UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN

A STRUCTURAL ANALYSIS OF DICTIONARIES AS SEMANTIC NETWORKS

TESIS PARA OPTAR AL GRADO DE
DOCTOR EN CIENCIAS, MENCIÓN COMPUTACIÓN

CAMILO FERNANDO GARRIDO GARCÍA

PROFESOR GUÍA:
CLAUDIO GUTIÉRREZ GALLARDO

MIEMBROS DE LA COMISIÓN:
AIDAN HOGAN
GUILLERMO SOTO VERGARA
GILLES SÉRASSET

Este trabajo ha sido parcialmente financiado por
CONICYT PCHA/Doctorado Nacional

SANTIAGO DE CHILE
2021

Resumen

El lenguaje es una cualidad humana que nos permite, entre otras cosas, comunicarnos organizarnos y compartir conocimiento. Es un sistema dinámico en una constante evolución. Existen diversas metodologías lingüísticas para abordar la investigación del lenguaje. Un enfoque para investigar aquellas preguntas ha sido estudiar grandes redes de significados. Construir aquellas redes puede lograrse usando distintas fuentes de datos. En este trabajo nos concentramos en un tipo de fuente: Diccionarios.

Las palabras son bloques de construcción básicos de estructuras de significados más complejos, mientras que la red de sus relaciones puede ser considerado el esqueleto que los mantiene unidos. Los diccionarios proveen una imagen del léxico y, por lo tanto, son una fuente útil para obtener tales esqueletos de significado. La información semántica de los diccionarios puede ser organizada como redes usando un modelo técnicamente simple, pero conceptualmente poderoso donde nodos representan las entradas del diccionario y los arcos representan la relación "es usada para definir". El argumento simple es el siguiente: La definición de una palabra involucra recursivamente nuevas palabras, por lo tanto, nuevas acepciones y significados. De esta manera, un diccionario puede ser visto naturalmente como una red.

En esta tesis estudiamos y analizamos redes de diccionario para respaldar la hipótesis que las redes léxicas subyacentes en diccionarios son buenas fuentes de material para estudiar lenguaje natural y obtener valiosos resultados. Mostramos como las redes de diccionario pueden tomar un rol muy relevante al momento de estudiar la semántica del léxico, por ejemplo, al estudiar la evolución y proximidad de nubes de significados o la búsqueda de palabras relevantes y centrales.

Nos enfocamos en tres aspectos de redes de diccionario. Primero, sistematizamos la estructura de tales redes. Descubrimos que comparten una estructura común global y local. La homogeneidad de sus propiedades no parece estar asociada al idioma del diccionario ni tampoco al año de su publicación. Segundo, analizamos la evolución del diccionario español a través de los años. Descubrimos que la incorporación y eliminación de palabras no afectan a esta estable estructura de red. Los cambios son detectables en los subgrafos que rodean a cada nodo, dándole a cada palabra una red de significados de sutil evolución. Tercero, nos concentramos en cómo puede extraerse conocimiento lingüístico usando redes de diccionarios. Descubrimos que diccionarios son valiosas fuentes de *proximity data* (grado de cercanía entre palabras) y que puede extraerse usando sólo técnicas estándares de análisis de redes.

Los resultados nos muestran el tipo de conocimiento que las redes de diccionario pueden entregar. Éstos no podrían haber sido obtenidos analizando sólo el texto o las definiciones individualmente. Conjeturamos que los resultados obtenidos en este trabajo son válidos no sólo para inglés y español (los idiomas centrales de esta tesis) ya que no sacamos partido de ninguna característica particular de estos idiomas. Sería interesante continuar este estudio en otros idiomas así como extenderlo a diccionarios especializados.

Abstract

Language is a human trait that allows us, among other things, to communicate, organize, and share knowledge. It is a dynamic system in a constant evolution. There are several linguistic methodologies to approach the investigation of language. One such approach to investigate these questions has been studying big networks of meanings. Building such networks can be achieved using several types of sources. In this work we concentrate on one such source: dictionaries.

Words are basic building blocks of more complex meaning structures, while the network of their relationships can be considered as a skeleton that holds them together. Dictionaries provide snapshots of the lexicon and therefore provide a useful source to obtain such skeletons of meaning. The semantic information of dictionaries can be organized as networks using a technically simple though conceptually powerful model where nodes represent dictionary entries and edges represent the relation "is used to define". The simple argument goes as follows: 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.

In this thesis, we study and analyze dictionary networks to support the hypothesis that underlying lexical network in dictionaries are good sources of material for studying natural languages and yield useful results. We show how dictionary networks can play a highly relevant role when studying the semantics of lexicon, for example, in evolution and proximity of clouds of meanings or the pursuit of relevant and central words.

We focus on three aspects of dictionary networks. First, we systematize the structure of such networks. We found that they share a common global and local structure. The uniformity of their properties does not appear to be associated to the language of the dictionary nor the year of publication. Second, we analyze the evolution of the Spanish dictionary over the years. We found that the incorporation and removal of words do not alter this stable network structure. Changes are noticed in subgraphs that surround each node, giving each word a subtle evolving semantic network of meanings. Third, we focus on how linguistic insights can be extracted using dictionary networks. We found that dictionaries are a rich sources of proximity data (degree of closeness of concepts) and that such data can be extracted using only standard network analysis techniques.

The results show us the kind of linguistic insights that dictionary networks can provide. These insights could not be obtained by just analyzing the text or definitions individually. We hypothesize that the results obtained in this work are valid beyond English and Spanish (the main languages of this thesis) as we do not exploited any particular feature to these languages. It would be interesting to continue this study in other languages as well as to extend it to specialized dictionaries.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

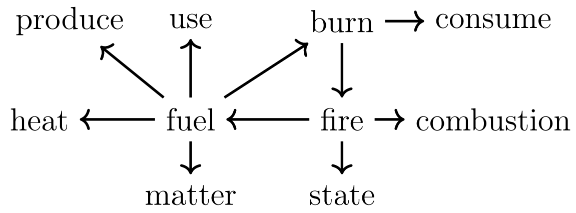
Language is an important part of our lives. It is a unique human trait that differentiates us from other species and marks a major moment in our evolutionary line [97]. It allows us to communicate, organize, and share knowledge. It has been an important factor in the speed of the evolution of humanity and its development [107]. Language is also an integral part of our society and culture. There are still discussions about whether it is culture that molds language or it is language that influences culture. A life without language seems virtually impossible for humans.

Language is not a fixed and static system. Like everything in nature, it is in a constant state of evolution. Several characteristics of language change over time: phonetic changes; pronunciation changes, lexical changes, semantic changes, and syntactic changes. New words are added to the lexicon, due to the introduction of a new, previously incommunicable concept, or to increase the different ways of mentioning an existing concept. Some words experience some slight changes in their meanings to adapt to new cultural trends. And a few words are not used anymore. Language is thus in a constant state of change. It ceases to change only when it has ceased to be spoken or has become a dead language

Even though language is an important part of our lives, it is still an intriguing product of the human mind. There is a lot that we do not know about its nature, structure, and evolution. Researchers have been studying language for a long time, trying to understand and explain how language works.

With the availability of computers in the sixties it became feasible to handle and thus study big networks of meanings. A primary model was that of words and bindings among them. Several models support this approach, such as the semantic memory of Collins and Quillian in 1969 [22] and the frames of Filmore et al. in 1976 [41]. The former is one of the first models of semantic memory where concepts were nodes linked by propositions (e.g., the nodes for *canary* and *animal* were connected via an *is-a* link). In the latter, concepts are related to semantic frames: a collection of facts or a coherent structure of related concepts that specify features (attributes, functions, interactions, etc.)

The building of such networks can be done from different types of sources. In this work



Fire: Fuel in a state of combustion.

Fuel: Any matter used to produce heat by burning.

Burn: To consume with fire.

Figure 1.1: The network built from the entries *fire*, *fuel*, *burn*, and their definitions.

we concentrate on one such source: dictionaries. As Ober and Shenaut write, “the intended metaphor is that of entries in a dictionary or encyclopedia bound together into intricate constellations of meaning as they recur in the bodies of definitions.” [84]. The linguist Kenneth Litkowski [59] was among the first researchers to raise awareness of such underlying lexical networks in a dictionary and argued that dictionaries have not been given their due as sources of material for natural language. In this thesis we pursue this hypothesis.

The model Litkowski proposed in 1978 has the advantage of being technically simple though conceptually powerful. He proposed that the semantic information of dictionaries be organized as networks where nodes represent dictionary entries and edges represent the relation “is used to define”. The naive argument goes as follows: 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, ..., w_n that defines it. Roughly speaking, the nodes of the network will be the words and for each word w there will edges $w \rightarrow w_{\mathrm{i}}$ for each $w_1, ..., w_n$ defining it (modulo standard normalization: repetitions, lemmatization, etc.) (see the example in Figure 1.1).

Although this model abstracts away many relevant pieces of information present in dictionaries like the syntactic structure of sentences, the different senses of a word, grammatical roles, etc., we will see that it allows powerful analysis. In fact, the actual construction is subtle as we will see and its simple structure allows using standard machinery for network analysis. Essentially, what is going on, is that the network of a dictionary permits to explore and study “global properties” of it, i.e. those topological properties that emerge from the network of words and that cannot be captured “locally”, that is, by only considering isolated words and their definitions [7, 8, 74, 81]. A simple example of a global property is the strongly connected component (there is a direct path from each word to any other word). In the context of dictionaries, this means starting from any word w one can go through definitions and loop back to w (the path *fire*, *fuel*, *burn*, *fire* in Figure 1.1).

A natural question arises: to what extent can dictionary networks (a particular form of organizing the lexicon of a language) shed light and knowledge on the lexical network of a language? In this work, we study and analyze dictionary networks. We show that they have distinctive properties as compared to other types of networks such as social networks. For example, we show that they are highly resilient, have a graceful decay, a shared local structure, among other features. We also show that their characteristics depend neither on the edition of the dictionary, nor on the language, suggesting that there are inherent properties of dictionary networks. We also show that these networks can provide other type of linguistic knowledge, such as identifying interesting categories of words; the high resilience to change of networks of meaning; a slow evolution at a global scale; etc. We think that one can safely hypothesize that these may be properties of lexical networks in general. If this is the case, dictionaries could be of much help as a model for the study of properties of lexical networks of human languages.

1.1 Objectives

In this dissertation, we tackle the problem of the lack of evidence towards the hypothesis suggested by Litkowski: The theory of directed graphs is a suitable framework for systematizing definitions from dictionaries in order to extract their semantic content and yield useful results.

The main goal of this thesis is to show how dictionary networks play a fundamental role and that without them it would be very difficult or virtually impossible to achieve the same results, establishing a groundwork about the potentialities of their structure and the relationship between words in their definitions.

We aim for this work to be part of the groundwork for any structural theory of dictionaries. We expect that these kinds of approaches, as hinted at by some samples, help linguists to study the semantic relationship of words. This work should give insights to research works on how to approach this kind of networks: it should describe some of their common properties, how they are measured and why there are believed to be important. We focus our study on two dictionaries: The Online Plain Text English Dictionary and the *Diccionario de la Lengua Española* (Dictionary of the Spanish language). The former is based on the 1913 US Webster's Unabridged Dictionary and the latter is the most authoritative dictionary in Spanish first published in 1780.

We divide our work in three objectives.

1.1.1 The structure of dictionary networks

The first objective is to determine and systematize the basic structure of networks underlying dictionaries. To achieve this objective we identify common global and local properties between different dictionary networks, such as clustering coefficient, average length path, and component sizes. We observe how classical centrality measures behave in this domain and how they correlate to linguistic levels of word relevance. We identify which are the most important traits that differentiate dictionary networks from networks from other disciplines and propose a model for random generation of these networks.

1.1.2 The evolution of dictionary networks

It is clear that the languages evolve; some words are added and some are not used anymore. The second objective is to identify if the structure of dictionary networks evolve too and how they change over time. We analyze the structure of the *Diccionario de la Lengua Española* in the span of almost 100 years. We study the same common properties —global and local— that we find in the development of the first objective. We find whether these properties change following a pattern, randomly, or do not change at all.

1.1.3 Linguistics insights using dictionary networks

The third objective is to exhibit how linguistic insights can be extracted from dictionary networks. Particularly, we focus on automatically obtain proximity data from dictionaries. Proximity is a notion that indicates the level of which things belong together psychologically. Usually, obtaining proximity data requires experts, a large amount of people, or large corpora, becoming prohibitive to many languages. Since, dictionaries are a fair representation of the vocabulary of a language and collective heritage of native speakers, we exploit the implicit network representation of a dictionary, using classical measures of similarity between the vertices of a graph to obtain clouds of lexical proximity.

1.2 Contributions

The contributions of this thesis dissertation are summarized as follows:

- A study about the share common global and local structure of dictionary networks.
- The necessity of a notion of centrality in a linguistic context.
- A model to generate random dictionary networks.
- A study about the evolution and the resilience of dictionaries over time.

As a result of the work conducted in the present dissertation, the following articles have been published:

- Camilo Garrido and Claudio Gutiérrez. Dictionaries as Networks: Identifying the graph structure of Ogden's Basic English *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers*, December 11-16, 2016, Osaka, Japan
- Camilo Garrido, Claudio Gutierrez and Guillermo Soto. A New Class Of Proximity Data Obtained From Dictionary Networks. *The 41th Annual Meeting of the Cognitive Science Society, CogSci 2019* (Poster)
- Camilo Garrido, Claudio Gutierrez and Guillermo Soto. The Semantic Network of the Spanish Dictionary during the last century: structural stability and resilience. *Electronic lexicography in the 21st century: Proceedings of eLex 2019 conference. Lexical Computing, 2019.*
- Camilo Garrido, Claudio Gutierrez and Guillermo Soto. Dictionaries as networks: Revisiting Litkowski's hypothesis *Computational Linguistics Journal, MIT Press.* (In revision; submitted February 2020)

1.3 Outline and Structure of the thesis

This dissertation is structured as follows:

- **Chapter 2: Literature review.** This chapter describes research and work relevant to this thesis, related to dictionaries as sources of knowledge, dictionary modeling as networks, lexical databases, semantic networks and synonym extraction.
- **Chapter 3: Dictionary network models.** We discuss different approaches for modeling dictionaries as networks. We also discuss the advantages and limitations of a comprehensive dictionary network model and the minimal model. We systematize the basic structure of dictionary networks. We observe that they share common global and local properties.
- **Chapter 4: Data collection of dictionaries.** This chapter details the process of collecting the data of the dictionaries used in this thesis, as well as the pre-processing and cleaning of this data.
- **Chapter 5: Ogden's Basic English.** In this chapter we show how linguistic-relevant words form part of a strongly connected core, while classic measures of relevance in networks fail to capture them.
- **Chapter 6: The structure of dictionary networks.** This chapter analyzes the presence of triads in dictionary networks and shows their relevance characterizating the networks' structure. We propose a model for random generation of these networks.
- **Chapter 7: The Spanish dictionary during the last century.** In this chapter, we study the evolution of the semantic network of the Spanish dictionary during the last century by analyzing the permanence and changes of its structural properties. We discuss the stability of global structural properties and the relevance of local properties in the evolution of lexicon.
- **Chapter 8: A new class of proximity data.** In this chapter, we present a method to automatically obtain proximity data from dictionaries. We present how we can take advantage of the network representation of a dictionary to obtain clouds of lexical proximity. We describe the evaluation process with native speakers and discuss the results.
- **Chapter 9: Conclusions.** Finally, we present conclusions, contributions and discuss limitations of the proposed solution as well as the direction of future research.

Chapter 2

Related Work

In this chapter, we present the state of the art regarding dictionary networks, including dictionaries as a source of knowledge and structural analysis from a network point of view. We also present applications of dictionaries as sources of knowledge and efforts in extracting semantic information from dictionaries.

We performed a comprehensive review of all topics related to dictionary networks that we were aware of. Some of them go beyond the scope of this thesis, but we thought that the reader would benefit from a broad overview of the emergence of the systematic study of dictionary networks.

2.1 Language as a network

Words in human language interact in non-random ways. Several researchers have modeled and exploited these interactions to analyze different aspects of language. For example, the co-occurrence of words in sentences reflect language organization in a subtle manner [40]. This co-occurrence can be described in terms of a graph of word interactions. Such structure displays two features. First, the network follow a small-world structure. Second, they follow a scale-free distribution of degrees. These researchers consider that these observations might be of use to study the evolution and social history of lexicons. Dorogovtsev and Mendes [29] also believe that human language may benefit from its description as a complex network of linked words. They propose a stochastic theory of the evolution of human language based on the interaction of words and preferential attachment. At each step, a new word appears and it interacts (connects) preferentially with old words.

Other researchers have focused on the relationship where two words are connected if they express a similar concept [75]. They showed that such a network follows a small-world structure with a small average shortest path and a scale-free distribution. In addition to the applications for the study of the structure and evolution of languages, they believe that this network can have applications in cognitive science. The human mind is associative and connects similar concepts to retrieve information.

All these properties on language networks are potentially universal features [98]. For instance, these networks are sparse, have a small-world structure, and are heterogeneous (follow a power law). They point to the convenience of developing new network models to analyze and inspect questions of language.

These works make us wonder if the universality of language networks extend to dictionaries. They motivate us to compare the structure and measures of the networks derived from dictionaries built with different objectives and also different languages.

2.2 Structure of Dictionary Networks

Litkowski [59] 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. The goals of his work were, first, to describe how to use the dictionary itself to move towards identification of primitives, and second, to show how this process can be used to provide the capability for discriminating among word senses and characterizing knowledge contained in a definition. He presented three models to represent a dictionary. The basic model uses nodes to represent the word y occurring in the entry being defined as well as those words x occurring in its definition. The edges represent the relation *x is used to define y*. Each word in a definition is lemmatized, since the inflected form of a word does not appear as a main entry in the dictionary. The second model is an enriched version of the first, and incorporated the different senses of the words in a definition as different nodes. The third model considered the nodes as concepts. For instance, let us consider the definition of *broadcast* as "the act of spreading abroad". If "abroad" has two senses (abroad$_1$ and abroad$_2$) then there should be two nodes, one representing "the act of spreading abroad$_1$" and other for "the act of spreading abroad$_2$". It would be no longer valid to say that a node represents a definition, rather they say it represents a "concept". For more details on these models, please see Section 3.1.1

Following these ideas, Batagelj et al. [9] performed a network analysis over two online computer dictionaries. Both of them focus on concepts related to computer science and information studies, such as programming languages, operating systems, networking. They presented the top nodes according to centrality measures, degree, among other network properties. No further analysis was made.

