Dictionaries as Networks: Identifying the graph structure of Ogden's Basic English

Camilo Garrido and **Claudio Gutierrez**
Center for Semantic Web Research & Department of Computer Science,
Universidad de Chile
{cgarrido, cgutierr}@dcc.uchile.cl

Abstract

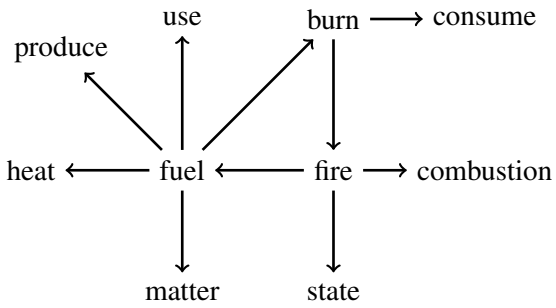
We study the network structure underlying dictionaries. We systematize the properties of such networks and show their relevance for linguistics. As case of study, we apply this technique to identify the graph structure of Ogden's Basic English. We show that it constitutes a strong core of the English language network and that classic centrality measures fail to capture this set of words.

1 Introduction

Dictionaries are rich sources to investigate the semantic structure of natural language. The purpose of dictionaries, writes Wilks et al. (1993), "is to provide definitions of senses of words and, in so doing, supply knowledge about not just language, but the world." The definition of a word involves recursively new words, and thus, new senses and meanings. In this way, a dictionary can naturally be viewed as a network where each word w is related to the set of words $w_1, \ldots, w_n$ that define it: for each word w so defined, consider the relationship $w \to w_i$ for each $i = 1, \ldots, n$, and proceed recursively with all the entries of the dictionary. The idea is not new and was already proposed by K. C. Litkowski (1978).

Clearly this basic idea must be refined. There are words that are in inflected form (e.g. verbs); that are the same but have different meanings (e.g. singer: the machine, the musician, etc.); that are in plural or singular; that are the same adjective with different gender, etc. In order to make the network conceptually coherent, one should define classes of words; for example, all the inflected forms of the verb "play" define one class whose representative is the word "play". There are several other simple processing decisions to be made. This naive version can be further refined by incorporating more elaborated linguistic features, like labeling the edges according to parts of speech to which they point, *e.g.* nouns, verbs, adjectives, adverbs, etc., or giving different nodes and weights to different meanings of a word, and so on. The surprising fact is that even using a naive approach, the network obtained gives highly relevant and interesting information about the language. See Figure 1 for a small example.

Although the idea of using mathematical and computational tools to capture the semantics information in dictionaries has been broadly explored (Amsler, 1980; Calzolari, 1984; Wilks et al., 1988), the idea



FIRE: Fuel in a state of combustion.

FUEL: Any matter used to produce heat by burning.

BURN: To consume with fire.

Figure 1: Entries in the dictionary for the words *burn*, *fire*, and *fuel*, and their corresponding subgraph built from them.

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 3565–3576, Osaka, Japan, December 11-17 2016.

of exploiting the inherent network structure of dictionaries has not been pursued systematically. As we mentioned, the idea was proposed several decades ago (Litkowski, 1978), but only recently, with the explosion of network studies and hardware availability, there have been some works in this direction (we discuss them in the Related Work section).

The aim of this paper is to provide evidence of the fruitfulness of studying dictionaries as networks. We show that dictionaries (in general) have similar structure from the point of view of networks, and as expected, their structural properties differ from networks obtained from other areas. We claim that dictionary networks have particular properties (strong connectivity, resilience, component analysis, etc.) that shed light on the structure of the languages and deserve to be studied in depth. We found out that classical tools for studying and analyzing networks –particularly those popularized by Social Network Analysis (Wasserman and Faust, 1994)– like centrality measures, subgroups, affiliation, etc. are not always meaningful in this new realm, and that their successful application to this linguistic setting requires to be reworked. For example, it is not evident that they help capturing notions such as "most relevant" or "non-relevant" words in a dictionary that are important, for example, for building small dictionaries, basic sets of words for beginners, etc.

In order to test these and other ideas in practice we chose as a study case *Ogden's Basic English*, a set of 850 words selected by the linguist C. K. Ogden to serve as a basic language (Ogden, 1930) . In order to study it from a network point of view, we built a network out of an English dictionary. We chose the *Online Plain Text English* (OPTED) because it is reliable, contains 94.5% of Ogden's words, and is open data, thus, allowing anyone to replicate our experiments. We then applied different graph-theoretical notions and techniques to this network, aiming to capture Ogden's set of words.

Our study shows, using *only* graph-theoretical tools, that Ogden's set of words is part of a strongly connected core of the English dictionary, a subset of words that directly use each other in their definitions. We then show that it is not formed by the "most central" words (according to classic ranking measures), but by a combination of high ranked words plus others that play the role of "covering" the rest of the network, that is, being "close" to most words in the dictionary.

The main contribution of our work is to add evidence about the value of using dictionary networks to study linguistic properties of languages.

2 Related Work

The community agrees that dictionaries are a source of lexical knowledge (Calzolari, 1984; Dolan et al., 1993). This knowledge can be used for the development of NLP techniques, establishing usage relation between words or hierarchy relations like hypernyms or "part_of". They also can be used for the creation of pocket dictionaries and many other applications.

