Then all senses of S_1 are targets of **Gener** links originating from of all senses of S_2
  And all senses of S_2 are targets of **Hypo** links originating from of all senses of S_1.

This ensures that there is no explosion of the number of invalid **Gener** and **Hypo** links. After doing some testing with different thresholds, we reached the conclusion that a synset that happened to be the hypernym of more than 15 other synsets had the greatest chance to contain true generic terms (in our sense).[7]

With this strategy, we caught in our nets "only" 111,032 **Gener** relations and the same number of **Hypo** relations. Without the ">15" constraint, numbers would have been much higher and en-LN data much less accurate.

[7]For instance, the WordNet sense `car#1` belongs to a synset that is the hypernym of 31 other synsets. It has thus been identified as good candidate for generic term; as a result, the corresponding lexical unit is the **Gener** of 66 other lexical units in the en-LN. In contrast, `desk#1` belongs to a (singleton) synset that is the hypernym of only 3 other synsets; no **Gener** link has been pulled from the DESK lexical unit.

A smaller set of **Gener** and **Hypo** relations (69,696) has been extracted from instance→type relations between nominal synsets (`wn_ins.pl` file) based on the following principle:

If synset S_1 is a type of synset S_2 (that is its instance)
Then all senses of S_1 are **Gener** of all senses of S_2
And all senses of S_2 are **Hypo** of all senses of S_1.

To conclude this section, notice that the strategies applied for extracting **Syn**$_\cap$, **Gener** and **Hypo** relations—which implies symetric relations—are chiefly responsible for the very high proportion of "mutual arcs" in the graph—see section 3.1 below, that presents a topological comparison of the fr- and en-LNs.

2.3 Accessing the resulting en-LN

Once the interpretation of WordNet information into a lexical system structure is performed, we are able to access and navigate through the en-LN with the *Dicet* lexicographic editor, designed for lexicographic work on the fr-LN. In actual fact, we are now able to edit and transform the newly generated en-LF using our lexicographic approach.

In order help the reader have a more concrete grasp of how different the English lexical system is from WordNet, we provide in Figure 2 below a lexicographic view of the first sense of the GEEK vocable. For a presentation of the specificity of lexicographic editing by means of the Dicet editor, see (Gader et al., 2012).

3 Graph properties

The aim of this section is three-fold:

1. to determine to what extent the fr-LN and the en-LN differ in terms of mathematical organization;

2. to formally characterize the structure of both networks as so-called *Hierarchical Small World Networks*, which is the expected graph type for lexical systems;

3. to use the full-scale nature of the en-LN, inherited from WordNet, to anticipate future formal properties of our "adolescent" fr-LN.

Section 3.1 presents a formal characterization of the fr-/en-LNs from the viewpoint of their graph structure. Topological analyses of both graphs allow us to mathematically compare their formal structure. Section 3.2 summarizes this comparison in layman terms and draws conclusions from formal differences that have emerged.

3.1 Formal topological analysis

Structural properties of our lexical systems were studied using *pedigree.py*, a Python script developed by Emmanuel Navarro (Gaillard et al., 2011). This script performs topological analyses—called *graph pedigrees*—, that allow for rigorous graph characterization and comparison. More specifically, we seek to determine if the fr-/en-LNs are *Hierarchical Small World Networks* (Watts and Strogatz, 1998; Newman, 2003; Gaume, 2004).

Hierarchical Small World Networks, hereafter HSWN, exhibit four properties:

1. low density, i.e. small number of arcs compared to the number of nodes;

2. high global clustering coefficient, i.e. high number of connected neighbor nodes;

3. distribution of degrees (probability distribution of number of arcs associated to a node) that follows a power law;

4. low average path length, i.e. small average minimal number of arcs between two nodes for each possible pairs.

Table 5 below shows the pedigree of our two lexical systems.

The current fr-LN comprises 9.9 times less nodes (*n*) than the English network—straight from the oven—, for 27.1 times less arcs (*m*). To determine if these densities are low, we compare *m* to n^2 and $nlog(n)$. n^2 is the maximum amount of arcs that can exist for a given number of nodes and a unique relation type.[8] It is about 432×10^6 for the fr-LN and 43×10^9 for the en-LN. From this point of view, their densities are low. $nlog(n)$ represents the order of magnitude of HSWN's density (Gaume, 2004). It is about 89,773 for the fr-LN, which is twice the current amount. For the en-LN, it is about 1,100,267, which is close to what we measured.

[8] In our case, there are 662 different relations involved in the fr-LN and 12 in the en-LN. The maximum amount of arcs increases proportionaly.



Figure 2: (Partial) lexicographic view of GEEK **1**.

	fr-LN	en-LN
n	20,791	206,976
m	34,920	946,208
<k>	3.3406	5.9029
Directed	true	true
Mutuals	15,576	942,795
Loops	46	1
Single	3,540	19,756
Multiples	432	124
ncc	14,295	34,342
C	0.1058	0.1031
Out degree distribution		
a	-2.0243	-1.8479
r^2	0.9572	0.8453
LCC		
n_lcc	1,788	144,294
m_lcc	5,973	851,748
C_lcc	0.2816	0.0980
L_lcc	13.0861	10.1479

Table 5: Pedigree of the fr-/en-LNs.