Dictionaries use words to define words; therefore at some point there will be circularities in the definitions. Using dictionary networks built from the synsets (set of synonymous words) of the eXtended WordNet [66, 71, 106], Levary et al. [58] studied loops and self-reference in the definition of words. Synsets were designated as nodes with a directed edge drawn from a node to all of the synset nodes that occur in the definition. They have a great amount of short loops (length ≤ 6), much more than a randomized graph with the same degree distribution. They found that loops are not strictly isolated but are often linked to form larger, yet still semantically coherent, strongly connected components.

Because of the circularities in dictionaries, the meaning of some words must be learned by experience. In their research, Massé et al. [63] addressed the problem of finding a small vocabulary that allow us to learn, only by definitions, the meaning of a larger vocabulary, assuming that we already know the meaning of the words in the small vocabulary. They provided an algorithm to obtain such sets. The principle was to keep removing “sink” words (i.e words that are not used to define others) until no more words can be removed and then do the same with strongly connected components.

Picard et al. [89], using the previous research, reduced a dictionary to grounding kernels: a set of words of about 10% of the dictionary from which all the other words could be defined. The grounding kernels correspond to the result of recursively removing the sink nodes. They have an internal structure with a strongly connected kernel core. A kernel core corresponds to the vertices belonging to the sources of the SCC-quotient graph. A strongly connected component (SCC) is a subset of nodes where every node is reachable from every other node in the subset. The SCC-quotient graph is obtained by collapsing each strongly connected component into a single node. There is an edge between two collapsed nodes (A, B) if there is an edge between nodes (u, v) in the original graph and $u \in A$ and $v \in B$. These kernels allow for defining a hierarchy of definitional distance, and showing it correlates with psycholinguistic variables.

On similar lines, and motivated by the question *How many words are sufficient to define all other words?*, Vincent-Lamarre et al. [110] studied the structure of several English dictionaries. They found that, in theory, for a dictionary about 90,000 words, a person only needs to know by previous experience about 1,400 words to learn the meanings of all the rest. They also discussed the necessity of definitions in dictionaries and life. Not all meanings can be learned by experience. Oftentimes times meaning is transmitted by other people to describe something or communicate knowledge.

Finally, these ideas about dictionary networks can be extended to similar types of linguistic objects, such as semantic networks, lexical networks, thesauri, etc. The work of Steyvers and Tenenenbaum [101] presents an analysis of 3 types of semantic networks: WordNet [71], word association norms [78], and Roget’s Thesaurus [93]. The undirected network of WordNet is built from the relationship between words and their word-meanings. In the directed network from word association norms, a word A points to a word B if A evoked B as an associative response for people. In the directed network from Roget’s thesaurus, a connection is made between a word node and a category node when the word falls into the semantic category. The researchers focused on a statistical analysis rather than on other structural principles such as networks motifs [74, 22]. They showed that these networks have a small-world structure, characterized by sparse connectivity, short average path lengths between words, and strong local clustering. Definitions and examples of these network measures can be found in the work of Newman [81]. Following the idea that these regularities reflect the mechanisms by which semantic networks grow, they proposed a growing model for semantic networks, since more classical models like those of Barabasi and Albert [7] do not correspond to the behavior of the studied semantic networks. The main concern of these model is generating appropriate small-world statistics and power-law connectivity distribution.

Treating thesauri as directed graphs has provided new insights into its macro structure. De Jesus Holanda et al. [25] analyzed the distribution of outgoing and incoming links of an English thesaurus, providing a fitting function for growth. They suggest that words in a thesaurus should be ranked in terms of their connectivity index rather than the standard method (using frequency of words in a corpus).

The work of Litkowski [59] is one of the main inspirations for this thesis. We use the relation *x is used to define y* as the main association in the networks we build and we ponder about his models and expand the ideas to an ideal dictionary model (see Section 3.1.2). There are two main differences between this thesis and the works presented in this section. First, this thesis attempts to maintain the analyses of dictionary networks focused on the dictionary and their definitions. Some of the works cited in this section show interesting features of dictionaries. However, they present them in contrast or referencing particular external sources of data in a specific language. For instance, Picard [89] uses psycholinguistic data and Levary [58] uses historical and etymological data. Both of those datasets are English datasets. It prevents us to attempt similar results, since Spanish is one of the main focuses of this thesis. The second difference refers to the datasets used in the analyses. We focus only on traditional dictionaries where the words of a language are defined. Some of the works mentioned use thesauri, dictionaries of a specific domain (such as ODLIS: Online Dictionary of Library and Information Science), free-association norms and synset-only WordNet [9, 101, 25]

2.3 Lexical Networks

Lexical networks focus on the representation of lexical knowledge as networks. They differ mainly from dictionary networks on how they are built, the sources of data, and the goals they aim to achieve. The most relevant work in this line is WordNet [72, 71, 38]. WordNet is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations. In other words, there are nodes that represent words as we know them (a set of characters) and nodes that represent a concept or meaning. Multiple word are connected to a single word-meaning node if the words are synonymous. A word is connected to multiple word-meaning nodes if it is polysemous. Other important lexical networks are BabelNet [77] and FrameNet [5]. BabelNet is an automatic effort in building a very large, wide coverage multilingual semantic network. This resource is created by linking Wikipedia and WordNet. And FrameNet is a lexical database of English, which aims to be both human- and machine-readable. It is based on the idea that the meanings of most words can best be understood with "characteristic features, attributes, and functions of a denotatum, and its characteristic interactions with things necessarily or typically associated with it." [1].

In this work, we do not make use of lexical networks. They are built manually by experts or automatically integrating manually created datasets. Every lexical network has their own structure and rules. We aim to develop and work with methods that apply to the structure of a dictionary, a widespread structure in many languages and periods of time. Although we make comparisons with WordNet in this thesis, we do not utilize it in its original lexical network structure. We use the pairs word-definition as a regular dictionary does.

2.4 Free-Association Norms

A free-association task consists of people being exposed to a cue word and recording the first response word that comes to their mind. For example, fire → burn; cake → sweet. A collection of responses to each cue given by a large group of speakers, together with the frequency of each response, constitutes a free-association norm. Such norms are believed to reflect the strength of the links between words in the lexicon of an average speaker, and studies in psycholinguistics often rely on this information to explain various cognitive processes related to lexical semantic memory. Free-association norms are useful for constructing maps of the lexical knowledge that is most accessible to people sharing a language and a cultural heritage [78]. Linguistic analyses of the responses can be helpful to determine the likelihood of searching words within a particular domain of information [80]. They also are useful to predict memory performance and other cognitive behaviors [79].

The English language has a predominant presence in the available free-association datasets. Some datasets focus on covering as most words of the lexicon as they can [78], and others focus on having a broad ranging age of participants in the development of the dataset [30]. However, there are efforts on contributing to other languages, such as Spanish [39], French [109], and Dutch [24].

The research on free-association not only focuses on producing such datasets; some researchers focus on modeling and performing a statistical analysis with the intention of explaining cognitive processes in human semantic memory. [65].

As free-association norms are useful for constructing maps of the lexical knowledge, they are useful to validate and contrast the lexical information extracted from dictionary networks.

2.5 Extracting semantic information from dictionaries

The work of Spärck Jones in 1967 [99] is one of the first studies involving dictionaries. She realized that the words and their definitions formed loops. In other words, if you pick a word in a definition and then a word in its definition and so on, you can reach the very first word. Calzolari [14] addresses this issue from an empirical point of view. Following the focus on definitions, the work of Reichert et al. [92] prepared a complete inverted index of the definition text of the dictionary that they called "the concordance index". It listed all defined vocabulary under the words which occurred in their definitions and provides a way of finding all definitions which contain a given genus term.

Dictionaries began to gain attention with the development of machine-readable dictionaries (MRDs). An MRD is a dictionary stored as machine (computer) data instead of being printed on paper. It is an electronic dictionary and lexical database. Olney et al. [86, 87] rendered the Merriam-Webster dictionary into an MRD.

Amsler in 1980 [3] presented some efforts to exploit dictionaries and extract information for applications in computational linguistics. He investigated the possibility of building taxonomies based on the structure of the definition of words. He also offers some insight on the frequency of the vocabulary and semantic ambiguity. A little after, he continued analyzing dictionary definitions to extract a taxonomy for English nouns and verbs [2]. He found that at the top of the hierarchy are the most general elements, senses of words such as "cause", "thing", and "being". These words are all interrelated in the hierarchy, stating that this is evidence of the existence of primitive concepts.

Calzolari and Picchi [15, 16] focused on detecting patterns in the definition of words to extract two kinds of relations: Hyponyms and "restriction". Hyponym is a well-known relation in linguistics that refers to a word of more specific meaning than a general or superordinate term applicable to it. The relation of "restriction or modification" refers to the part of a definition that link the *genus* and the *differentia*. The *genus* corresponds to the part that repeats on other definitions and serves to form families. The *differentia* corresponds to the rest of the definition. For example, the genus is "a motor vehicle" in the following definitions:

Car: a motor vehicle with four wheels.
Motorcycle: a motor vehicle with two wheels.

The extraction of these relations was based almost exclusively on the position of the genus term in the definitions. Chodorow [20] also addressed the issue of extracting hyponym–hypernym relations. His automated mechanism for finding the genus terms is based on the observation that the genus term for verb and noun definitions is typically the head of the defining phrase. This reduced his task to finding the heads of verb phrases and noun phrases. In the same line, Markowitz et al. [62] processed semantically significant patterns in dictionary definitions, concentrating on the start of definitions, but they also demonstrated that important information can be extracted from syntactic and graphemic elements, such as parentheses. The information they extracted involves lexical relationships: taxonomies and set memberships, selectional restrictions, and special subcategories of nouns, verbs, and adjectives.

In a more distinct line, Wilks et al. [114] provided tools that take as input the forms of information given in the Longman Dictionary of Contemporary English (English definitions, syntax codes, subject and pragmatic codes) and provide either: (i) a clustered network of words based on the co-occurrence; (ii) a formalized set of definitions of sense entries in a nested predicate form, where the predicates are a "seed set" of senses; (iii) frame-like structures containing a formalization of the English definitions using English words as predicates.

How the form definitions are built and arranged plays a fundamental role when extracting semantic information. The works presented in this section are mostly focused individually on definitions. They extract information one definition at the time. In contrast, this thesis focuses on the structure and relationship between words and their definitions. Changes in one definition directly affects the semantic information extracted from dictionary networks.

2.6 Structural network analysis

Networks analysis is an important discipline today. The work of Newman [81] is one of the first to give a broad overview of the subject. His work reviews a variety of techniques and models to help us understand or predict the behavior of networked systems such as the Internet, social networks, and biological networks. The techniques include the small-world effect [113], degree distributions, clustering coefficients [8], network correlations, random graph models [7, 32], models of network growth and preferential attachment [33], and dynamic processes taking place in networks. He showed the value of these properties for a variety of different networks taken from the literature. In the same line, Kunegis [54] presents a collection of over 160 network datasets with their corresponding analysis (degree distribution, average shortest path length, distance distribution, etc.).

Complex networks are particular types of networks having patterns of interconnections occurring at higher numbers than those in randomized networks. These patterns or network motifs help to uncover structural design principles. Milo et al. [74] found that networks that belong to the same field have common patterns, and these, in turn, are distinct from networks from different fields. For example, ecological food networks have common patterns that are different from the patterns found in engineering networks. On the other hand, Itzkovitz et al. [50] analyzed the subgraph distribution in random networks. They presented approximate equations for the average number of subgraphs for random sparse directed networks. Most of these subgraphs are of size 3. If we focus on subgraphs of size 3, we have to study the triad census. The triad census corresponds to the frequency of sixteen isomorphism classes of subgraphs of three nodes. Faust [36] provides the necessary background to understand this property. She also examined the triad census as a mean for investigating structural patterns [35], alerting that "caution should be taken in interpreting higher order structural properties when they are largely explained by local network features." A triad census is a summary that retains important information about local features of the network and allows one to test hypotheses about the prevalence of structural properties such as transitivity or intransitivity. Triadic tendencies are important structural features of social networks, particularly in social networks [37]. In order to facilitate the computation and comparison of these tendencies, Holland and Leinhardt [47, 48] provided formulas for the expected numbers in the triad census given the dyad census, i.e., the number of pairs of nodes with no link between them (null), a link in one direction (asymmetric), and a link in both directions (mutual).

We cannot study dictionary networks without a proper structural network analysis. We take inspiration in the works mentioned in this section to gain a better understanding of dictionary networks. We mainly follow the classic network analysis methodologies [81], models of network growth [33, 101], and the triad census [36].

2.7 Applications of dictionaries as sources of knowledge

Dictionaries have manifold applications. There are three areas where dictionaries and their automatization have proved particularly useful: automatic synonym extraction, word sense disambiguation, and word embeddings. Although these lines of researches are not closely related to the objective of this thesis, we added these topics to show the value of dictionaries and areas where the knowledge obtained from dictionary networks may contribute.

Synonyms are used in several NLP tasks, such as information retrieval, machine translation, and generation of texts [117, 90, 55], hence the efforts in extracting synonyms in an automatic approach. The advantage of using dictionaries lies in the coverage, and the availability of such resources. For instance, Muller et al. [76] exploited the semantic distance between concepts in a dictionary. They strongly distinguished two types of objects in the derived network: the definiendum, the word or expression being defined; and the definiens, the proper definition. Once they built the graph, they computed a semantic similarity measure between the nodes. This distance was used to separate candidate synonyms for a given word. Wang et al. [111] utilized the regularity in the composition of dictionary definitions. They provided a better controlled environment for synonym distribution than free-text corpora. They found that using culture specific dictionaries result in the extraction of culture specific synonyms. Some works, rather than producing a defined set of words, produce a ranking of candidate synonym words. One approach is applying a method for finding authoritative pages to a dictionary network, comparing the problem of extracting synonym to searching similar pages on the web [11]. Another approach applies spreading activation: a method in cognitive science for searching associative networks [53].

The main feature of word sense disambiguation using dictionaries is their coverage. They perform independently of annotated data and exploit the graph structure of semantic networks to identify the most suitable meanings, rivaling supervised methods [49]. Usually, researchers use the dictionary as corpora for their methods. However, some approaches apply procedures over networks build from word definitions. Some of these approaches involve using spreading activation models [108], co-occurrence assuming that words that are similar in meaning tend to occur in similar linguistic contexts [88], and semantic proximity measures between words in a dictionary [45].

Vectorial representations of words have grown to play an important role in natural language processing [103]. Learning such representations relies on the distributional hypothesis – words appearing in similar context must have similar meaning and, thus close representations– over large and unlabeled corpora. Dictionaries and their definitions provide lexical and phrasal semantics [46], compositionality [95], and multilingual information [26], used by researchers to improve and extend word representations. There are some efforts in learning word embeddings from definitions only [83, 105, 13]. Although dictionaries are small corpora compared to the corpora used in standard methods [69, 28, 18], vectors derived from dictionaries contain different semantic information from displaying similarities not found in classic word embeddings.

Chapter 3

Dictionaries as Networks

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. In this chapter, we discuss different approaches to build dictionary networks. We examine a comprehensive and a minimal model.

3.1 Construction and Structure

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 along the lines we explained previously, gives rise to different types of nodes and edges. Nodes could have types of n.; n.pl.; a.; v.; v.t.;v.i.; adv.; etc. Edges also can be of different types, according to 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 [59]).

On the other extreme, one can build a simple (naive) network without any typing of 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 hope to keep enhancing this baseline with further linguistic annotations.

3.1.1 Approaches for modeling dictionaries as networks

Litkowski [59] was one of the first to advocate the view of dictionaries as networks and presented several models. The goals of his work were, first, to describe how to use the dictionary itself to move towards identification of primitives, and second, to show how this process can be used to provide the capability for discriminating among word senses and characterize knowledge contained in a definition. He presented four models to represent a dictionary.

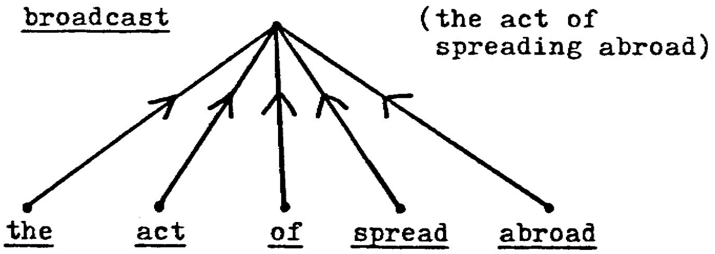


Figure 3.1: Litkowski's first model for dictionaries as networks.

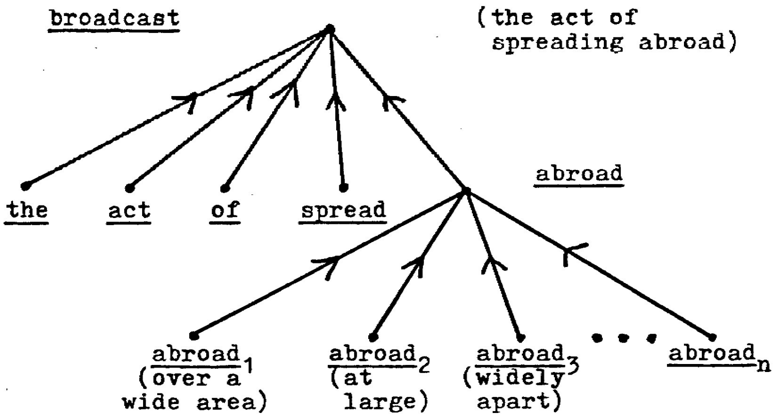


Figure 3.2: Litkowski's second model for dictionaries as networks.