One of the first uses of dictionaries was to develop Machine-Readable Dictionaries (Zingarelli, 1970) (MRDs). With MRDs and the concept of lexical databases, the importance of the information and knowledge that dictionaries contain began to gain attention. Amsler(1980) presents some efforts to exploit dictionaries and extract information for applications in computational linguistics. He investigates the possibility of building of taxonomies based on the structure of the definition of words. He also offers some insight on the frequency of the vocabulary and semantic ambiguity. Calzolari (1984) detects some patterns among lexical entries: hyponyms and restriction relations. Later, Calzolari et al. (1988) focused their efforts on extracting semantic information from dictionaries. They state "The dictionary is now considered as a primary source not only of lexical knowledge but also of basic general knowledge". They parse the entries and try to organize the knowledge with functions like Hypernym, Relation, Qualifier, etc. Wilks et al.(1988) discuss the importance of dictionaries for NLP tasks, in particular, the value of transforming machine readable dictionaries (MRDs) into machine tractable dictionaries (MTDs). They show three approaches for this: Obtaining and using co-occurrence statistics, producing a lexicon and extracting a Key Defining vocabulary.

At the same time MRDs began to get attention, so did modeling the dictionaries as networks. K. C. Litkowski (1978) was one of the first to state the importance of studying and exploiting dictionary networks, both as sources of material for natural language and to unravel the complexities of meaning.

He presented three models to represent a dictionary. The first model uses nodes to represent words and edges to represent the relation w_a *"is used to define"* w_b. The second model extends the first one, adding relations between words and senses. In the third and final model, nodes represent concepts and edges represent different relations between them (senses, part of sentence, etc). Definitions are broken down into subphrases. For example, "Broadcast: the act of spreading abroad" may be broken into "the act", "of", "spreading abroad", where these subphrases may be broken into smaller pieces.

After the seminal this work of Litkowski (1978), several researchers have used dictionary networks to study or extract information about the language.

Dolan el al.(1993) developed an automatic strategy to exploit dictionaries to construct a source of lexical an common sense information based on *hypernyms*, *locations*, *part_of*, and other relations. Although a network can be formed with those relations, the methods of the system to extract such relations between words is not clear. Picard et al.(2009) address the following question: "How many words do you need to know in order to be able to learn all the rest from definitions?" They approach this question representing dictionaries as networks. Levary et al.(2012) show that if we follow the definition of a word over and over, one typically finds that definitions loop back upon themselves. They also show that the loop is an essential element of the growth process of networks. They showed that words within these loops tend to be introduced into the English language at similar times. And, the evolution of these networks follow the "rich-get-richer" growth. Mihalcea(2004a) used networks derived from WordNet to test disambiguation.

There are other forms of building networks of words and using graph ideas in word analysis, *e.g.* co-occurrence of words in certain windows (bigrams, etc.) (Dorogovtsev et al., 2001; Mihalcea, 2004b).

Finally, there is a line of research that investigates the relationship between semantic networks and graph measures. Abbott et al. (2012) compare the functioning of the human mind when searching for memories with a random walk in a semantic network. They conclude results that can help clarify the possible mechanisms that could account for PageRank predicting the prominence of words in semantic memory. Yeh et al. (2009) used random walks to determine the semantic relatedness between two elements. They conclude that random walks performed with personalized PageRank is a feasible and potentially fruitful means of computing semantic relatedness for words and texts. Hughes et al. (2007) introduce a new measure of lexical relatedness based on the divergence of the stationary distributions computed from random walks over graphs extracted from WordNet. Steyvers and Tenenbaum (2005) conjecture about semantic networks stating that "these structures reflect the mechanisms by which semantic networks grow." All of these works served as sources of inspiration for our research.

3 Dictionaries as Networks

> "Ordinary dictionaries have not been given their due, either as sources of material for natural language understanding systems or as corpora that can be used to unravel the complexities of meaning and how it is represented. If either of these goals are ever to be achieved, I believe that investigators must develop methods for extracting the semantic content of dictionaries (or at least for transforming it into a more useful form). [...] A suitable framework appears to be provided by the theory of labeled directed graphs (digraphs)." (Litkowski, 1978).

If words are viewed as basic building blocks of more complex meaning structures, the network of their relationships can be considered as the skeleton that holds them together. Dictionaries are one of the primary sources to obtain such skeletons of meaning.

A network (or graph: both used synonymously) is defined by the nature of its nodes and the of relationships that connect its nodes. A dictionary viewed as a network on the lines we explained above, gives rise to different types of nodes and edges. Nodes have types of n.; n.pl.; a.; v.; v.t.;v.i.; adv.; etc. Edges also can be of different types, according the role or the place of the word in the definition. For example, consider the following three entries of the word "act", each with a different type:

Act (n.) A formal solemn writing, expressing that something has been done.

Act (v. i.) To exert power; to produce an effect; as, the stomach acts upon food.

Act (v. t.) To perform; to execute; to do.

Also, the words occurring in these definitions play different grammatical roles, can occur more than once, etc. All of these features should be included in a faithful network of a dictionary, ideally one from which one can reconstruct the dictionary (see some insights in (Litkowski, 1978)).

On the other extreme, one can build a simple (naive) network without any typing on nodes and edges, that is, just words pointing to words represented in some standard form (e.g. lemmatized). There is a compromise between these two extreme approaches: as usual, the simpler the better (for network analysis, more tools available; for comparison with other fields, particular features do not help) at the cost of losing some subtle linguistic properties. In what follows we develop the simplest possible approach, with the idea of showing the potentialities of the method, and hoping to keep enhancing this baseline with further linguistic annotations.