The fr-LN is a work in progress. It includes a high proportion of single nodes (17%), which implies a high number of strongly connected components (*ncc*)[9] and explains its small largest connected component (*LCC*). The network increases in arcs more quickly than in nodes, due to the organization of the lexicographic work. In addition, the amount of single nodes decreases. Table 6 shows this evolution over five months.

	June 2013	Oct. 2013	Evolution
Arcs	25,932	34,920	+35%
Nodes	18,057	20,791	+15%
Single	3,614	3,540	-2%

Table 6: Evolution of the fr-LN.

The en-LN has not undergone any evolution yet. However, it will be manually transformed in the future. Some arcs will be added and its proportion of single nodes (2%) will decrease. Some arcs will also be modified or deleted. For example, its unique loop is a WordNet error and will be eliminated from the en-LN (as it should be from WordNet).[10]

The French network, which contains a wide variety of links (662), has 44.6% of mutual arcs—i.e. arcs $a \rightarrow b$ for which a reverse arc $b \rightarrow a$ exists. They are many more in the en-LN (99.5%), due to the nature of the 12 lexical links encoded.

To estimate how nodes and arcs are locally or-

[9]Single nodes are considered to be strongly connected components.

[10]This is a `derivationally related form` arc connecting `unicyle`$_N$ to itself, that is present in both the on-line WordNet 3.1 and the Prolog database we used.

ganized in networks, one needs to examine their global clustering coefficients (*C*).

Nodes	Arcs	*C* Lexical	*C* Random
20,791	34,920	0.1058	0.00016
206,976	946,208	0.1031	0.00004

Table 7: Global clustering coefficients.

For our networks, *C* may seem small. However, as Table 7 shows, they are higher than for similar classical random graphs (Newman, 2003). In other words, these networks hold sets of highly connected lexical units.

To assess the structure of these lexical agregates in the network, we need to examine the distribution of degrees and the average path length (*L*).

As our networks are oriented, we focus only on their out distribution of degrees. Both networks follow a power law with a good correlation coefficient (r^2). In Table 5 above, a stands for the coefficient of the best fitting power law of these distributions. Such a distribution is in the same range as for typical HSWNs. This means that a few number of lexical units are highly connected to a slightly higher number of other lexical units, themselves connected to a slightly higher number of other lexical units. To put it differently, our networks contain lexical hubs and are hierarchically structured.

Bollobás and Riordan (2004) have shown that the *L* of HSWNs does not exceed log*n*/ log log*n*. Such a value for *L* means that it is possible to move rapidly from a node of the network to another.

As our networks have more than one component, measure of their *L* is problematic (Newman, 2003). In fact, it is difficult to define a path length between two non-connected nodes. A possible alternative can be to consider the *L* of *LCC* (*L_lcc*).

	n_lcc	log*n*/ log log*n*	L_lcc
fr-LN	1,788	6.350	13.0861
en-LN	144,294	7.240	10.1479

Table 8: Average path lengths.

Table 8 shows that *L_lcc* of our networks are higher than expected. For the French network, *LCC* is very small and probably not representative of the whole network. For the English network, the problem is different. The original structure of WordNet keeps separate the synsets of the four major parts of speech (with marginal transversal connections). It is reasonable to believe that some structuring lexical relations between agregates belonging to different parts of speech are missing.

To conclude, our fr- and en-LNs seem to be both structured as HSWNs, but have an average path length higher than they should have.

3.2 In layman terms

As indicated in section 3.1 above, the en-LN is substantially larger than the current fr-LN. In contrast, lexical relations are more diverse in the latter.

Despite such differences, the global structure of both networks appear to be similar. They seem to represent the same type of lexical organization. In both cases, senses are organized in highly connected subsets and some lexical units assume a pivotal role. These characteristics appear consistent with a semantic field structure. Further investigation is required to learn more about highly connected components, like the nature of links and lexical units involved. Some new similarities might then emerge.

The question of a fast and easy access between lexical agregates remains. More detailed observation would be required to determine why such an access is not possible in the LNs. The en-LN is made up mostly of paradigmatic links. Maybe this characteristic is the cause of our trouble. But this explanation does not hold in the case of the fr-LN. In a study of WordNet, where the different parts of speech are structured together, Sigman and Cecchi (2002) propose to introduce polysemous links to improve access between lexical units. We are currently implementing the weaving of such links in the fr-LN and more will be known soon about their incidence on the global structuring of LNs.

4 So what gives?

In form of conclusion, we summarize the practical interests of performing the WordNet → en-LN compilation. Two main points should highlighted.

First and foremost, as stated in section 2.3, we are able to wade through the en-LN and edit it using our "graph weaver:" the Dicet lexicographic editor. This is essential to us as we believe that the lexical system model ought to be extensively tested as an alternative to more ontological approaches to lexical knowledge structuring, such as WordNets.