The first model (Figure 3.1) uses nodes to represent the words of the entry being defined and the words composing the definition. The edges represent the relation *x is used to define y*. Each word in a definition is lemmatized, since the inflected form of a word does not appear as a main entry in the dictionary.

The second model (Figure 3.2) incorporated the different senses of the words in a definition as nodes. Letting each point in the first model represent all the definitions of an entry compresses the semantic content of each node. Incorporating nodes for each definition increases the degree of semantic richness in the model.

For the third model (Figure 3.3), Litkowski argued that his first two models do not portray any of the meaning of the dictionary, but rather indicate where particular relationships exist.

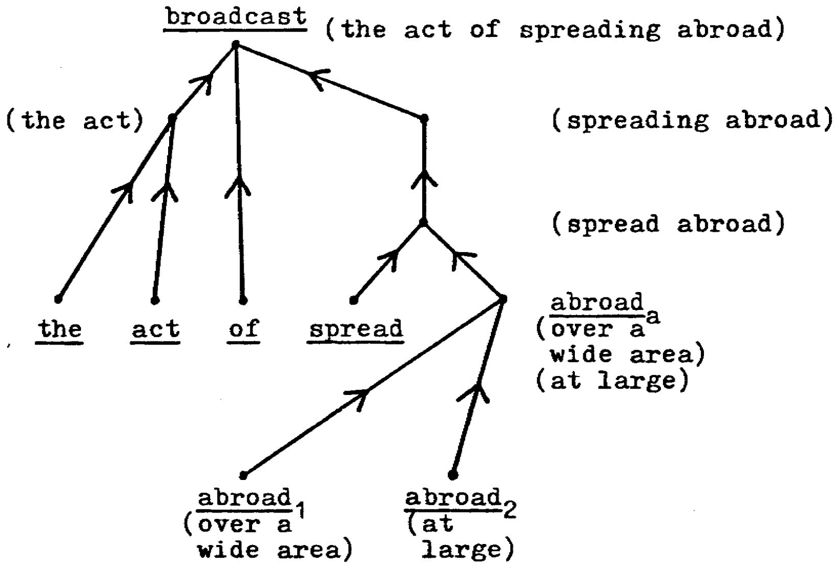

broadcast (the act of spreading abroad)
(the act)
(spreading abroad)
(spread abroad)
the
act
of
spread
abroad_a (over a wide area) (at large)
abroad_1 (over a wide area)
abroad_2 (at large)

Figure 3.3: Litkowski's third model for dictionaries as networks.

This model introduced the syntactical structure of definitions, breaking down definitions into subphrases and then into words.

The fourth model considered the nodes as concepts. For instance, let's consider the definition of *broadcast* as "the act of spreading abroad". If "abroad" has two senses (abroad$_1$ and abroad$_2$) then there should be two nodes, one representing "the act of spreading abroad$_1$" and other for "the act of spreading abroad$_2$". It would be no longer valid to say that a node represents a definition, rather it would represent a "concept".

3.1.2 A comprehensive model that we propose

There are some essential dimensions that should be considered if one would like to have a comprehensive network dictionary model. First, the network basic building blocks should be the words and their relationships, and should allow us to analyze words, definitions, and other types of relationships as a unified whole, in a "global" manner. This should not preclude other types of additional structures. Second, we must be able to reconstruct the (original) dictionary from the network, that is, current texts that define a sense. In this regard, the network should not lose information. For example, consider the words that need each other, that is a word x occurs in the definition of word y, and vice versa, y occurs in the definition of x. This raise several questions like which sense of words are being used in a given definition, in which context it occurs, etc.

In the following we systematically present the features of such a comprehensive dictionary network.

Nodes in the network model

There are three types of nodes in the dictionary network model: Entries, Senses, Words. Entry nodes correspond to entry words in the dictionary. Sense nodes are similar to synsets

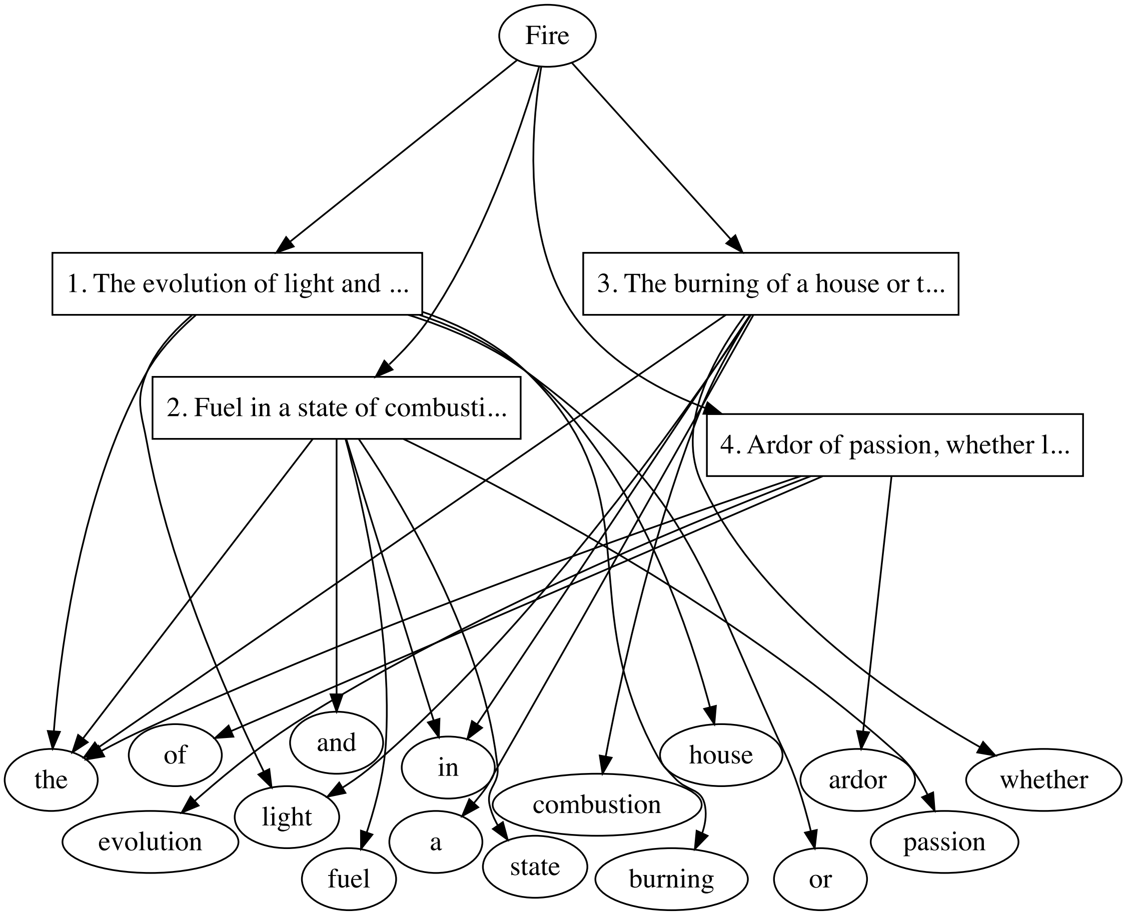

Fire
1. The evolution of light and ...
2. Fuel in a state of combusti...
3. The burning of a house or t...
4. Ardor of passion, whether l...
the
of
and
in
house
ardor
whether
evolution
light
a
combustion
passion
fuel
state
burning
or

Figure 3.4: Different types of nodes of a Comprehensive Dictionary network from the words in the definition of *fire* (Figure 3.4).

in WordNet, the difference is that in this model the senses are taken from the definitions rather than creating sense and assign it to a word form. Following the definition of *fire* (Figure 3.7a), there would be 8 sense nodes: one for "*The evolution of light and heat in the combustion of bodies; combustion; state of ignition.*", one for "*Fuel in a state of combustion, as on a hearth, or in a stove or a furnace.*", and so on. Finally, we have word nodes. Word nodes correspond to words that appear in the definitions. Word nodes entry nodes disjoint. It is not the same "**fire**" (entry) and "**fire**" (word). From the definition of *quick*:

> **Quick (superl.)** Speedy; hasty; swift; not slow; as, be quick.

we would extract 8 word nodes: *speedy*, *hasty*, *swift*, *not*, *slow*, *as*, *be*, and *quick*.

Edges in the network model

Between the 3 types of nodes we identify 4 types of directed relationships:

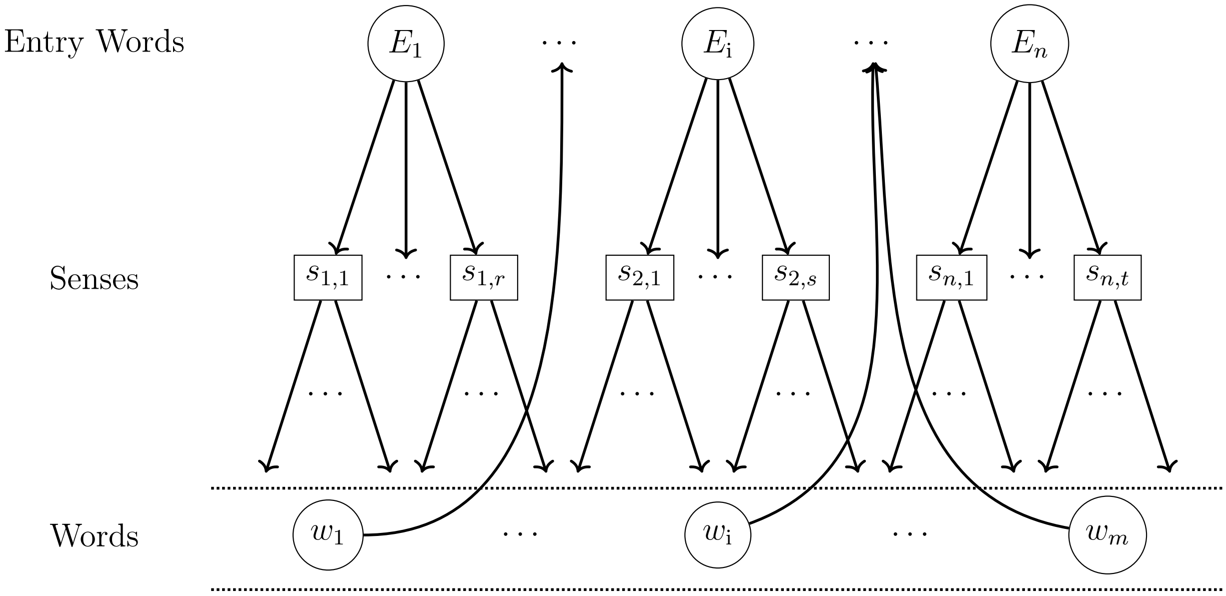


Figure 3.5: Diagram of ideal model for dictionaries as networks.

1. Entry words and senses.
2. Senses and words.
3. Words and senses.
4. Words and entries.

The definition of a word in the dictionary is composed by senses. To keep track of these relationships each entry node points to its corresponding sense nodes. The link between entry node and sense nodes represents the relation "X has a sense Y", meaning that each entry node in the dictionary has a directed link to the corresponding sense nodes. Following the definition of "*fire*" (Figure 3.7a), there would be a link from "*fire*" to "*The evolution of light and heat in the combustion of bodies...* ", from "*fire*" to "*Fuel in a state of combustion, as on a hearth...*", and so on. There also are links from sense nodes to word nodes, since we want to keep track of the words that are used in each sense. For example, the sense node "*Fuel in a state of combustion...*" would point to word nodes "*fuel*", "*in*", "*a*", "*state*", etc. We also need a link between words and senses. Many words are polysemous and we need to know which meaning the word is communicating. Each word sense points to the sense node corresponding to the meaning it want to convey. Following the sense in the last example "*Fuel in a state of combustion...*", we need to clarify if the word "*state*" is a noun about the condition of substance, social position, territory, or nation; or if it is used as a verb expressing a declaration or something set in definite form. The last relationship, words and entries, is needed because in many languages words change to adjust to grammatical tense, number, gender, subject, etc. Taking advantage on the fact that entry words often represented the lemmatized form of words, each word node in the model points to the corresponding entry word according to how it is used in the sentence of the sense. For example, in the sense of "*One of the five terminating members of the hand*" the word node "*terminating*" would point to the entry node "*terminate*" and the word node "*members*" would point to the entry node "*member*".

Order in the network model

Up to this point we have fulfilled two of the three main ideas that a ideal model should follow: The proposed model is based on networks and the main and basic relationship is "x is defined by y". In order to fulfill the third idea, "we must be able to reconstruct the dictionary from the model", we have to introduce an order to the edges of the network model.

If we consider the network derived from the definition of *swift* (Figure 3.6), it would not be a problem to identify the text (words) corresponding to the entry and the sense.

> **Swift (v. i.)** Moving a great distance in a short time; moving with celerity or velocity; feet; rapid; quick; speedy; prompt.

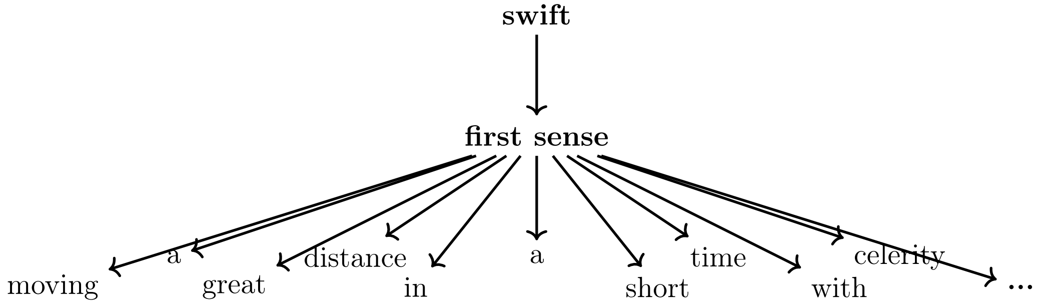


Figure 3.6: Network derived from the definition of *swift*

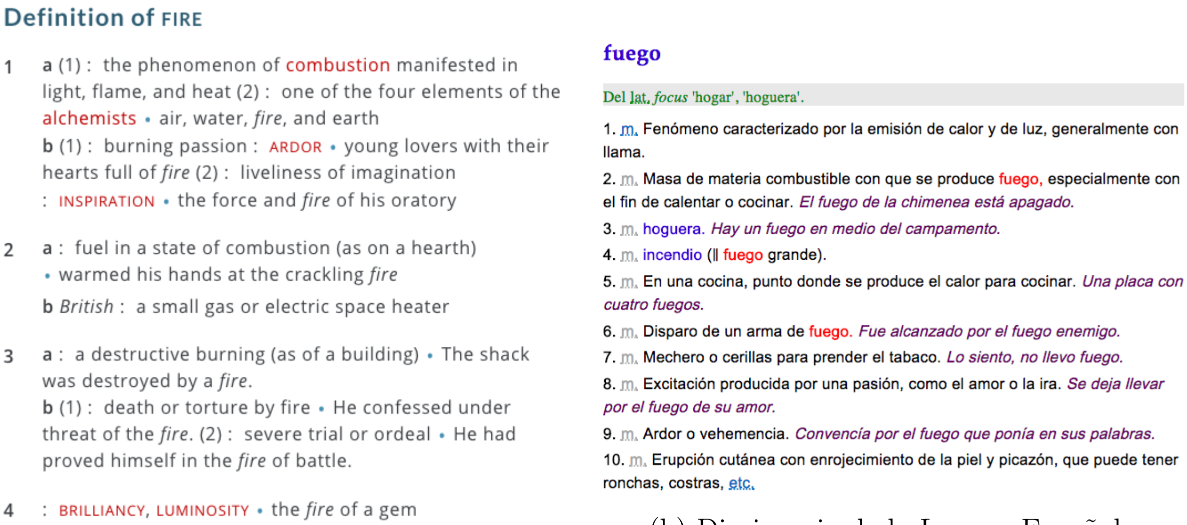
However, if we wanted to reconstruct the exact sentence of the sense, we could not. We do not know if *moving* is the first word in the sentence, nor if the word *short* is followed by the word *distance*. Since we do not know the order of the node we could reconstruct the sentence as "moving a *short* distance in a *long* time" or "moving a *long* distance in a *short* time". This situation also happens when a word is polysemous, where one entry word has more than one sense.

To be able to reconstruct the dictionary from the network, we need to preserve the order in the relationships entry-sense and sense-word. It is not necessary to preserve order in word-sense and word-entry relations since that information does not appear in the original dictionary.

Dictionaries are often a semi-structured sources of information. They have different structures depending the language and publisher, but patterns can be found in order to identify the elements to build the dictionary network. For example, we can find similarities on the position of the entry word or how senses are arranged, even in dictionaries of different languages (Figure 3.7). Entry words are at the top, senses are one per line, etc. This semi-structured nature of dictionaries with the assistance of the available natural language processing tools allow us to automatize the process of extracting the nodes and the edges to form the dictionary network.

Implementing the model

When extracting the word-sense relationship the following issue arises: How do we identify to which sense-node a word has to connect? The automatization of assigning a sense to each



Definition of FIRE

1 a (1) : the phenomenon of combustion manifested in light, flame, and heat (2) : one of the four elements of the alchemists • air, water, *fire*, and earth
b (1) : burning passion : ARDOR • young lovers with their hearts full of *fire* (2) : liveliness of imagination : INSPIRATION • the force and *fire* of his oratory

2 a : fuel in a state of combustion (as on a hearth) • warmed his hands at the crackling *fire*
b *British* : a small gas or electric space heater

3 a : a destructive burning (as of a building) • The shack was destroyed by a *fire*.
b (1) : death or torture by fire • He confessed under threat of the *fire*. (2) : severe trial or ordeal • He had proved himself in the *fire* of battle.