3.1 Building the Basic Network

For this work we implemented the following procedure to build the networks:

1. Model or Design. Consider all types of words as a single type: forget if they were nouns, verbs, adverbs, etc. Merge the entries that correspond to the same word into one definition, e.g. *Singer (n.) A machine for sewing cloth.* and *Singer (n.) One who, or that which, singes*. Forget the role and place of occurrence of a word, as well as its number of occurrences, inside a sentence (i.e. transform the defining text of a word in a set of words).

2. Clean. Remove the terms that are inflected forms, e.g *singing: from Sing*. Remove prepositions and articles. They appear too often in any text, so they would add noise to the graph. Lemmatize each word occurring in the definitions (transform nouns into singular; verbs into the infinitive; adjectives into their male singular form). Remove any word that does not appear in the dictionary, *e.g.* prefixes and suffixes like *Ex-* and *-able*.

3. Mathematical model of the dictionary. Build the graph over the previous data. At this point, the dictionary D has become a universe of words W and a set of pairs $(w, \mathrm{def}(w))$, where $w \in W$ is an entry in D and $\mathrm{def}(w) \subseteq W$ is the set of words occurring in the definition of w.

4. Build the Network. From the data in (3), construct a directed graph $G = (V, E)$, where the nodes are $V = \{w|(w, S) \in D\}$ and the edges $E = \{(w, w')|(w, S) \in D \text{ and } w' \in \mathrm{def}(w)\}$. For example, from the entry "*Eaglet (n.) A young eagle, or a diminutive eagle.*" we get the edges *(Eaglet,young), (Eaglet, eagle)* and *(Eaglet, diminutive)*.

The OPTED dictionary. We applied the above methodology to the *The Online Plain Text English Dictionary*[1] (OPTED) and the *Diccionario de la Real Academia Española*[2] (DRAE, Royal Spanish Academy Dictionary). We chose OPTED because is a public and free-access dictionary, based on Webster's Unabridged Dictionary, and an important and recognized dictionary. On the other hand, we chose DRAE because it is the most authoritative dictionary of the Spanish language. The first edition of the DRAE was published in 1780, and the current, twenty-third edition, was published in 2014.

The OPTED network has 95,095 nodes and 979,523 edges. The nodes are composed of 58,750 nouns and 12,261 verbs. The remaining 24,084 nodes correspond to adjectives and adverbs. The RAE network has 89,767 nodes and 1,152,301 edges. The nodes are composed of 54,767 nouns and 12,046 verbs. The remaining 22,954 nodes correspond to adjectives and adverbs.

To make a good description of the dictionary network, we analyzed its different features. First, we present a set of basic properties and compare them to other kinds of networks (social, information, etc.). Second, we show a component analysis. And third, we present other characteristics obtained with graph machinery.

3.2 Dictionary Networks compared to other networks

We do the comparison with other types of networks based on classic parameters used to describe networks (Newman, 2003). Table 1 shows basic parameters for three different dictionary networks, and another three networks built by humans.[3]

[1] `http://www.mso.anu.edu.au/~ralph/OPTED/`

[2] `http://www.rae.es/`

[3] We use *igraph* `http://igraph.org/` as network analysis package and Stanford CoreNLP(2014) for lemmatizing the words in the dictionary.

	n	m	z	l	α	$c1$	$c2$	r
OPTED	95 095	979 523	20.601	4.64	2.63 / 3.13	0.009	0.217	-0.0081
DRAE network	89 767	1 152 301	25.673	3.26	2.39 / 2.74	0.044	0.201	-0.0092
WordNet	84 967	1 134 957	26.715	2.99	2.84 / 2.99	0.029	0.203	-0.0157
ca-HepPh	11 204	235 268	41.997	4.67	1.76 / 1.76	0.659	0.690	0.630
cit-HepTh	27 400	352 542	25.733	4.28	2.72 /4.14	0.120	0.329	0.002
p2p-Gnutella04	10 876	39 994	7.355	4.64	- /3.55	0.005	0.008	-0.0083

Table 1: Basic measures for networks. OPTED is an English dictionary network. DRAE is a Spanish dictionary network. WordNet is a dictionary network built from WordNet. ca-HepPh is a collaboration network from the e-print arXiv. cit-HepTh is the Citation graph from the e-print arXiv. p2p-Gnutella04 is a sequence of snapshots of the Gnutella peer-to-peer file sharing network. Details for the last three networks are in (Leskovec, 2014).