Second, thanks to WordNet, we now have at our disposal a lexical unit-based access to the English

lexicon that can be used to explore structural behavior of full-scale lexical systems, in anticipation of the fr-LN reaching lexicographic maturity.

Acknowledgments

Work on the fr-LN is supported by a grant from the Agence de Mobilisation Économique de Lorraine (AMEL) and Fonds Européen de Développement Régional (FEDER). We are grateful to Veronika Lux-Pogodalla and two GWC2014 reviewers for their comments on the first version of this paper. Emmanuel Navaro has been very helpful in providing feedback on the use of Pedigree.py.

References

B. T. Sue Atkins. 1996. Bilingual Dictionaries: Past, Present and Future. *Proc. of Euralex'96*, Martin Gellerstam, Jerker Järborg, Sven-Göran Malmgren, Kerstin Norén, Lena Rogström, Catalina Röjder Papmehl (Eds.), Gothenburg University, Department of Swedish, 2006, 515–590.

Morton Benson. 1989. The Structure of the Collocational Dictionary. *International Journal of Lexicography*, 2(1):1–14.

Béla Bollobás and Oliver Riordan. 2004. The Diameter of a Scale-Free Random Graph. *Combinatorica*, 24(1):5–34.

Nabil Gader, Veronika Lux-Pogodalla and Alain Polguère. 2012. Hand-Crafting a Lexical Network With a Knowledge-Based Graph Editor. *Proc. of the Third Workshop on Cognitive Aspects of the Lexicon (CogALex III)*, The COLING 2012 Organizing Committee, Mumbai, 109–125.

Benoit Gaillard, Bruno Gaume, and Emmanuel Navarro. 2011. Invariants and Variability of Synonymy Networks: Self Mediated Agreement by Confluence. *Proc. of TextGraphs-6: Graph-based Methods for NLP*, ACL, Portland, 15–23.

Bruno Gaume. 2004. Balades Aléatoires dans les petits mondes lexicaux. *I3 Information Interaction Intelligence*, 4(2):39–96.

Alvaro Graves and Claudio Gutierrez. 2005. Data Representations for WordNet: A Case for RDF. *Proc. of Global WordNet Conference 2006*, Petr Sojka, Key-Sun Choi, Christiane Fellbaum, Piek Vossen (Eds.), Jeju Island, 2006, 165–169.

Huang Xiao and Zhou Chang-le. 2007. An OWL-based WordNet lexical ontology. *Journal of Zhejiang University SCIENCE A*, 8(6):864–870.

Jaap Kamps. 2002. Visualizing WordNet structure. *Proc. of Global WordNet Conference 2002*, Central Institute of Indian Languages , Mysore, 2002, 182–186.

Veronika Lux-Pogodalla and Alain Polguère. 2011. Construction of a French Lexical Network: Methodological Issues. *Proc. of the First International Workshop on Lexical Resources, WoLeR 2011. An ESSLLI 2011 Workshop*, Ljubljana, 2011, 54–61.

Igor Mel'čuk. 1996. Lexical Functions: A Tool for the Description of Lexical Relations in the Lexicon. *Lexical Functions in Lexicography and Natural Language Processing*, Leo Wanner (Ed.), Language Companion Series 31, John Benjamins, Amsterdam/Philadelphia, 37–102.

Igor Mel'čuk. 2006. Explanatory Combinatorial Dictionary. *Open Problems in Linguistics and Lexicography*, Giandomenico Sica (Ed.), Polimetrica, Monza, 225–355.

George A. Miller, Claudia Leacock, Randee Tengi and Ross T. Bunker. 1993. A semantic concordance. *Proc. of the ARPA Human Language Technology Workshop*, Princeton, 303–308.

Mark E.J. Newman. 2003. The structure and function of complex networks. *SIAM REVIEW*, 45:167–256.

Anne Osherson and Christiane Fellbaum. 2010. The Representation of Idioms in WordNet. *Proc. of Global WordNet Conference 2002*, CFILT, IIT Bombay, Mumbai, 2010.

Alain Polguère. 2007. Lexical function standardness. *Selected Lexical and Grammatical Issues in the Meaning-Text Theory. In Honour of Igor Mel'čuk*, Leo Wanner (Ed.), Language Companion Series 84, John Benjamins, Amsterdam/Philadelphia, 43–95.

Alain Polguère. 2009. Lexical systems: graph models of natural language lexicons. *Language Resources and Evaluation*, 43(1):41–55.

Mariano Sigman and Guillermo A. Cecchi. 2002. Global organization of the Wordnet lexicon. *Proc. Natl. Acad. Sci.*, 99(3):1742–1747.

Dennis Spohr. 2012. *Towards a Multifunctional Lexical Resource. Design and Implementation of a Graph-based Lexicon Model*, De Gruyter, Berlin/Boston.

Duncan J. Watts and Steven H. Strogatz. 1998. Collective dynamics of 'small-world' networks. *Nature*, 393:440–442.