4 : BRILLIANCY, LUMINOSITY • the *fire* of a gem

(a) Merriam-Webster

fuego

Del lat. *focus* 'hogar', 'hoguera'.

1. m. Fenómeno caracterizado por la emisión de calor y de luz, generalmente con llama.
2. m. Masa de materia combustible con que se produce fuego, especialmente con el fin de calentar o cocinar. *El fuego de la chimenea está apagado.*
3. m. hoguera. *Hay un fuego en medio del campamento.*
4. m. incendio (‖ fuego grande).
5. m. En una cocina, punto donde se produce el calor para cocinar. *Una placa con cuatro fuegos.*
6. m. Disparo de un arma de fuego. *Fue alcanzado por el fuego enemigo.*
7. m. Mechero o cerillas para prender el tabaco. *Lo siento, no llevo fuego.*
8. m. Excitación producida por una pasión, como el amor o la ira. *Se deja llevar por el fuego de su amor.*
9. m. Ardor o vehemencia. *Convencía por el fuego que ponía en sus palabras.*
10. m. Erupción cutánea con enrojecimiento de la piel y picazón, que puede tener ronchas, costras, etc.

(b) Diccionario de la Lengua Española.

Figure 3.7: Definitions.

word in the sentences of a definition can become a difficult task. This problem could be addressed using methods and algorithms from the literature, since they have made a lot of improvement over the last years. However, there still are difficulties that do not allow us to automatize or keep at minimum the human supervision. One of the difficulties is that the source for the dictionary network and sense disambiguation techniques are different. It is hard to resolve the sense into the senses of the dictionary we are working. Also these tools and methods for sense disambiguation are often more developed and validated for Western languages, in particular English.

3.2 The minimal model

In an ideal dictionary network, all the grammatical, semantic, phonological, and historical information should be included as entities and relationships from which one can reconstruct the dictionary. On the other extreme, one can build a simple (naive) network without any typing on nodes and edges, that is, just based on words pointing to words represented in some standard form (e.g. lemmatized). At first, this approach might seem impractical since there is a lot of information missing. However, there is a compromise between these two approaches. A simple dictionary network does not have to deal with the issues presented in section 3.1.2 and facilitates the network analysis and the comparison with other fields at the cost of losing some subtle linguistic properties. Some researchers [21, 89, 58] have followed this simple approach in their works. They do not focus on the linguistic features of words but rather on the network properties.

3.3 Building the Minimal Model

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. 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, sings.* 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, conjunctions, interjections, pronouns (personal, demonstrative, possessive, etc.) and articles. They appear too often in any text, so they would add noise to the graph. Lemmatize each word occurring in the definitions. As stated by Crystal [6], a lemma is the item which occurs at the beginning of a dictionary entry; more generally referred to as the headword. It is an abstract representation, subsuming all the formal lexical variations which may apply. In other words, we transform nouns into their singular form; verbs into their infinitive form; adjectives into their male singular form. The idea, as in lexical networks [41], is to avoid to have multiple objects that only differ in the syntactic function but have the same meaning. We also remove any word that does not appear in the dictionary, *e.g.* prefixes and suffixes like *Ex-* and *-able*, as they also have a syntactic function.

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 after steps (1) and (2).

4. Build the Network. From the data in (3), construct a directed graph $G = (V, E)$, where the nodes are $V = \{w \mid (w, S) \in D\}$ and the edges $E = \{(w, w') \mid (w, S) \in D \text{ and } w' \in S\}$. 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).*

We applied the above methodology to the *The Online Plain Text English Dictionary*[1] (OPTED) and the *Diccionario de la Lengua Española*[2] (DLE, Spanish Language Dictionary). We chose OPTED because is a public, free-access, important and recognized dictionary, based on the Webster's Unabridged Dictionary. On the other hand, we chose DLE because it is the most authoritative dictionary of the Spanish language. The first edition of the DLE 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 DLE 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.

The networks obtained are available at https://github.com/hirohope/dictionary_networks/.

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

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

	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
DLE 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 3.1: Basic measures for networks. OPTED is an English dictionary network. DLE 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 can be found in [57].

3.4 Differences and distinctive features

There are manifold types of networks (biological, transport, communications, informational, social, etc.), each family having its characteristic properties. In this section, we study dictionary networks and compare them to other types of networks. We follow classic network analysis methodologies [81], focusing on component analysis, centrality, cliques and triadic configuration.

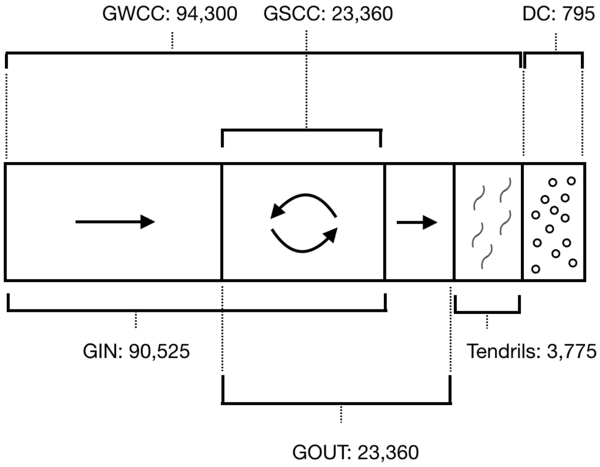
3.4.1 Basic Measures

Table 3.1 shows basic parameters for three different dictionary networks and three other networks built by humans.[3]

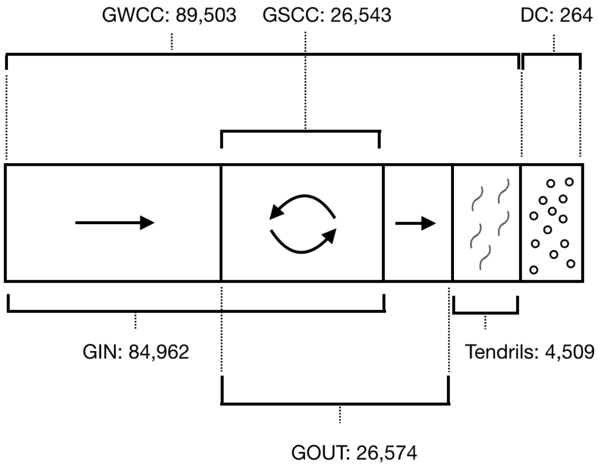
The three dictionaries have similar and inherent values for basic parameters (as compared to other types of networks), and their properties are similar to semantic networks. Steyvers and Tenenbaum [101] observed for the latter: "they have a small-world structure, characterized by sparse connectivity, short average path lengths between words, and strong local clustering." Additionally, dictionary networks show high resilience, meaning that removing some words will cause little disruption in the network, since with high probability there will be other good relations to supply the loss. Details follow.

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 $c2$ $(= \frac{1}{n}\sum_{\mathrm{i}} c_{\mathrm{i}}$ where $c_{\mathrm{i}} = \frac{\text{number of triangles connected to vertex i}}{\text{number of triples centered on vertex i}})$

[3] We use *igraph* http://igraph.org/ for network analysis and the *Stanford NLP* [61] for lemmatizing the words in the dictionary.



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



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

Figure 3.8: 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.

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.

Regarding network resilience (which correlates with high connectivity): The standard measure is vertex attack tolerance VAT [64], *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 $V - S$ and S is a non-empty subset. We determine that VAT is 0.245 for OPTED and 0.3 for DLE. Compared to other scale-free networks [64] (HOTNet 0.06, big barbell 0.08, star 0.11, C3 0.15, barbell 0.2, PLOD 0.25, wheel 1.0), dictionary networks show surprisingly high resilience. More on this in section 7.1.

3.4.2 Component Analysis

When describing the topology of networks, components are a relevant feature. They are defined as groups of nodes that are connected. In what follows, we divide the graph in the following main components:

- *Giant Weakly Connected Component* (GWCC): the biggest subgraph where all vertices are connected to each other by some path, ignoring the direction of edges.
- *Disconnected Components* (DC), that consists of separate small connected components not present in the GWCC.

- *Giant Strongly Connected Component* (GSCC): the biggest subgraph where every vertex is reachable from every other vertex. Usually the most relevant part of the network.
- *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).
- *Giant out-component* (GOUT): the set of nodes that have paths that starts in the GSCC. In other words, the set of words that are used to define those in GSCC.
- *The Bidirectional Component*: the subgraph of GSCC where for every edge (a, b) if the edge (b, a) also exists.
- *The Bidirectional Strongly Component*: the biggest subgraph of the Bidirectional Component where all vertices are connected to each other by some path.
- *Tendrils*: nodes that have no access to GSCC and are not reachable from it.

For the case of the OPTED and DLE dictionary networks, the relative size of each component is similar in both (see Figure 3.8). The Giant Strongly Connected Component corresponds to around 30% of the whole network. The Bidirectional Component stays around 17% of all the words. Finally, the Bidirectional Strongly Connected Component covers about 11% of the network.

3.4.3 Centrality Measures and Cliques

One of the main advantages of studying dictionaries as networks is the possibility of "discovering" global structures or properties of them (i.e. those depending on the whole network). A simple example of a property in a dictionary would be the frequency of words (that can be computed using the plain text). This is a standard measure of how important or "central" a word is, but there are others of particular relevance, such as classical centrality measures. Centrality measures are one of the most important notions in the study of complex networks. They are a quantitative measure that aims to reveal the importance of a node. Each measure has its own definition of "importance", trying to capture different aspects of influence of the node over the network or vice-versa. We chose four among the most popular: *degree* (most central nodes are those with higher number of adjacent nodes without considering the direction of edges); *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); and *PageRank* (essentially tells the probability of being at the given node after a random walk of arbitrary length starting from a arbitrary node in the graph). All of them are network properties impossible to compute with the dictionary as a simple text or list of definitions. We present them here to highlight their different behaviour and correlation. Figure 3.9 shows that these centrality measures in dictionary networks have weak correlation among them. We use Jaccard index over other correlation measures, such as Spearman's rank or Pearson's rho, to focus on the words as a set rather than the their order. To give a taste of word rankings using these measures, we list in Table 3.2 "top" words for different measures previously mentioned. Natural questions arise: Is there a notion of "most" relevant word in a dictionary network? What linguistic features are these different measures capturing?

Another productive topic of graph machinery applied to dictionaries is similarities among nodes. For example, one would expect that groups of words strongly tied in the network would be linguistically related. To illustrate the potentialities of the idea we explored cliques (a set of words such that for each pair u, v of words, $u \to v$ and $v \to u$, or in our representation, v occurs in u's definition and vice versa). We show that (bidirectional) cliques tell something about similarity of words. Big cliques are rare in general networks and also in dictionaries. In OPTED there is no K_6 (cliques are denoted K_n where n is the number of nodes), seven K_5 (shown in Figure 3.10), 174 K_4 and 2,641 K_3. In DLE we find no K_5, four K_4 and 243 K_3.

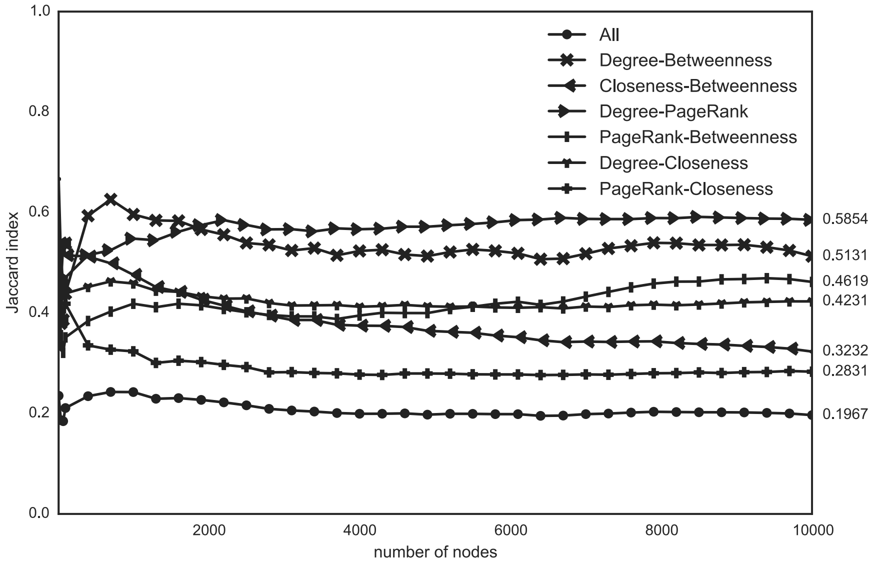


Figure 3.9: Common words of top rankings under different centralities, measured by Jaccard index ($\frac{|S_1 \cap S_2|}{|S_1| \cup |S_2|}$) for different number of nodes (0 to 10,000). For example the top ranked words for Degree and PageRank have 58.54% of their universe in common. All together they have 19.67% in common.

3.4.4 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 to the network core refers to a central and densely connected set. On 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 [23]: "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 [85] supports this intuition. Unfortunately, to the best of our knowledge, there is no categorization of the core structure of dictionary networks.

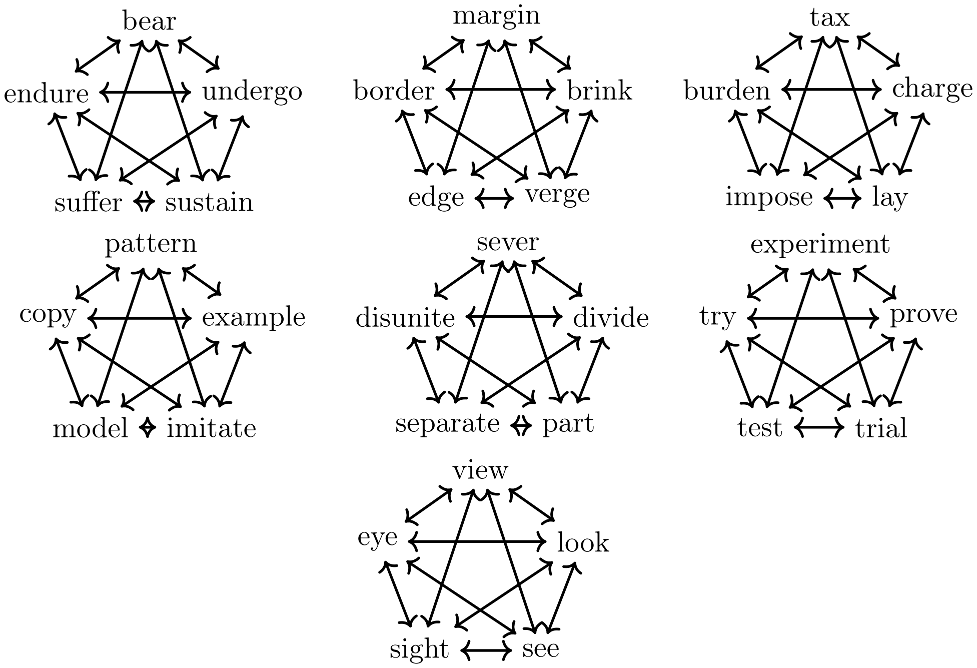


Figure 3.10: The seven cliques of size 5 in OPTED (there are no bigger cliques). Recall that if words u, v occur in a clique, v occurs in the definition of u and vice versa. Note their semantic closeness.

3.5 Word sense disambiguation of definitions

As we mentioned earlier in Section 3.1.2, an ideal dictionary network model incorporates different senses of a word, letting each vertex represent a sense rather than a word with all their senses merged as simpler models do. An automatic process is the main challenge when implementing a comprehensive model.

In this section, we study the use of word embeddings as an approach for unsupervised word sense disambiguation in Spanish dictionary definitions. This is an attempt to identify and evaluate the difficulties in the construction of a more complete dictionary network model.

Recently, in order to improve the performance of word sense disambiguation methods, researchers have tried to take advantage of the potential of word embeddings [19, 104, 94]. Word embeddings are vector representations for words or concepts in a low dimensional space, learned from the contexts in which words appear in large corpora of text. They capture regularities in language and contextual distributions. Supervised word sense disambiguation is based on the hypothesis that contextual information provides a good approximation to word meaning [73].

The main idea of our approach is to represent the sentence of the ambiguous word as a vector, as well as the sentences that define each sense of the ambiguous word. We select the sense whose vector is the more similar to the vector of the sentence in which the ambiguous word appears. Preliminary results fail to show insights to improve the dictionary network model or the model analysis. They show that word embeddings fail in providing substantial

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 DLE

#	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 3.2: Top words in OPTED and DLE 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; this is not anymore true for larger lists of words.

improvements compared to other unsupervised methods, such as Lesk and a uniformly random sense selection. This seems contradictory compared with the current state of the art on word sense disambiguation that has made incredible progress in the latest years. Therefore, we disregarded word sense disambiguation in the rest of the work. Nevertheless, we believe that this experience can be useful to design better experiments in the future.

3.5.1 Our disambiguation method

The main idea of our disambiguation method is to select the word's sense most similar to the sentence in which the word appears. The underlying assumption is that similar senses occur in similar contexts.

Let w be the word to be disambiguated and S the sentence in which w appears. We calculate the vector representation $v_{sentence}$ of the sentence S as the average of the vector representation of each word in the sentence:

$$v_{sentence} = \sum_{w_{\text{i}} \in S} \frac{V(w_{\text{i}})}{N}$$

where N is the number of words in S and $V(w_{\text{i}})$ is the vector representation of w_{i} in a pre-trained embeddings dataset. Likewise, we compute the vector representation $v_{sense_{\text{i}}}$ of each sense of the ambiguous word w. Then, we compute the cosine similarity between the sentence vector and each sense vector as follows:

$$sim(v_{sentence}, v_{sense_{\text{i}}}) = \frac{v_{sentence} \cdot v_{sense_{\text{i}}}}{||v_{sentence}||\ ||v_{sense_{\text{i}}}||}$$

We average the vectors to compute the vector representation of a sentence because we consider that all the words should contribute equally. Other methods favor other functions, such as fractional or exponential decay, considering just a window around the target word and the importance of a word is inversely proportional to the distance [49]. However, in a dictionary, all words support the definition providing context [76].