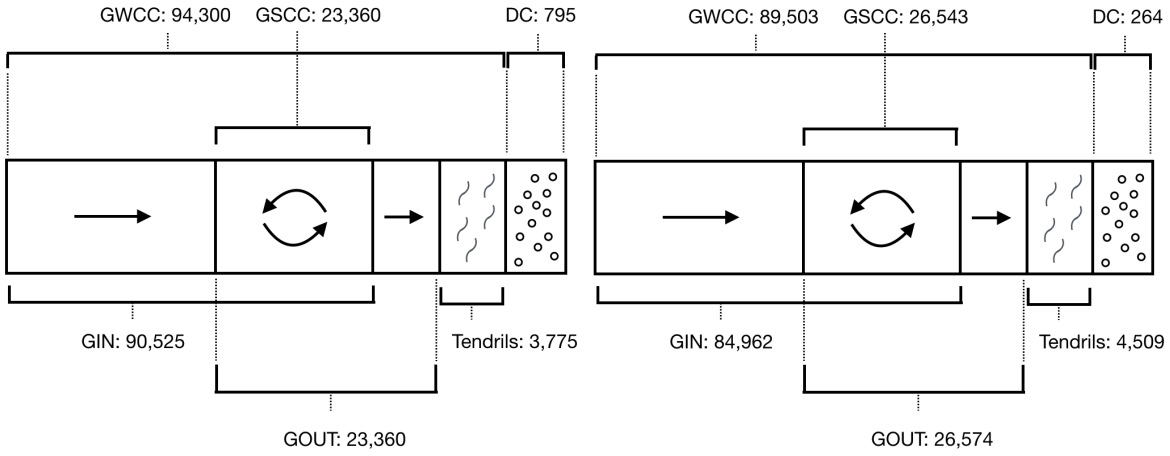
The number of nodes n tells the "size" of the network; m is the number of edges that allows for an estimation of its density, the fraction $0 \leq \frac{m}{n(n-1)} \leq 1$. Our three dictionary networks have m about 10 times n. The mean degree z gives an idea of the distribution of the edges on vertices. The mean vertex-vertex distance l tells how related/close the pairs of nodes are. The numbers in the table indicate that dictionaries have the small-world property. The parameter α refers to the exponent of the degree distribution function ($p_k \sim k^{-\alpha}$, where p_k is the fraction of the nodes that have degree k, in/out-degree) when the network (as in this case) follows this type of distribution ("power law"). It means that there are few nodes with a high degree and a large tail of low-degree nodes. The clustering coefficients $c1(= \frac{6 \times \text{number of triangles}}{\text{number of paths of length 2}})$ and $c2$ $(= \frac{1}{n}\sum_i c_i$ where $c_i = \frac{\text{number of triangles connected to vertex } i}{\text{number of triples centered on vertex } i})$ refer to the degree to which vertices tend to cluster together. In terms of network topology, the clustering coefficient refers to the presence of triangles in the network, being $c1$ a global coefficient and $c2$ a local one. In the language of social networks, the friend of your friend is likely to also be your friend. In our setting, two words having a common (non frequent) word in their definitions are likely to be related. The r coefficient indicates whether the high-degree vertices in the network associate (have links) preferentially with other high-degree vertices or not. $r = 1$ means high connectivity among them; $r = -1$ means low connectivity.

It is interesting to observe that the three dictionaries have similar parameters (as compared to other types of networks), and their properties are similar to semantic networks. Steyvers and Tenenbaum (2005) observed for the latter: "they have a small-world structure, characterized by sparse connectivity, short average path lengths between words, and strong local clustering."

Another measure is network resilience, which correlates with high connectivity. The standard measure is vertex attack tolerance VAT (Matta et al., 2014), *i.e.* behaviour of the network after removal of some nodes, defined as $\min_{S \subset V}\{\frac{|S|}{|V-S-C|+1}\}$, where C is the largest connected component in $V - S$. We determine that VAT is 0.245 for OPTED and 0.3 for DRAE. Comparing to other scale-free networks (Matta et al., 2014) (HOTNet 0.06, big barbell 0.08, star 0.11, C3 0.15, barbell 0.2, PLOD 0.25, wheel 1.0), dictionary networks are placed among the most resilient, meaning that removing some words will cause little disruption, since with high probability there will be other good relations to supply the loss.

3.3 Component Analysis

Components are classic features when describing the topology of networks. The graph is divided in two main parts: the *Giant Weakly Connected Component* (GWCC), the biggest weakly connected component present in the graph (v is connected to w if there is a undirected path from v to w), and the rest, the *Disconnected Components* (DC), that consist of separate small connected components. GWCC consists of three parts: the *Giant Strongly Connected Component* (GSCC) (strongly connected means that for each pair of nodes v, w, there is a directed path from v to w and vice versa), usually the most relevant part of the network; the *Giant in-component* (GIN), the set of nodes that have paths to GSCC (in our setting, words that in their definitions recursively use words in GSCC and are not used to define those in GSCC); and the *Giant out-component* (GOUT), the set of words that are used to define those in GSCC. Finally, the *Tendrils* are nodes which have no access to GSCC and are not reachable from it.



(a) OPTED (English, 95,095 words)

(b) DRAE (Spanish, 89,767 words)

Figure 2: Component Analysis showing similar structures for English and Spanish dictionary networks. The core part of the network (GSCC) is composed of words that are entangled –recursively use themselves in their definitions–, and amounts to approx. 25-30% of all entries in the dictionary.

3.4 Other characteristics obtained with graph machinery

One of the most basic measures to study words in text is consider their frequency of occurrence. The dictionary as network allows the use of other measures, in particular, classical centrality measures in the literature: Degree, PageRank, betweenness centrality, and closeness centrality. As shown in Figure 3, each of them captures different features as they have little correlation. To give a taste of the results, we list in Table 2 "top" words for different measures previously mentioned.

Another productive topic of application is the search for similarities among words. To illustrate it we show that big (bidirectional) cliques, which are rare in a dictionary, are formed by words with similar meanings. In OPTED there is no K_6, seven K_5 (shown in Figure 4), 174 K_4 and 2,641 K_3. In DRAE no K_5, four K_4 and 243 K_3.