3.5.2 Experiment

We evaluated extrinsically the performance of our method in a sentiment analysis task. Experiments have shown that the use of word sense disambiguation has resulted in an improved sentiment analysis of micropost data (e.g. tweets) that outperforms systems built without incorporating word sense disambiguation [102]. We opted for a extrinsic evaluation because, to the best of our knowledge, there is no gold standard dataset for word sense disambiguation in Spanish that matches the senses in the Spanish Dictionary. Available Spanish datasets for word sense disambiguation map words to senses in translations of WordNet or have their own dictionaries. These datasets prevent us from evaluation directly on the Spanish Dictionary.

Task

We integrated word sense disambiguation into the sentiment analysis classification as follows: Given a sentence, replace some of the words in the sentence with the definition of their corresponding sense. For example, given the sentence *I went to the bank to deposit money*, if we replace the word *bank* with the sense used in the sentence we get the sentence *I went to the 'financial establishment that uses money deposited by customers for investment, pays it out when required, makes loans at interest, and exchanges currency' to deposit money.*

The assumption is that replacing a word with its corresponding sense will improve the polarity of the sentiment classification. Because of the ambiguity of natural language, many words have different meanings and different connotations, both positive and negative. For example, the word *childish* has a negative connotation implying an adult behaving immaturely in its sense *silly and immature*, but also it implies lively and energetic in its sense *like a child.* Also, specifying the meaning of words with several senses might help to polarize the sentence they appear. For example, the verb *cut* have more than 15 senses: to wound, to divide a pack of cards, to intersect, etc.

We test 3 word sense disambiguation methods:

1. **Our method**: Using the vector representation to choose the most similar sense.
2. **Lesk**: Given an ambiguous word and the context in which the word occurs, Lesk algorithm returns a sense with the highest number of overlapping words between the context sentence and different definitions from each sense.
3. **Random**: Selecting a uniformly a random sense.

And for each disambiguation method, we test 7 strategies to choose the words in the sentence to be replaced with their sense definition:

1. **All**: All the words in the sentence.
2. **Top-1**: Only the words in the sentence that have the greatest number of senses.
3. **Top-3**: Three words in the sentence that have the greatest number of senses.
4. **Bottom-1**: The word in the sentence with the least number of senses (at least two).
5. **Bottom-3**: Three words in the sentence that have the least number of senses (at least two senses).
6. **Random-1**: A uniformly-random word from the sentence.
7. **Random-3**: Three different uniformly-random words.

As classification method to determining the sentiment of a sentence, we use the Google Natural Language Processing Sentiment API[4]. This API, using a pre-trained model, classifies documents and sentences assigning a score r with $-1 \leq r \leq 1$. A score $0.25 < r \leq 1$ means positive. A score $-0.25 \leq r \leq 0.25$ means neutral. And a score $-1 \leq r < -0.25$ means negative.

[4] https://cloud.google.com/natural-language/

Method	Accuracy	F1-Score
Base	0.4975	0.5261
Our method	0.4118	0.4389
Lesk	0.4043	0.4316
Random	0.4101	0.4199

Table 3.3: Average accuracy and macro-averaged precision, recall and F1 measures for sentiment analysis classification in Spanish sentences.

Dataset

As benchmark for the experiment, we considered the corpora provided in the test set for the Task-1 of TASS 2018[5]. TASS is the workshop and shared task "Sentiment Analysis at SEPLN". It has been held under the umbrella of the International Conference of the Spanish Society for Natural Language Processing (SEPLN).

Task-1 focuses on the evaluation of polarity classification systems of tweets written in Spanish. We use this dataset because it resembles dictionary definitions. Both of them are short texts: Tweets have up to 240 characters while 98% of the Spanish Dictionary senses have less or equal to 240 characters. Also, TASS dataset contains tweets written in the Spanish language spoken in Spain, Peru and Costa Rica, a Spanish diversity also found in the Spanish Dictionary.

There are 2000 tweets in the dataset. Each of them is annotated with 3 different levels of intensity: Positive, Neutral, and Negative; –tweets with no annotation were removed–.

In our disambiguation method, we use pre-trained Spanish word embeddings computed using FastText [12] and the Spanish Billion Word Corpus [17]. These embeddings were provided by the Computer Science Department of the University of Chile[6].

3.5.3 Results

Table 3.3 shows the overall results of the experiment for all the replacement of words for their senses and word disambiguation methods. We can observe that the baseline –no replacement of any word for its sense– has better accuracy and F1 measure than our method, Lesk algorithm, and random. It is better to not make any replacement at all. Nevertheless, our main focus is not to improve the performance of sentiment analysis, but to determine if the use of word embeddings improve the disambiguation of words compared to other methods. In average, our method reports better accuracy and F1 measure than Lesk algorithm and Random. However, the differences are not significant. Our method reports an accuracy of 0.4118, that is an improvement of less than 1% compared to Lesk algorithm and Random. The same occurs with the F1 measure.

The results of the experiment divided by the different methods to replace words by their senses are shown in Table 3.4. We observe that our method outperforms Lesk algorithm

[5] http://www.sepln.org/workshops/tass/2018/

[6] https://github.com/dccuchile/spanish-word-embeddings

and Random in 4 replacement settings: all words in the sentence, the three words with the greatest number of senses, the word with the least number of senses, and the three words with the least number of senses. Then again, the accuracy improvement of our method is not significant. In the remaining replacement settings of the experiment – the word with the least number of senses, a random word, and three random words– the method with the best scores is Random.

All three methods do not show a strong inclination for the number of senses of the word being disambiguated. Replacing a word with several senses produces the same results when replacing words with few senses or even a random word.

The best scores –not taking into account the baseline– are obtained using the random disambiguation strategy. Replacing a random word of the sentence by one of its random selected sense, give us the best accuracy. And replacing the word with the greatest number of senses in the sentence, by one of its random selected sense, give us the best F1-Score.

We hypothesized that a possible reason for these results is that definitions of the dictionary are written in a specific form and style. For example, several words are defined as "a word pertaining or related to another word". The word *childish* ("infantil" in Spanish) is defined as "pertaining or related to infancy or children". Since we utilized pre-trained Spanish word embeddings, words such as pertaining or related might deteriorate the vector representation of senses. Incorporating the Spanish dictionary definition texts into the word embedding training might help to improve the results.

3.5.4 SemCor Control Experiment

Given the previous results, we made another experiment in order to discard that the cause of the low performance of our method is due to the data used for experimentation. We want to discard issues with the language and social media nature of the corpus.

In this experiment we disambiguate words from the SemCor dataset 3.0[7]. SemCor is a corpus manually annotated with senses from the English WordNet.

We disambiguated 7000 words using three methods: Our method, Lesk algorithm, and Random. For our method we used the pre-trained English word embeddings provided by FastText [8] [12] and WordNet 3.0 as the source for sense definitions.

Since, each method allow us to rank the senses of the word being disambiguated, we evaluated the top-1 and top-3 accuracy for each method. Our method ranks the senses by similarity, Lesk ranks them by the size of the intersection, and Random ranks them in uniformly random order.

Table 3.5 shows the top-1 and top-3 accuracy for each method. We observe that our method has a better top-1 accuracy than Lesk and Random. However, as in the previous experiment, the improvement is not significant. Additionally, Random has a better top-3 accuracy than both our methods and Lesk.

[7] http://web.eecs.umich.edu/~mihalcea/downloads.html#semcor

[8] https://fasttext.cc/docs/en/english-vectors.html

Replacement	Method	Accuracy	F-Score
	Base	**0.4975**	**0.5261**
all	Ours	0.3150	0.3208
	Lesk	0.2685	0.2404
	Random	0.2870	0.2776
top1	Ours	0.4475	0.4788
	Lesk	0.4580	0.4815
	Random	0.4590	**0.4901**
top3	Ours	0.4125	0.4443
	Lesk	0.3875	0.4080
	Random	0.4095	0.4367
bottom1	Ours	0.4530	0.4822
	Lesk	0.4460	0.4714
	Random	0.4460	0.4695
bottom3	Ours	0.4150	0.4443
	Lesk	0.4110	0.4310
	Random	0.4085	0.4332
random1	Ours	0.4450	0.4727
	Lesk	0.4600	0.4863
	Random	**0.4615**	0.4871
random3	Ours	0.3945	0.4295
	Lesk	0.3990	0.4205
	Random	0.3995	0.4274

Table 3.4: Measures for each replacement method of a word by its disambiguated sense in Spanish.

3.5.5 Discussion

We found that there is practically no difference between disambiguation using word embeddings (describing the context and the senses as the average vector representation and choosing the most similar) and choosing a random sense. Word embeddings has been exploited to improve several NLP tasks, including supervised word sense disambiguation systems, showing good results. This makes us believe that there are several aspects of the experiment that can be improved.

Our next step is to test if our system is failing to capture the semantic information of the context of the disambiguated word or is failing to represent the meaning of the senses given the writing style of dictionaries.

	Method	Accuracy
Top-1	Our method	0.233
	Lesk	0.218
	Random	0.229
Top-3	Our method	0.544
	Lesk	0.517
	Random	0.553

Table 3.5: Top-1 and top-3 accuracy for word sense disambiguation of English words.

Chapter 4

Data Collection of Dictionaries

In this chapter, we describe the process of collecting, cleaning and preparing the data of the dictionaries used throughout this work. We detail the source of each dictionary, its format, and other particulars needed to apply the dictionary network methodology. Even though it is a pure technical work, we thought that this work could be useful for other researchers in order to pre-process the dictionary data for numerous other dictionaries available, especially, the historical ones.

The core of our work was the Spanish Dictionary, and we used two English ones to have insights of similar phenomena in other languages. The networks data obtained from these dictionaries are available at https://github.com/hirohope/dictionary_networks/.

4.1 The Online Plain Text English Dictionary

The Online Plain Text English Dictionary[1] (OPTED) is a public domain English dictionary based on the 1913 US Webster's Unabridged Dictionary. It comes in simple plain text format, divided into 26 files, one for each letter of the English alphabet. The dictionary defines more than 95,000 English words. They are divided into approximately 58,750 nouns, 12,270 verbs, and 24,100 adjectives and adverbs.

Each line in the dictionary corresponds to the definition of a word-sense. It starts with the word-form to be defined. It is followed by some abbreviations in parenthesis providing some basic linguistic information about the word. For example, *n.* if the word is a noun, *adj.* if the word is an adjective, *v.t.* if the word is a transitive verb, among others. After that comes the proper definition of the word-sense. Some examples are shown in Figure 4.1.

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

Draw *(v. t.)* To form a sketch or a picture of; to represent by a picture; to delineate.
Draw *(v. t.)* To pull from a sheath, as a sword.
July *(n.)* The seventh month of the year, containing thirty-one days.

Figure 4.1: Excerpt of three word-sense definition of the OPTED.

4.2 English WordNet

WordNet[2] is an online lexical database developed at Princeton University's Cognitive Science Laboratory over many years by linguists, lexicographers, students, and software engineers.[72, 71, 38]. WordNet has four components that are of interest for our work:

- *Lemmas*: words forms represented in their familiar orthography.
- *Synsets*: instances grouping synonymous words that express the same concept. Some lemmas have only one synset and some have several.
- *Definitions*: each synset contains a brief definition in the classical manner of dictionaries.

Based on these three components we can obtain the definition of all word-meanings of each lemma, in the same manner as in a classical dictionary. We utilized the version of WordNet provided by the Natural Language Toolkit [3] which is based on WordNet 3.0. This version contains about 155,000 lemmas. They are divided approximately into 117,000 nouns, 11,500 verbs, 22,100 adjectives, and 4601 adverbs.

4.3 The Spanish Language Dictionary *(Diccionario de la Lengua Española)*

The Diccionario de la Lengua Española (DLE) is the most authoritative dictionary of the Spanish language. The DLE is a dictionary issued periodically since 1780 by the Spanish Royal Academy [4]. New versions of the dictionary present updated lexicon and linguistic and editorial reorganizations[5].

We work with three versions of the DLE, the editions from 1925, 1956, and 2014. The 2014 version of the dictionary defines more than 93,000 Spanish words. These words are distributed in approximately 54,800 nouns, 12,000 verbs, and 22,954 adjectives and adverbs. The 1956 version defines about 76,000 words and the 1925 version defines about 67,000 words.

We collected the physical copies of these editions and digitized them using a optical character recognition. Until the twenty-first edition in 1992, the DLE was published exclusively on paper. Since then, the RAE provides a free electronic consultation version of the DLE on

[2] https://wordnet.princeton.edu/
[3] https://www.nltk.org/
[4] https://www.rae.es/
[5] http://www.rae.es/diccionario-de-la-lengua-espanola/presentacion

its website[6] that allows to search a specific word and obtain the definition of its word-senses. This version is insufficient to provide the data for our work. The system restricts too many queries and they do not maintain a publicly available complete list of words that are defined in the dictionary.

Automatic digitization processes are not flawless. Two kind of errors were produced: misrecognition or characters and division of words. Misrecognition errors are when a digitized word differs from the actual word in one or more characters. For example, the character **"ó"** was recognized in some cases as the character **"6"**. A list of the frequent misrecognition errors is presented at the end of the section. These kind of errors were easily identified since they produce non-existent words. We use the Spanish Billion Words Corpus [17] to check if a word is real or non-existent. Division of words occurs when the automatic digitization processes misinterpret the distance between two characters as a blank space, dividing the words into two or more words. For example, **"P r e s t i g i o"** instead of **"Prestigio"** or **"Preocup ante"** instead of **"Preocupante"**. This kind of error were more problematic, since they sometimes divide the word into two or more non-existent and/or existent words. In the previous example, the word **"Preocup"** is a non-existent word and the word **"ante"** is an existent word. In order to repair these errors, we iterated over every non-existent word. We checked if the concatenation of some words around the non-existent words allows us to replace it with a real word. We used a window of size 5 around the non-existent word. In the case that the concatenation generated more than one real word, we choose the most frequent word with respect to the Spanish Billion Words Corpus.

Frequent misrecognition errors:

- '**o**' instead of '**.**'.
  E.g.: *'Persona que controlao'* instead of *'Persona que controla.'*
  E.g.: *'De dos sílabaso'* instead of *'De dos sílabas.'*

- '**.**' instead of '**,**'.
  E.g.: *'Producir algo. darle el primer ser.'* instead of *'Producir algo, darle el primer ser.'*
  E.g.: *'limpia. fija y da esplendor'* instead of *'limpia, fija y da esplendor'*

- '**,**' instead of '**.**'.
  E.g.: *'Astron, Conjunto enorme de estrellas.'* instead of *'Astron. Conjunto enorme de estrellas.'*
  E.g.: *'tr, Hacer molduras en algo,'* instead of *'tr. Hacer molduras en algo.'*

- '**í**' instead of '**f**'.
  E.g.: *'reíerencia'* instead of *'referencia'*
  E.g.: *'transíerencia de calor'* instead of *'transferencia de calor'*

- '**I**' instead of '**l**'.
  E.g.: *'ecoIógico'* instead of *'ecológico'*

[6] https://dle.rae.es/

E.g.: *‘respuesta Iógica’* instead of *‘respuesta lógica’*

- ‘**6**’ instead of ‘**ó**’.
  E.g.: *‘canci6n’* instead of *‘canción’*
  E.g.: *‘xil6fono’* instead of *‘xilófono’*

- ‘**l**’ instead of ‘**í**’.
  E.g.: *‘magnlfico’* instead of *‘magnífico’*
  E.g.: *‘pellcano’* instead of *‘pelícano’*

- ‘**ft**’ instead of ‘**ñ**’.
  E.g.: *‘maftana’* instead of *mañana’*
  E.g.: *‘rebafto’* instead of *‘rebaño’*

- ‘**o**’ instead of ‘**c**’.
  E.g.: *‘oombate’* instead of *‘combate’*
  E.g.: *‘oalendario’* instead of *‘calendario’*

- ‘**b**’ instead of ‘**n**’.
  E.g.: *‘nicbo’* instead of *‘nicho’*
  E.g.: *‘cbocolate’* instead of *‘chocolate’*

4.4 Thoughts and observations

It is better to work with a digital dictionary (machine-readable) rather than a physical copy. A physical copy required to be digitized prior to any processing and the digitization process has several downsides that need to be handled. As we mention before, automatic digitization processes are not flawless. There will be errors that will lead to loss of information. A proper digitization of a dictionary is not a simple step in a project, but a project itself. A useful resource to find digitization errors is a corpus of the language being processed. Digitization errors often lead to the formation of non-existent words and corpora provides a ground truth to check the existence of a word. Another useful tool is a part-of-speech tagger. In rare cases, a word will be mistaken as another existent word. Using a tagger, we can check if the sentence makes sense. Although digital dictionaries have fewer complications, physical copies have advantages. Physical copies allows us to work with a greater number of datasets. A valuable project, very helpful to the community, would be to have a public repository of proper digitized dictionaries. It would facilitate and encourage the study and analyses of dictionaries as networks and sources of lexical knowledge.

Chapter 5

Ogden's Basic English

Ogden's Basic English is an English-based controlled language created by Charles Kay Ogden in 1930 [85]. It is a simplified subset of the English language, according to Ogden, "a system in which everything may be said for all the purposes of everyday existence". This subset consists of 850 words[1]. Ogden stated the following:

> 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. But the chief reason why it is possible to do so much with the limited word-list is because Basic has been able so completely to do without 'verbs.'

Clearly the main arguments for the choice of the words are linguistic.

In this chapter, we study the relation of this group of words with the structure of the dictionary network. As we previously saw, the structure of dictionary networks is an important feature. If they carry any linguistic knowledge, it has to be reflected in its structure. We focus on two structural features: first, the concept of relevance of a word or centrality in the case of a network, and second, the notion of connectivity. 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 and, primarily, on how dictionary networks handle the basic information of language. For our experiments we use

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

the The Online Plain Text English Dictionary (OPTED). We summarize here the principal findings regarding Basic English and its core. A more comprehensive discussion and analysis is published in our work [43].