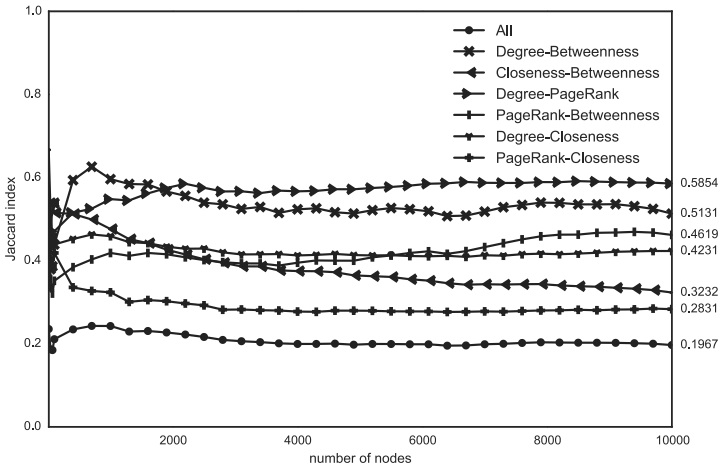


Figure 3: Common words of top rankings under different centralities, measured by Jaccard index ($\frac{|S_1 \cap S_2|}{|S_1| \cup |S_2|}$) for different number of nodes (0 to 10,000). For example the top ranked words for Degree and PageRank have 58.54% of their universe in common. All together they have 19.67% in common.

3.5 Core/Periphery Structure

Another feature that would help us understand the structure of dictionaries is the core/periphery characterization. This concept refers to the categorization of the nodes of the network. The nodes corresponding

Top Words OPTED

#	Deg	Pag	Clo	Bet
1	be	be	be	see
2	have	have	have	make
3	see	see	make	part
4	make	not	see	alt
5	use	make	part	form
6	pertain	manner	use	state
7	act	act	form	be
8	also	use	act	call
9	state	part	person	use
10	not	state	set	set
11	form	alt	call	take
12	part	person	state	act
13	call	thing	also	scale
14	alt	pertain	give	have
15	quality	place	take	manner
16	manner	form	point	point
17	person	word	run	body
18	place	certain	out	place
19	same	quality	place	line
20	body	time	right	give

Top Words DRAE

#	Deg	Pag	Clo	Bet
1	decir	algo	decir	hacer
2	persona	decir	ser	dar
3	otro	ser	persona	decir
4	ser	otro	otro	acción
5	tener	no	algo	estar
6	hacer	persona	tener	tener
7	algo	hacer	hacer	efecto
8	acción	tener	estar	persona
9	estar	cosa	cosa	medio
10	perteneciente	acción	dar	agua
11	relativo	estar	no	parte
12	no	dar	como	punto
13	cosa	como	más	cuerpo
14	efecto	efecto	parte	ser
15	como	relativo	acción	tiempo
16	parte	perteneciente	alguno	cosa
17	dar	pertenecer	medio	relativo
18	muy	parte	poder	derecho
19	más	poder	muy	mano
20	alguno	alguno	poner	estado

Table 2: Top words in OPTED and DRAE under diverse centrality measures: Degree (Deg), PageRank (Pag), Closeness (Clo), and Betweenness (Bet) Centrality. Note that there is a high degree of common notions among the top ranked words in the English and Spanish dictionaries.

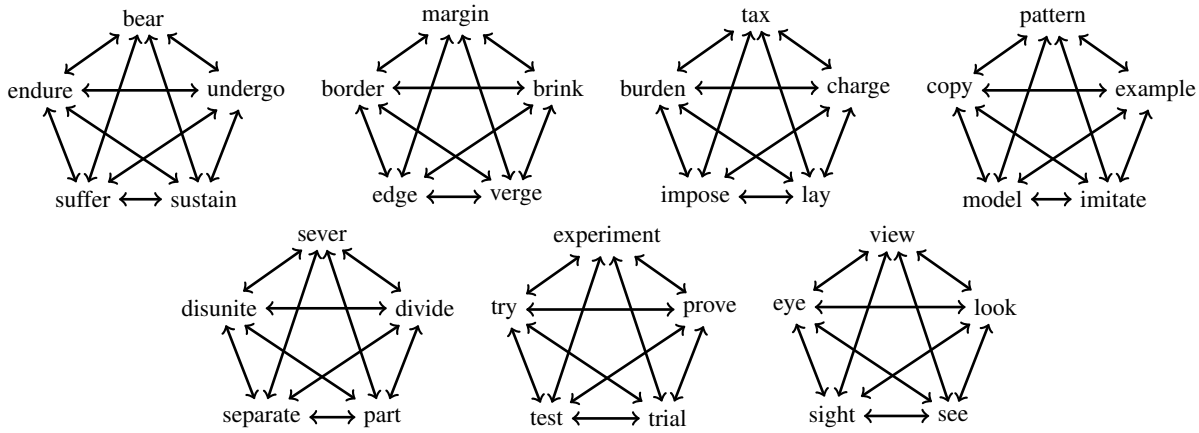


Figure 4: The only seven cliques of size 5 in OPTED (there are no bigger cliques). These words use each other in their definitions. Note their semantic closeness.

to the network core refers to a central and densely connected set. In the other hand, the periphery denotes a sparsely connected and non-central set of nodes that are linked to the core.

There are several types of core structures(Csermely et al., 2013): "traditional" core-periphery networks, rich-club networks, nested networks, bow-tie networks and onion networks. Intuitively, a dictionary network should follow one of these structures. The production of learner's dictionaries that uses a defining vocabulary to write all the definitions, or the simplification of languages through the definition of a small set of words(Ogden, 1930) supports this intuition. Unfortunately, to the best of our knowledge, there is no categorization of the core structure of dictionary networks.