The Online Plain Text English Dictionary contains 803 words of the 850 of the Ogden's vocabulary. The 47 Ogden words not present in the dictionary are words that were removed from the dictionary in the cleaning phase, for example, conjunctions and adverbs. These words are the following:

- a
- about
- across
- after
- against
- among
- and
- any
- as
- at
- before
- between
- but
- by
- down
- every
- for
- from
- he
- here
- I
- in
- like
- married
- mine
- near
- of
- off
- on
- or
- other
- over
- than
- that
- the
- there
- this
- through
- to
- together
- under
- up
- waiting
- where
- who
- with
- you

These words do have a definition in the dictionary. However, in the construction process of the dictionary network, we remove this kind of words (See 3.3). As we observe, this words are mainly prepositions, conjunctions and pronouns.

5.1 How central are Ogden's words?

Naively one would think that Ogden's words should have a good correlation with "central" nodes in the dictionary network. We investigated this with the classic centrality measures presented in section 3.4.3. We took the best k nodes for each centrality measure and for every $0 < k \leq 803$, and checked how many of Ogden's words are in each of these sets. The results (shown in Figure 5.1) indicate that none of the centrality measures does a good job capturing Ogden's Basic English.

The best performance is obtained by degree centrality that captures almost 48% of Ogden. Degree centrality is strongly correlated with the frequency of occurrence of words. Ogden stated explicitly that "what a word will do for us has little relation to the number of times it is used in newspapers and business letters". Even tough frequency in corpus does not have to be related to frequency in definitions, the result of this measure leaves a mixed feeling. On the other hand, PageRank has the worst performance, capturing only 38.6%. This is rather surprising as PageRank is one of the most popular centrality measures today for text on the Web, used in multiple areas like ranking web pages, sense disambiguation [68], keywords and sentences from text [67], among others.

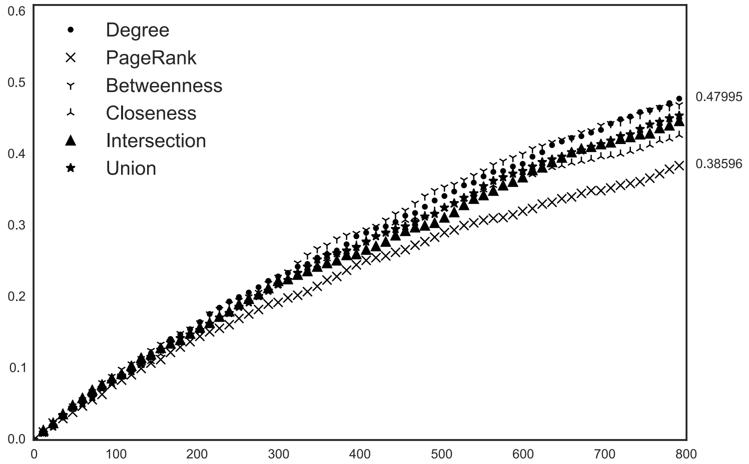


Figure 5.1: Ogden's Basic English words in top 800 words using different centrality measures. X-axis indicates k top-words and Y-axis, the percentage of Ogden's words in that set. Centralities by themselves are not a good method to capture the notion of importance that Ogden's Basic English represents.

Group Centrality. Refining the idea, one could hypothesize that the problem is with *individual* centrality since the meaning of words is essentially a network property and not an individual one. Ogden chose those words to work together, so we believe that analyzing the centrality of a set of words rather than individually, may lead us to identify if one of the centrality notions is more present in the Ogden's set. There is an extended notion of centrality, called *group centrality* [34], that captures "centrality" of groups, not individuals. In order to measure the centrality of a group of nodes, we can imagine that we compress the group of nodes into a single one, and then measure the centrality as always.

We experimented in this direction with groups of Ogden's words. We measured four group centrality measures based on the classic notions: group degree, group PageRank, group closeness, and group betweenness. As for the groups, we extracted from Ogden's set the top third and bottom third words according to PageRank. We chose PageRank since it seems the most promising to capture word senses [4, 115]. We extracted these two groups in an attempt to capture differences in the set and to determine if there were some evidence that there are some essential words in Ogden's set. As comparison and baseline, we used the most frequent words and a random selection from the OPTED dictionary, which is used in the construction of the dictionary network. For the most frequent words, we took the 803 words (size of Ogden) and we also divided them into the top third and bottom third. For the random words, we just took 268 words (a third of Ogden) selected uniformly among all the words.

The results of this experiment can be seen in Table 5.1.

For each of them we tested the four group centrality measures. Table 5.1 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 suggests that

	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 5.1: 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.

this group covers an ample part of the network, i.e. these words are "spread" around the network. This idea is slightly supported by the numbers given by 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 as for the baseline to compare to Ogden's top and bottom third we could not compute *real* group centralities due to lack of good algorithms and libraries (the problem is known to be NP complete [44]). By *real* group centrality we mean the group that has the highest value. That is different from how we did it in the experiment, taking the nodes with the highest value and then computing the group centrality of that group.

These experiments suggest that centrality measures inspired basically on social networks reality cannot be directly applied to lexicon. This points to the very question if there are reasonable notions of centrality in this world.

5.2 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 5.2. 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 only Ogden word not present in the SCC is the word "*tomorrow*", as this word does not appear in any definition in the

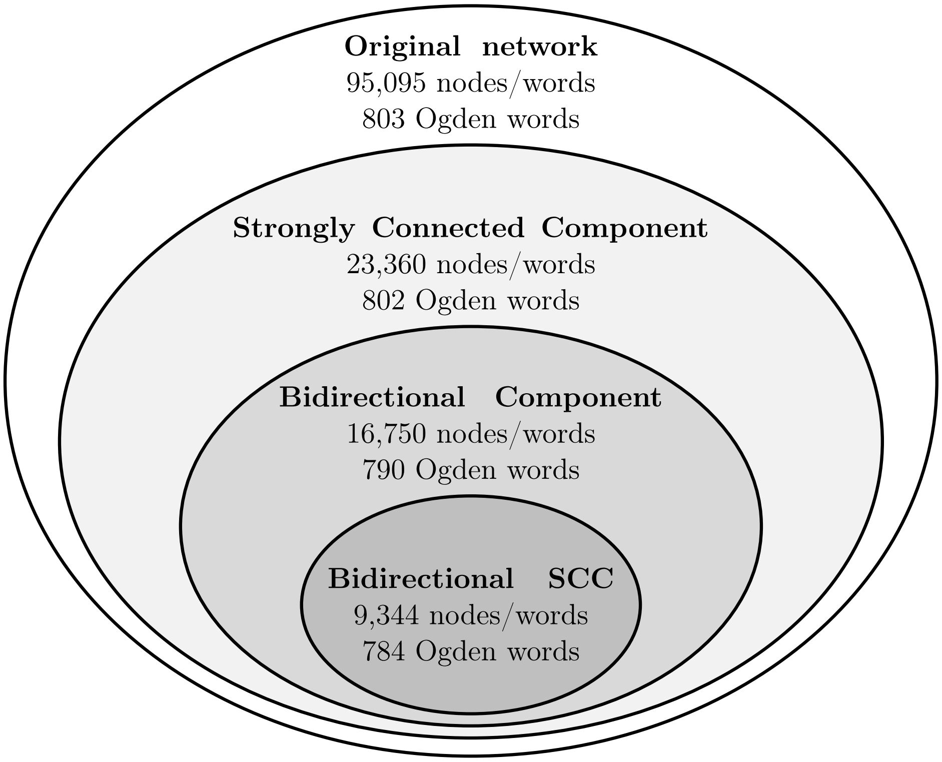


Figure 5.2: Connectivity analysis of components of OPTED network: The complete graph, Strongly Connected Component (words that recursively define themselves), Bidirectional Component (words that mutually need each other in order to be defined), and Bidirectional SCC. In the latter component only 3% of Ogden's words are lost), showing that Ogden's words strongly need each other.

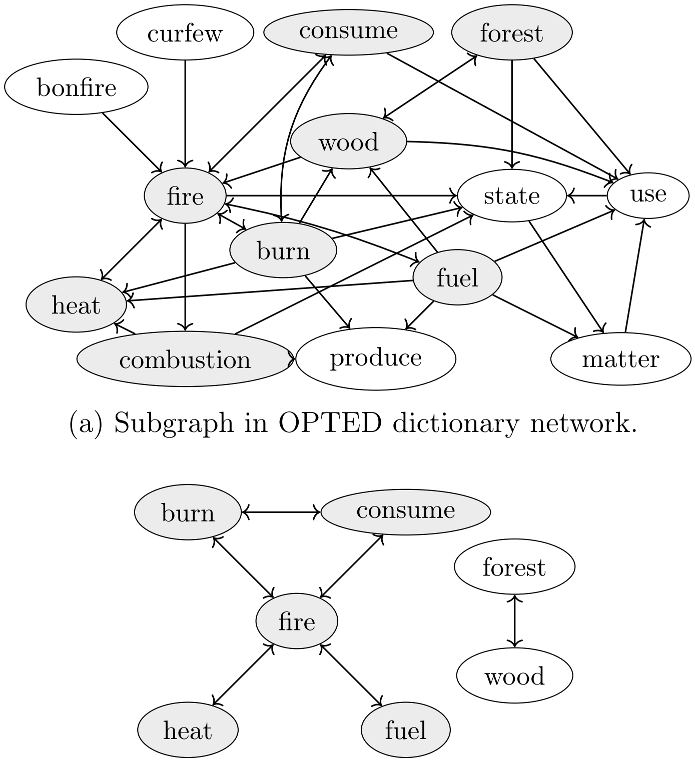
dictionary. The discarded words (approx. 3/4 of the total) are those that either are terminal (not used to define other words) or n-th level terminals (and terminals after eliminating the terminals and so on).

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. [89] 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.[58] 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



(a) Subgraph in OPTED dictionary network.

(b) Bidirectional Component of the subgraph (a)

Figure 5.3: Subgraph in OPTED dictionary network and its Bidirectional Component. Gray nodes in (a) and (b) belongs to the SCC of the respective graphs.

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.

Chapter 6

Triadic configuration

Earlier we showed that dictionary networks have a strong local clustering coefficient. In terms of network topology this means a high presence of triangles in the network. For example, two words having a common (non frequent) word in their definitions are likely to be related. This is a simple example of a triadic (triangle) configuration.

Triadic configurations are fundamental to many social structural processes. They provide the basis for a variety of social network theories and methodologies [36]. A triad is a subgraph of three nodes and the arcs between them. There are $\binom{n}{3} = \frac{n(n-1)(n-2)}{6}$ triads in a directed network with n the number of vertices in the network and 16 different isomorphic classes of triads (Figure 6.1).

We analyze the triadic configuration of dictionary networks and compare it with some social networks. Among the dictionary networks we analyze are the ones derived from WordNet, *The Online Plain Text English Dictionary* (OPTED) and 3 editions of the *Diccionario de la Lengua Española* (2014, 1956, and 1925). As for social networks we use 7 networks from the The Koblenz Network Collection [54]: the network of publications in the arXiv's High Energy Physics – Theory section, the Enron email network, the trust network from the online social network Epinions, the communication network of the Linux kernel mailing list, the reply network of technology website Slashdot, the results network of chess games from a Kaggle dataset, and the flight network collected by the OpenFlights.org project.

In Figure 6.2, we show the triadic configuration of these networks. We considered only connected triads (i.e. did not consider triads 003, 012 and 102 that have isolated vertices) because the low density of dictionary networks produces high amounts of those triads distorting the results.

We can observe that the most frequent triad in dictionary networks is the triad 021U (Figure 6.2a). This means that two unrelated words use the same third word in their definitions. It is a coherent result since, as we pointed earlier, there are a small group of high frequency words and more than two thirds of the vertices do not belong to the strong component of the dictionary network.

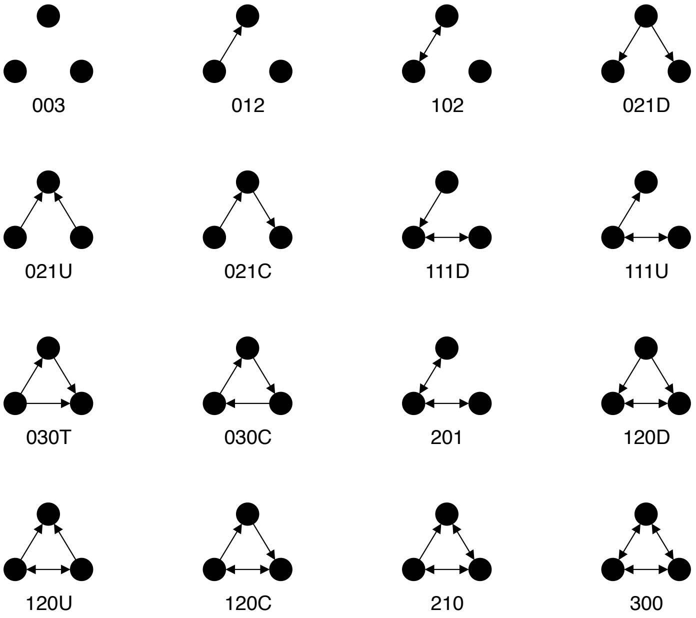
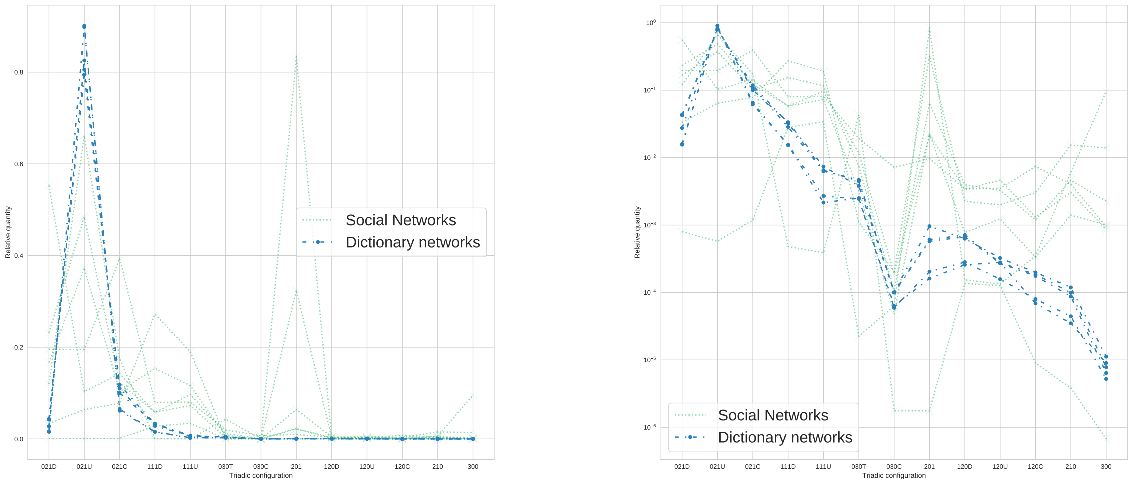


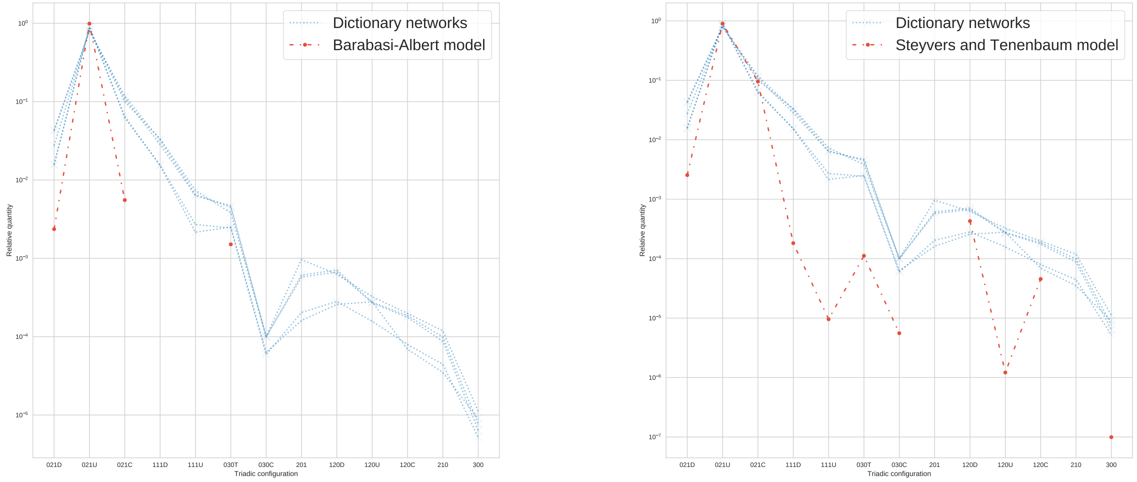
Figure 6.1: All 16 possible triads in a directed network.



(a) Triads census

(b) Triads census in logarithmic scale

Figure 6.2: Comparison of the patterns of triads between dictionary and social networks



(a) Barabási–Albert Model

(b) Steyvers and Tenenbaum Model

Figure 6.3: Triadic configuration of a generated network with directed Barabási–Albert model ($N = 23000$, $m = 12$) and the generated network with Steyvers and Tenenbaum's model ($N = 23000$, $m = 24$, $\alpha = 0.97$) compared to dictionary networks.

In order to avoid the noise produced by the most frequent triads, in Figure 6.2b we present the frequency of triads in a logarithmic scale. Note that dictionary networks follow a similar distribution among their triads, having their peak at the 021U triad. Then, the frequency starts to decrease as the triads become more connected. There is a drastic fall at the 030C triad (a complete cycle) and a small rise in the frequency as the triads with mutual links (in both directions) appear. The triad 300 is the least frequent in dictionaries, as it would mean that all three words use each other in their respective definitions, a fact that rarely occurs.

The triadic configuration of social networks presents different patterns. They have the same fall in the 030C triad (a complete cycle) as dictionary networks and a rise in the 201 triad (a mutual line), but they have strongly different patterns for the other triads as can be seen in Figure 6.2.

6.1 A model to generate random dictionary networks

There is a wide variety of models available for generating networks, but none of them seem to work for modeling triads occurring in dictionaries. Regarding their power law structure, the closest models to generate a dictionary structure are the directed version of Barabási–Albert model [7] and the model for semantic networks proposed by Steyvers and Tenenbaum in 2015 [101]. but they fail in almost all other facets, particularly in the triad structure. In this section we present a model for dictionaries.

Let us first review the problems of the most popular models. In the directed version of the Barabási–Albert model (recall that we modeled dictionaries as directed networks), nodes and edges are being added as follow: Each new edge starts at a new node. New nodes can only point to older nodes already added to the network. This prohibit the formation of cycles and pairs of nodes pointing to each other, being only possible to generate the triads 021U, 021D, 021C, and 030 (not taking into account the trivial triads 003 and 012). This model can generate only four types of triads: 021D, 021U, 021C, and 030T. Only if we consider a fully connected network to begin the construction, pairs of nodes pointing to each other can be generated. However, they will be the only ones. Thus the Barabási–Albert model for generating networks gives us a triadic configuration far from a configuration from a dictionary network (Figure 6.3a).

Steyvers and Tenenbaum model for generating networks [101] follows the same principles as the Barabási–Albert model, but there is a probability that each new edge may point in one of two possible directions, toward the existing node or toward the new node. The model allows the creation of cycles among the nodes, but, as the previous model, it also prohibits the creation of pairs of nodes pointing to each other. Thus, triads 201 and 210 cannot be generated. (Figure 6.3b). As happened with the directed version of the Barabási–Albert model, the exception for pairs of nodes pointing each other are the initial edges of the fully connected network. Thus these models do not generate the structure of dictionary networks. In the following, we present a model inspired by these ideas that models dictionary networks.

6.1.1 The proposed model

The model we will propose aims to replicate three traits of dictionaries. The first one is that dictionary networks follow a power law distribution. A power law distribution is achieved using preferential attachment (favoring connections to existing nodes with more connections). This is essentially step 2.

Second, when a new word is incorporated into the dictionary, this new word uses other words in its definition. However, it may happen that this new word is immediately used in the definition of an existing word or that the new word and an existing word use each other in their definitions. In network's terms, the connections of the new word may have different directions Using the probabilities p_f, p_b, and p_m, we can model this characteristic. Section 7.2 describes these circumstances in detail. These are the subtleties of step 2.

The third trait involves the ability of definitions to change over time (step 3). It is not difficult to see that this is the essential novelty of the model as compared to classical social networks models. If we only consider preferential attachment and the direction of the new connections, the relation between existing words are fixed. However, we observed that definitions change in time (Section 7.2). We modeled this behavior, adding uniformly random links between existing nodes.

Lexicographers will have no difficulty in recognizing that this process models the construction of new versions of a dictionary: In step 2 new words are added, with their corresponding definitions and also sometimes used in other definitions; additionally, and this is the crucial issue that allows to model the triads closed to those of dictionaries, in step 3 some old definitions are updated.

The model we propose for generating dictionary networks works as follows:

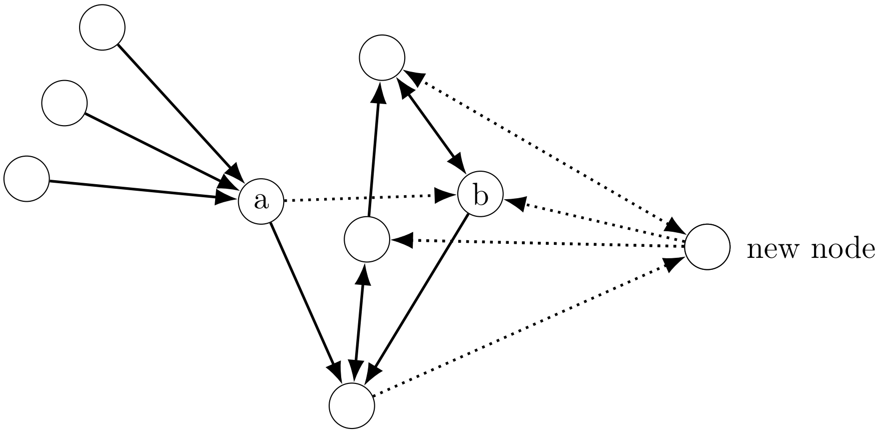


Figure 6.4: Snapshot of the model adding a new node to the network. Four edges are added (dotted lines): two pointing forward, one pointing backwards and one in both directions. Additionally, a random edges from node a to node b is added.

1. The generated network starts with a small fully connected network of m nodes.
2. At each step, a new node is added to the network. For each such new node, m existing nodes (same m as step 1) are chosen and connected to the new node. They are chosen with a probability proportional to the number of edges connected to them. Formally, each existing node v_i has the probability p_{v_i} to be connected to the new node, with

   $$p_{v_\mathrm{i}} = \frac{\mathrm{d}_{v_\mathrm{i}}}{\sum_j \mathrm{d}_{v_j}}$$

   where $\mathrm{d}_{v_\mathrm{i}} = \mathrm{d}^{\mathrm{in}}_{v_\mathrm{i}} + \mathrm{d}^{out}_{v_\mathrm{i}}$ correspond to the sum of the in- and out-degree of the node v_i at that step.

   The connection between the new node and the m_1 chosen existing nodes can occur in three differently ways:

   (a) with a probability p_f the edge will point from the new node to the existing node, i.e. the edge $(v_{new}, v_{existing})$ will be added to the network.

   (b) with a probability p_b the edge will point from the existing node to the new node, i.e. the edge $(v_{existing}, v_{new})$ will be added to the network.

   (c) with a probability p_m the new edge will point in both directions, i.e., the edges $(v_{existing}, v_{new})$ and $(v_{new}, v_{existing})$ will be added to the network. Naturally, it must be satisfied that $p_f + p_b + p_m = 1$.
3. Additionally, at each step, after the connections of the new node are created, we add m_2 new edges with a source and a target chosen uniformly between all the nodes in the network (including the new node).

In Figure 6.5 we can observe the triadic configuration of a network generated by our proposed model and the configuration of dictionary networks. We chose the parameters N, m_1, and m_2 to match the average number of nodes and edges in the strongly connected component of the dictionary networks analyzed in this thesis: Diccionario de la Lengua Española (Spanish), Online Plain Text English Dictionary, and WordNet (English). As for the probabilities p_f, $p_b =$, p_m, we observed on different editions of the Spanish dictionary that they had to follow these rough rules: $p_f > 0.9$ and $p_f >> p_b > p_m$. Using these rules, we manually tuned the values to reach the best fit.

The relative values of each triad follow a similar distribution to the distribution of dictionary networks, particularly the key 021U triad. They have peaks and slopes in the same triads. Triad 120U is the triad that displays the strongest difference. The proportion of 120U triads in the network generated with the proposed model is less than the proportion of dictionary networks. That this simple generation model generates a distribution so similar to dictionaries, except in one triad, tells us that essential features of a dictionary network structure are captured. It would be interesting to discover how to improve the model (e.g. regarding the 120U triad and other traits).

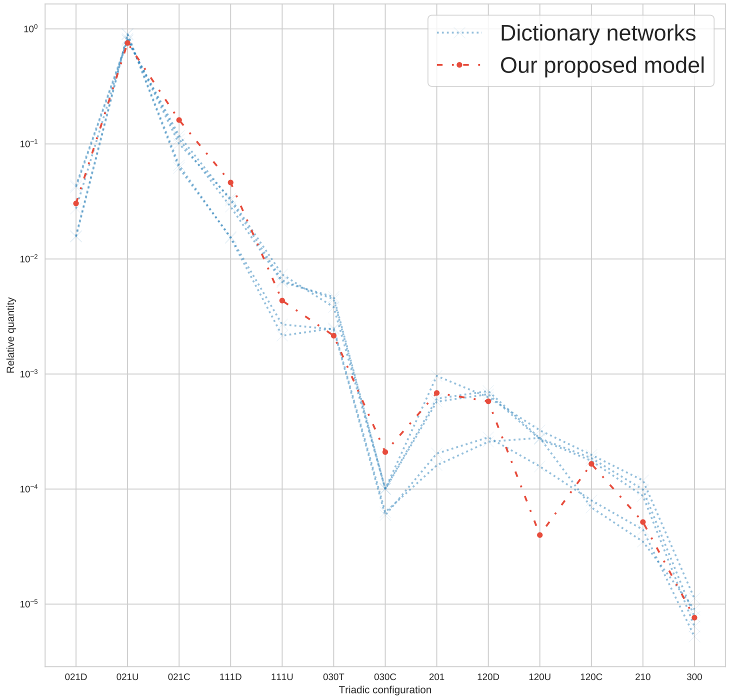


Figure 6.5: Triadic configuration of a generated network with our model compared to dictionary and social networks. $N = 23000$, $m_1 = 24$, $m_2 = 10$, $p_f = 0.93$, $p_b = 0.05$, $p_m = 0.02$.

6.1.2 A note on preliminary models we develop

In the process of developing the model to generate random dictionary networks, we constructed a couple of models. Although, they do not generate dictionary networks as well as our proposed model, they gave us insights and understanding to develop the proposed model. We believe they might be useful to document them here. We describe two of them.

Preferential Attachment plus random links.

As we mentioned, the main flaw in the preferential attachment model is that new nodes can only point to older nodes already added to the network, prohibiting the formation of cycles and pairs of nodes pointing to each other. A simple solution is to add m_2 directed links between random nodes at each step. This change allows the generation of cycles and pairs of nodes pointing to each other. In Figure 6.6 we can observe that this model generates all the types of triads. However, the distribution of the triads of this model does not quite match the distribution of dictionary networks. The frequency of the most frequent triads in dictionary networks are similar. However, as we move into less frequent triads, the distribution of the model becomes more irregular and underestimated. Adding the random directed links between nodes helped bringing the model closer to the distribution of dictionary networks.

Steyvers and Tenenbaum with bidirectional links.

As we mentioned the model proposed by Steyvers and Tenenbaum [101] cannot generate pairs of nodes pointing to each other. We extended the model allowing the new edges to point in both directions. Now each new edge has a probability p_m of pointing in both directions, in addition to the probability p_f of pointing from the new node to an older node and the probability p_b of point from an older node to the new node. Despite the fact that now the triads of the generated network can have cycles or be more dense, these kinds of triads are vastly under-represented. We can observe this in the triads in the right side of Figure 6.7. The probability of a new edge pointing in both directions in the version of this model closer to dictionary networks is $p_m = 0.01$. This low probability causes, for example, that triad 210 –two bidirectional links and one directional link– is almost not generated and triad 300 –a clique; three bidirectional links- to not be generated at all.

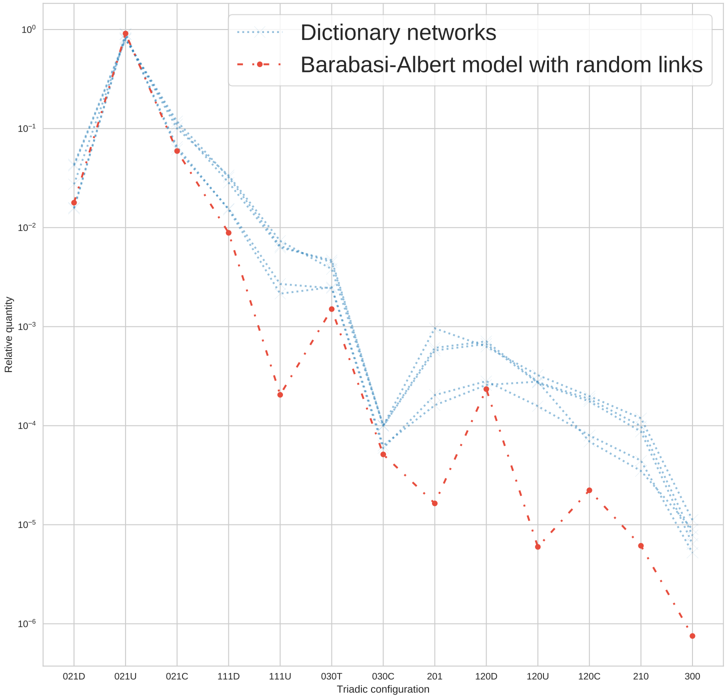


Figure 6.6: Triadic configuration of a generated network with an preliminary model (preferential attachment with random links) compared to dictionary and social networks. $N = 23000$, $m_1 = 24$, $m_2 = 6$

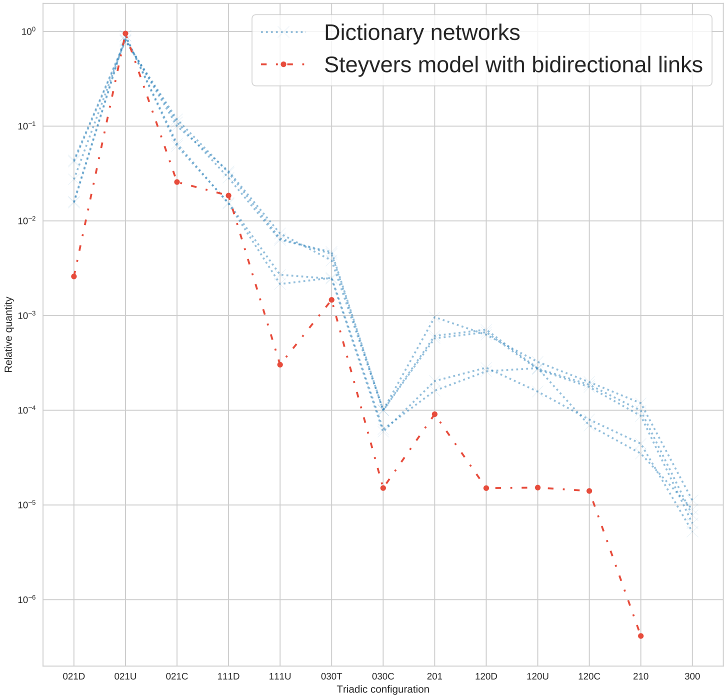


Figure 6.7: Triadic configuration of a generated network with an preliminary model (Steyvers and Tenenbaum with bidirectional links) compared to dictionary and social networks. $N = 23000$, $m = 12$, $p_f = 0.98$, $p_b = 0.01$, $p_m = 0.01$

Chapter 7

Evolution of *Diccionario de la Lengua Española*

In this section we study the network of The Spanish Language dictionary (*Diccionario de la Lengua Española*, DLE) and show that its basic network structure has remained stable and resilient over the years. We analyze three editions of the DLE: the 15th (published in 1925), the 18th (1956), and the current 23rd edition (2014). According to the Royal Spanish Academy, the 1925 and 2014 editions are especially significant. The former (1925) incorporated attention to different Spanish-speaking territories besides Spain, and describes simpler definitions. The latter (2014), the most recent version, besides updating the lexicon, modifies its structure to facilitate search, and incorporate other features, e.g. to show variations of entries and a consistent treatment in their male and female forms. To have an intermediate reference point, with a logarithmic interval between the extremes (30 and 60 years), we employed the 18th edition (1956). We used printed versions (no digital versions exist for 1925 and 1956). The details on how we processed these versions are in Section 4.

Despite its structural stability, there are changes in the successive versions of the DLE: new entries are incorporated, some entries are removed and some definitions are enriched or modified. In this section, we focus on these changes in the dictionary.

7.1 A stable and resilient structure

A first snapshot of the evolution of dictionary networks is given by basic network measures (See Table 7.1) [81]. the number of nodes (n) indicates the number of words in the dictionary. The dictionary grows about 15% every 30 years in this period. Edges (m) do not grow at the same rate, and the current dictionary has on average fewer edges per node (z) than previous years (meaning shorter definitions on average). Despite the changes in the number of nodes and edges, the average distance between entries (l) is not affected, staying around 4. The parameter α, the exponent of the degree distribution function ($p_k \sim k^{-\alpha}$), also remains almost unaffected along the years with value $\alpha \approx 2.6$. The clustering coefficients along the years are also very similar, both global (c^1) and local (c^2). Lastly, the degree correlation coefficient (r) drops over the years. This may be caused by lexicographic decisions between

	n	m	z	l	α	c^1	c^2	r
DLE 1925	60,823	1,058,012	17.39	4.03	2.59	0.019	0.227	0.042
DLE 1956	69,719	1,174,912	17.49	4.03	2.58	0.017	0.225	0.039
DLE 2014	87,255	1,076,377	12.34	4.09	2.65	0.015	0.224	0.002

Table 7.1: Basic measures for the networks of Spanish dictionary (DLE) along the years. The Online English dictionary OPTED and WordNet networks are show for comparison. n and m are number of nodes and edges respectively.

editions, e.g. the removal of adverbs with the suffix *-mente* or past participles of verbs.

For the Spanish dictionary network (Table 7.2), the Giant Strongly Connected Component for all three editions keeps it size around 30% of the whole network. The Bidirectional Component stays around 17% of all the words over the years. The Bidirectional Strongly Connected Component covers about 11% of the network. Finally, one of the strongest notions of connectivity is the subgraph induced by the strongly connected component of triangles. It represents less than the 3% of the network in each dictionary. The ratio of size of each component is consistent over time. The words composing the components is also consistent. Note that around 80% of the words of a component in 1925 remain in the same component in 2014 (Table 7.3).

Resilience in a network refers to the vulnerability or the ability of the network to resist link or node failures. This happens to be a relevant property in dictionary networks. There are several notions of resilience. Here, we use the variation of the size of the largest component as nodes are removed from the network. As we saw in Sections 3.4.2 and 5.2, connected components are an important feature of dictionary networks. We use two approaches to node removal: random choice and high in-degree nodes, the latter meaning the removal of words that occur the most in other definitions. As a baseline, we compare the behavior of dictionary networks with that of a random graph. We use the random graph model proposed by [7] based on the idea of preferential attachment. It is frequently used for language networks comparison [29, 101].