4 Ogden's Basic English

Ogden's Basic English is an English-based controlled language created by Charles Kay Ogden in 1930. It is a simplified subset of the English language. According to Ogden, it is "a system in which everything may be said for all the purposes of everyday existence" (Ogden, 1930). This subset consists of 850

words[4]. The rationale of the choice of words is explained as follows (Ogden, 1930):

> The greater part of the words in use are shorthand for other words. Most common words are colored by our feelings, the words express judgment of our feelings in addition to their straight forward sense. It is generally possible to get to the factual level without much trouble.
> By putting the word to be tested in relationship with other possible words, questions can be framed in the form, "What word takes the place of the word in the middle in this connection?" Puppy is a Dog and time, young. Bitch is a Dog and sex, female. There are thirty lines for thirty sorts of questions.
> Questions of what a word will do for us has little relation to the number of times it is used in newspapers or letters.
> The number of 850 was found with 600 names of things, 150 are names of qualities, and the last 100 are the words which put the others into operation and make them do their work in statements.

Clearly the main arguments for the choice of the words are linguistic. In what follows, we will attempt to capture these words by purely graph-theoretical methods, thus shedding some light on the essential structure of Ogden's basic vocabulary in the network of the language. For our experiments we use the OPTED dictionary, that contains 803 words of the 850 of the Ogden's vocabulary.

4.1 Centrality Measures

The first naive hypothesis is that Ogden's set has good correlation with "central" nodes in the dictionary network. We investigated this with four classic centrality measures: *Degree* (most central nodes are those with higher number of adjacent nodes), *Closeness* (most central nodes are those that minimize the sum of the "distance" to other nodes in the graph), *Betweenness* (counts the number of shortest paths between all pairs of nodes passing through a given node), *PageRank* (essentially tells the number of steps taken to reach the node by a random walk starting from an arbitrary node).

We took the best k nodes for each centrality measure and every $0 < k \leq 803$, and checked how many of Ogden's words are in each of these sets. From the results (Figure 5) it follows that none of the centrality measures do a good job capturing Ogden's Basic English.

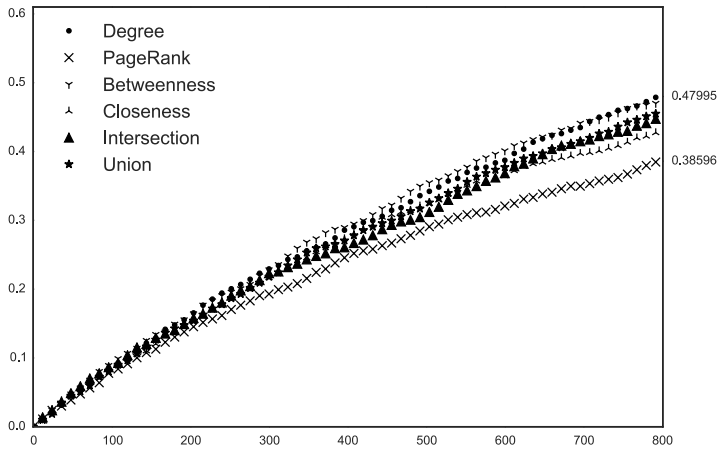


Figure 5: Ogden's Basic English words in top 800 words using different centrality measures. X-axis indicates k top-words and Y-axis, the percentage of Ogden's words in that set. Centralities by themselves are not a good method to capture the notion of importance that Ogden's Basic English represents.

The best measure in this task is degree centrality that captures almost 48% of Ogden. On the other hand, PageRank has the worst performance, capturing only 38.6%. In some sense we knew that degree centrality (which captures frequency) should perform poorly because Ogden stated explicitly that "what

[4]The list of the words can be seen in `http://ogden.basic-english.org/wordalph.html`

a word will do for us has little relation to the number of times it is used in newspapers and business letters". More surprising is the performance of PageRank, one of the most popular centrality measures today, used in multiple areas like ranking webpages, sense disambiguation (Mihalcea, 2004a), keywords and sentences from text (Mihalcea, 2004b), among others.

Group Centrality. Refining the idea, one could hypothesize that the problem is with *individual* centrality. The meaning of words is essentially a network property and not an individual one. There is an extended notion of centrality, called *group centrality* (Everett and Borgatti, 1999), that captures "centrality" of groups, not individuals. Unfortunately it is still not well developed, algorithmically.

We performed some experiments in this direction with groups of Ogden's words. We ranked Ogden's set using PageRank (seems the most promising to capture word senses (Abbott et al., 2012; Yeh et al., 2009)) and formed two groups, one with the top third and the other one with the bottom third of Ogden. As comparison and baseline, we extracted two sets of the same size from the set of words in the OPTED dictionary, one using the top nodes based on frequency, and the other one using a random selection. Results can be seen in Table 3.

	Degree	PageRank	Closeness	Betweenness
Ogden's	10 568	0.0310	0.5547	$4.06 \cdot 10^8$
Frequency	11 460	0.0314	0.5522	$4.40 \cdot 10^8$
Random	5 670	0.0129	0.5277	$2.10 \cdot 10^8$

(a) Top third set from 803 nodes (268 nodes).

	Degree	PageRank	Closeness	Betweenness
Ogden's	8 486	0.0157	0.5464	$2.95 \cdot 10^8$
Frequency	10 314	0.0199	0.5589	$3.41 \cdot 10^8$
Random	3 394	0.0097	0.5122	$1.34 \cdot 10^8$

(b) Bottom set from 803 nodes (268 nodes).

Table 3: Group Centrality for subsets of 803 words (nodes) chosen from three different sources: Ogden's set of words; selected from the OPTED dictionary by best frequency; chosen from OPTED at random.

For each of them we tested the four group centrality measures. Table 3 sheds some light on the existence of different types of roles in Ogden's set of words. The top third Ogden is rather aligned with classic centrality in the network (PageRank, many connections, in the middle of paths, etc.). On the contrary, the bottom third of Ogden behaves very much like random selection regarding PageRank and strongly diminishes its degree. This points to a role of covering an ample part of the network or being "spread" around the network. Though only slightly, this is further supported by the numbers of closeness. The closeness value of Ogden's bottom third is smaller than Ogden's top third (contrary to frequency that increases). The numbers are far from being conclusive due to the limits of the experimentation. As a baseline to compare to Ogden's top and bottom third, we had to use individual rankings, because we could not compute the actual (and ideal) group centralities due to lack of good algorithms and libraries (the problem is known to be NP complete (Garrido, 2016)).

In conclusion we state (although cannot explain well its rationale at this stage) that centrality measures inspired basically on social networks cannot be directly applied in this area. This points to the need for more sophisticated types of centrality measures for semantic networks (if the notion makes sense at all in the area), and in particular for dictionary networks.

4.2 Strong components of graphs

There are graph-theoretical notions about what the "core" (kernel) of a graph is, mainly using connectivity notions. For our dictionary network they seem promising under the hypothesis that connectivity (relationship) between and among groups of words is at the base of language.

We already saw in the component analysis (which holds for any network) that for our purposes one easily can get rid of more than 2/3 of the words in the OPTED dictionary by eliminating those words that are not used to define others (i.e. are "terminal" in some sense).

One can conduct a finer analysis as shown in Figure 6. From the whole OPTED network (which contains 803 words of Ogden) one can get the *strongly connected component* (SCC), those words that, by means of a cycle, are "used" in some sense to define themselves recursively. It has $23,360$ nodes and 802 of Ogden (99.87%). The discarded words (approx. 3/4 of the total) are those that either are terminal (not used to define other words) or n-th level terminals (and terminals after eliminating the terminals and so on).

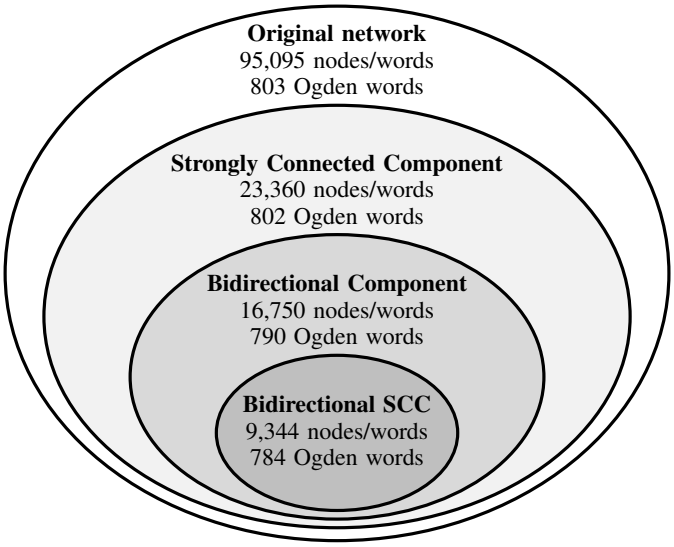


Figure 6: Connectivity analysis of components of OPTED network: The complete graph, Strongly Connected Component (words that recursively define themselves), Bidirectional Component (words that mutually need each other in order to be defined), and Bidirectional SCC. In the latter component only 3% of Ogden's words are lost), showing that Ogden's words strongly need each other.

Next, we consider a strong notion of connectivity: two words are connected if they are mutually used in the definition of the other (*e.g. fire* and *light*). Considering the subgraph induced by this relation, the Bidirectional Component (BC), one gets $16,750$ words, which contain 790 of Ogden (96.89%).

From here one can consider the biggest strongly connected component of BC (there are many small islands in BC), called BSCC in the figure, that has $9,344$ nodes and 784 words of Ogden (97.63%). This shows that Ogden is strongly correlated with these graph theoretical notions.

Picard et al.(2009) explored a notion of core (grounding kernel, which essentially recursively eliminates terminal words) and got a graph of 10% of the original graph. In size it matches our BSCC. Levary et al.(2012) used this notion in eXtended WordNet ($79,689$ nodes) and additionally collapsed synsets in one word, getting a core of $1,595$ nodes. In this core there are 314 Ogden words (52% of the part of Ogden they considered and 36.9% of total Ogden).

From these data, it seems that our BSCC is reaching the limit of the reduction of the English Dictionary (like OPTED) that can be obtained using only connectivity notions in order to capture most of Ogden's words (we are losing only 3% of all Ogden words). The challenge now is how to continue shrinking this graph while keeping most of Ogden's Basic English inside.

5 Conclusion

We provided evidence that dictionary networks share a common network structure and have a great potential to help understanding some properties of languages. We showed weaknesses and strengths of classical network notions in studying properties of dictionary/semantics networks. The results of this study highlight the need to devise more elaborated (than the classical ones) notions of centrality to understand and rank words and sets of words.

Acknowledgement The authors are funded by the Millennium Nucleus Center for Semantic Web Research under Grant NC120004. Garrido is also funded by CONICYT under grant CONICYT-PCHA/Doctorado Nacional/2015-21150149.

References

Joshua T. Abbott, Joseph T. Austerweil, Thomas L. Griffiths. 2012. Human memory search as a random walk in a semantic Network. Advances in Neural Information Processing Systems. 25, 3050-58, 2012.

Robert Alfred Amsler. 1980. The Structure of the Merriam-Webster Pocket Dictionary. PhD. Thesis. Department of Computer Sciences. University of Texas. Austin, Texas.

Vladimir Bagatelj, Andrej Mrvar and Matjaž Zaveršnik. 2002. Network analysis of dictionaries. University of Ljubljana, Institute of Mathematics, Physics and Mechanics. Dept. of Theoretical Computer Science. Preprint series, Vol. 40 (2002), 834.

Nicoletta Calzolari. 1984. Detecting Patterns in a Lexical Data Base. *Proceedings of the 10th International Conference on Computational Linguistics and 22nd annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 1984.

Nicoletta Calzolari and Eugenio Picchi. 1988. Acquisition of Semantic Information From an On-Line Dictionary. *Proceedings of the 12th conference on Computational linguistics-Volume 1*. Association for Computational Linguistics, 1988.

Elizabeth Costenbader and Thomas W. Valente. 2003. The stability of centrality measures when networks are sampled. *Social Networks*. 25 (2003) 283-307

Peter Csermely, András London, Ling-Yun Wu, and Brian Uzzi 2013. Structure and dynamics of core/periphery networks. *Journal of Complex Networks 1.2 (2013): 93-123*.

William Dolan, Lucy Vanderwende, and Stephen D. Richardson. Automatically deriving structured knowledge bases from on-line dictionaries. *Proceedings of the First Conference of the Pacific Association for Computational Linguistics*. 1993.

S. N. Dorogovtsev, J. F. F. Mendes. 2001. Language as an Evolving Word Web Proceedings of the Royal Society of London B: Biological Sciences. 2001, 268 2603-2606.

M. G. Everett and S. P. Borgatti. 1999. The Centrality of Groups and Classes. *Journal of Mathematical Sociology*, 23(3): 181-201.

Linton C. Freeman. 1978. Centrality in Social Networks. Conceptual Clarification. *Social Networks*. 1 (1978/79) 215-239.

Camilo Garrido, Ricardo Mora and Claudio Gutierrez. 2016. Group Centrality for Semantic Networks: a SWOT analysis featuring Random Walks. *ArXiv Preprints*, `https://arxiv.org/abs/1601.00139`

Thad Hughes, Daniel Ramage. 2007. Lexical semantic relatedness with random graph walks. EMNLP-CoNLL, 581-589.

Jure Leskovec and Andrej Krevl. 2014. SNAP Datasets: Stanford Large Network Dataset Collection. `http://snap.stanford.edu/data`

David Levary, Jean-Pierre Eckmann, Elisha Moses, and Tsvi Tlusty. 2012. Loops and Self-Reference in the Construction of Dictionaries. *Physical Review X*, 2, 031018 (2012).

Ken C. Litkowski. 1978. Models of the Semantic Structure of Dictionaries. *American Journal of Computational Linguistics*, Microfiche 81, Frames 25-74.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 55-60.

John Matta, Jeffrey Borwey and Gunes Ercal. 2014. Comparative Resilience Notions and Vertex Attack Tolerance of Scale-Free Networks. *ArXiv Preprints*, `http://arxiv.org/abs/1404.0103`

Rada Mihalcea, Paul Tarau and Elizabeth Figa. 2004. PageRank on Semantic Networks, with Application to Word Sense Disambiguation. *Proc. COLING '04*.

Rada Mihalcea and Paul Tarau. 2004. TextRank Bringing Order into Texts *Proc. Association for Computational Linguistics 2004*.

George A. Miller. 1995. WordNet: A lexical database for English. *Communications of the ACM*, 38(11):39-41.

Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network. *Artificial Intelligence*, 193:217-250.

Mark Newman. 2003. The structure and function of complex networks. *SIAM review*, 45(2):167-256.

Charles Kay Ogden. 1930. *Basic English: A General Introduction with Rules and Grammar*, London, K. Paul, Trench, Trubner & Co.

Olivier Picard, Alexandre Blondin Massé, Stevan Harnard, Odile Marcotte, Guillaume Chicoisne, Yassine Gargouri. 2009. Hierarchies in Dictionary Definition Space. *23rd. Annual Conf. on Neural Information Proc. Systems: Workshop on Analyzing Networks and Learning With Graphs*, Vancouver BC, Canada, 11-12 Dec. 2009. Preprint *arXiv:0911.5703*.

Alain Polguère. 2014. From Writing Dictionaries to Weaving Lexical Networks. *Int J Lexicography*, (2014) 27 (4): 396-418.

Mark Steyvers and Joshua B. Tenenbaum. 2005. The Large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Science*, 29 (1): 41-78.

Stanley Wasserman and Katherine Faust. 1994. *Social Network Analysis. Methods and Applications*, Cambridge Univ. Press, Cambridge, UK.

Yorick Wilks 1993. Providing machine tractable dictionary tools. *Semantics and the Lexicon*. Springer Netherlands, 1993. 341-401.

Yorick Wilks, Dan Fass, Cheng-ming Guo, James E. McDonald, Tony Plate, and Brian M. Slator. 1988. Machine Tractable Dictionaries as Tools and Resources for Natural Language Processing. *Proceedings of the 12th conference on Computational linguistics-Volume 2*. Association for Computational Linguistics.

Eric Yeh, Daniel Ramage, Christopher D. Manning, Eneko Agirre, Aitor Soroa. 2009. Wikiwalk: random walks on wikipedia for semantic relatedness. Workshop on Graph-based Methods for NLP, 4149, Stroudsburg, PA, USA.

Nicola Zingarelli. 1970 Vocabolario della lingua italiana, a cura di Miro Dogliotti, Luigi Rosiello & Paolo Valesio. *Bologna: Zanichelli* (1970).