It turns out that removing random nodes produces almost no damage at all. All three dictionaries and random graphs resist the attacks well. The size of the component decreases linearly with respect to the number of nodes removed. On the other hand, dictionary networks and random graphs behave very differently when removing high in-degree nodes. Dictionary networks resist more attacks than random graphs (Figure 7.1). Random graphs decline quickly. Removing just 10% of the high in-degree nodes is necessary to completely destroy

	1925	1956	2014
Original network	$60,823$	$69,719$	$87,255$
SCC	$18,307$	$21,538$	$26,989$
Bidirectional Component	$10,462$	$12,061$	$16,025$
Bidirectional SCC	$6,125$	$7,429$	$11,308$
Triangle SCC	$1,033$	$1,318$	$2,359$

Table 7.2: Component sizes of the Spanish dictionary networks in number of words.

	1925-2014	% of 1925
Original	$54,235$	89.1%
SCC	$15,514$	84.7%
Bidirectional Component	$7,665$	73.3%
Bidirectional SCC	$4,841$	79.0%
Triangle SCC	828	80.2%

Table 7.3: Number of words (and percentage) from 1925 that remain in the same component in 2014.

and scatter a random graph. That is not the case with dictionary networks. The giant component of dictionary networks decreases almost linearly until we remove about a third of the network. From that point forward, the giant component starts to decline rapidly, scattering completely when 37% of the high in-degree nodes are removed. It is important to note that resilience of connectivity of dictionary networks does not rely on frequently-used words that connect the network, but on the high connectivity among all words. One could express this by saying that it is very difficult to completely remove a cloud of close concepts; there will always remain other ways to express them. This seems to be a particular property of dictionary networks, as other real world networks do not show this behavior [52, 31, 82].

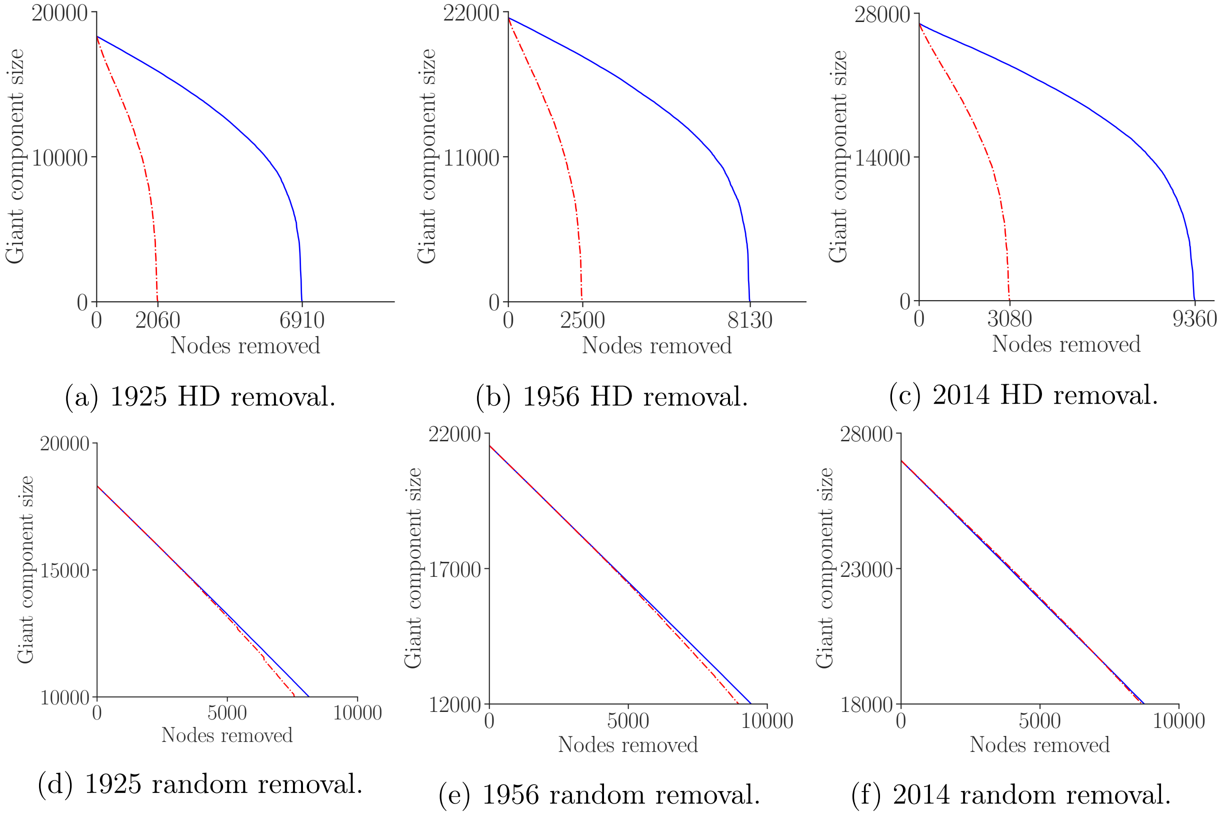
7.2 Definitional and interchangeable entries

The entries in a dictionary can be divided into two groups: *definitional* entries are words used to define other words and *interchangeable* entries correspond to words that do not occur in any definition at all. In network terms, definitional words are those that have inlinks and outlinks while interchangeable words have only outlinks. The fact that a word has only outlinks means that in some sense it is "disposable", that is, it could be replaced by the words in its definition [58], hence the name interchangeable.

In this section, we describe how these groups of words change between editions of the DLE and present a simple model portraying the probabilities of those variations.

If we study how incorporations and deletions of entries from one version of the dictionary to another occur, 8 possible outcomes show up (Figure 7.2). Definitional entries can (1) stay as a definitional entry, (2) become an interchangeable entry, (3) be removed from the dictionary. Likewise, interchangeable entries can (4) stay as an interchangeable entry, (5) become a definitional entry, or (6) be removed from the dictionary. Additionally, new entries are incorporated into the dictionary as (7) new definitional entries or (8) new interchangeable entries.

Most of the entries in a dictionary do not change their type between versions. In fact, in the DLE (versions approximately every 30 years) between 80%-90% of definitional entries stay as definitional and similar percentage of interchangeable entries stay as interchangeable (1 and 4 in Figure 7.2). When new words are added to the dictionary, most of them (76%-95%) enter as interchangeable (8 in Figure 7.2); only a few of them occur in definitions (7 in



(a) 1925 HD removal.

(b) 1956 HD removal.

(c) 2014 HD removal.

(d) 1925 random removal.

(e) 1956 random removal.

(f) 2014 random removal.

Figure 7.1: Sizes of giant component as nodes are removed. On the top, high degree (HD) node removal. DLE network (blue solid line) keeps its structure (giant component) as compared to a random network (red dotted line). On the bottom, random removal does not affect the size of the giant component in both DLE and random network.

Figure 7.2). On the other hand, almost all of the entries that are removed from the dictionary were interchangeable entries (6 in Figure 7.2).

In order to better describe the transitions among the types of words, we build a Markov chain using the empirical data of the transitions over the years (see Figure 7.3). A Markov chain is a stochastic model that describes the transitions between possible states using only its current state. It can be described as a directed graph with probabilities for edges and states for nodes. A word can be in one of three states. It can be a definitional, it can be interchangeable, or it can be "outside". The state outside means that the word is not in the dictionary. This model allows to estimate the probability of a word being in a state in future editions of the Spanish dictionary and the paths it is going to take. For example, a definitional word has a probability $p = 0.9$ of staying as definitional in 30 years in the future (one iteration). If we consider a span of 90 years (three iterations), a definitional word has a probability of $p = 0.729$ (calculated as $0.9 \cdot 0.9 \cdot 0.9$) of staying always as definitional. The model allows us to calculate the probability of more complex transitions. For example, the probability of a definitional word becoming interchangeable in one iteration and then being removed from the dictionary in the next iteration is $p = 0.0135$ (calculated as $0.07 \cdot 0.15$).

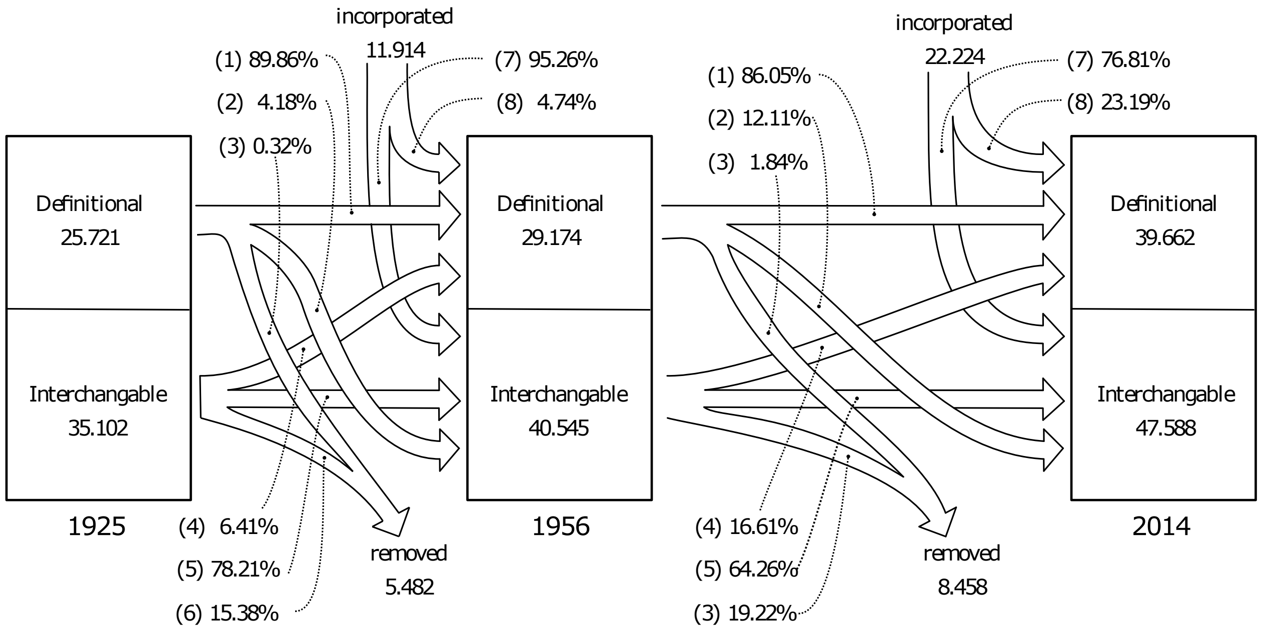


Figure 7.2: Changes in the entries of the DLE from 1925 to 1956 and 1956 to 2014. For a detailed explanation of the figure see Subsection 7.2.

7.3 Examples of simple local changes

There are changes that do not affect or change the overall structure of the network (as we saw in section 3). But they impact at the local level. In fact, these changes alter the structure of the vicinity of some words (not only those whose definition explicitly changes). We will illustrate these changes through some examples in order to offer insights on how the evolution of the network structure speaks about semantic features.

First, entering and outgoing words. *Aeropuerto* (airport) is an obvious case of an entering word that was not present in the 1925 edition. In fact, airplanes and other aerial words were emerging concepts at the time. In 1956, *aeropuerto* is already incorporated as a definitional entry. Later, in 2014, *aeropuerto* is still a definitional entry being used by 17 different words in their definitions, such as airfield (*aeródromo*), checkroom (*consigna*), and tower (*torre*). On the other hand, there are words that slowly were put aside in the dictionary. These words were definitional entries in 1925. In 1956, they became interchangeable entries, as they did not appear in any definition. And in 2014, they were completely removed from the dictionary. Examples are *Adolecente* (old form of adolescent); *fecundante* (who impregnates or fertilizes); *escaza* (an Aragonese word refering to a certain type of pot).

Second, words whose cloud of meaning changes. Consider the word *prostituta* (prostitute). The 1925 dictionary contains the definitional entry *prostituta* defined as *ramera* (whore). There is no definition for the male noun. However, the dictionary contains the interchangeable entry *"prostituto, ta"* (the suffix denotes it can be male or female). This entry refers to the irregular past participle of the verb *prostituir* (prostitute). In the 1956 dictionary, these entries remain with few changes. Both of them keep their definitions, but the entry *prostituta*

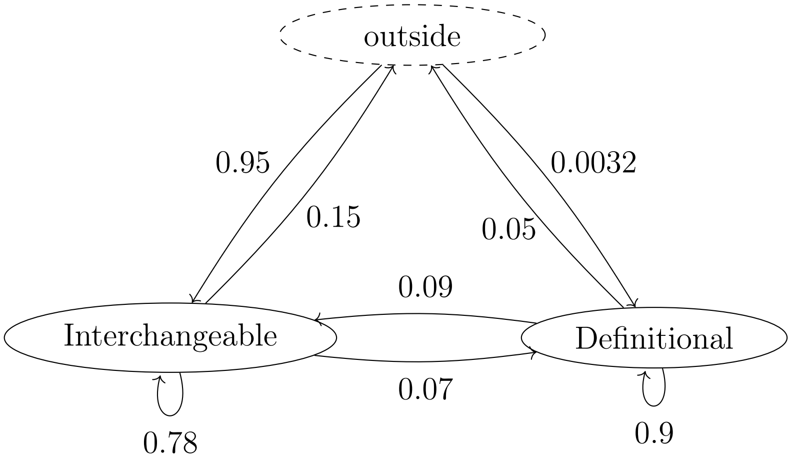


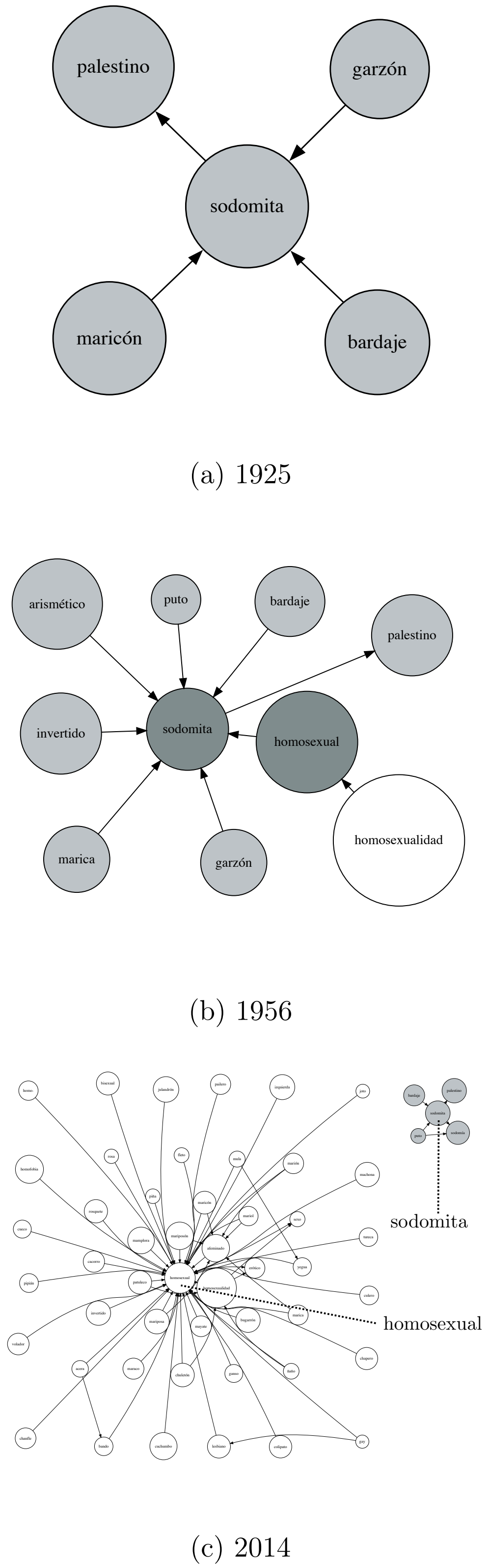
Figure 7.3: Markov chain that describes the probability of transitions among types of words every 30 years in the Spanish Dictionary.

became an interchangeable word. Most of the changes occurred in the 2014 edition. First, the entry *prostituta* was removed from the dictionary. Second, the entry *"prostituto, ta"* became a definitional entry. And third, the entry *"prostituto, ta"* no longer refers to the past participle, but to the noun, covering both the male and female forms. It also got a neutral gender and a less derogatory definition: a woman or man who engages in sexual acts for money.

7.4 More complex local changes

The above changes are not particularly surprising (one could guess them, although in the network they can be detected automatically!). There are more interesting cases that we think would be difficult or virtually impossible to detect without having a network, and thus, demonstrate in some sense the potentialities of the network methodology. A good example is the evolution in the relationship between the words *sexo* (sex) and *sexual* (sexual) and between *homosexual* (homosexual) and *sodomita* (sodomite).

The words *sex* and *sexual* are directly related since the definition of sexual is basically "of or pertaining to sex". However, it is interesting to observe how the relationships between their neighborhoods change. In 1925 (Figure 7.5a), the neighborhood of *sex* is noticeably larger than the neighborhood of *sexual*; moreover, *sex* was surrounded by biological terms, such as plant, walnut, sweet potato, male, female, hermaphrodite, etc. Later in 1956 (Figure 7.5b), the size of the neighborhoods became very similar as *sexual* occurs in more definitions. The neighborhood of *sexual* expanded to a particular subject. Words such as *sperm*, *egg*, *orgasm*, incorporated *sexual* in their definitions. There are many paths between *sex* and *sexual*, but this edition is the first one to have a word that connects them directly (i.e. there is a path of length 2): *masochism* is defined using both *sex* and *sexual.* Now, in 2014, both neighborhoods increase their size (Figure 7.5c), hence their semantic weight. The cloud around *sexual* becomes bigger than that of *sex* and both entries appear where more words connect directly, such as *sexuality*, *venereal*, and *transsexual.*

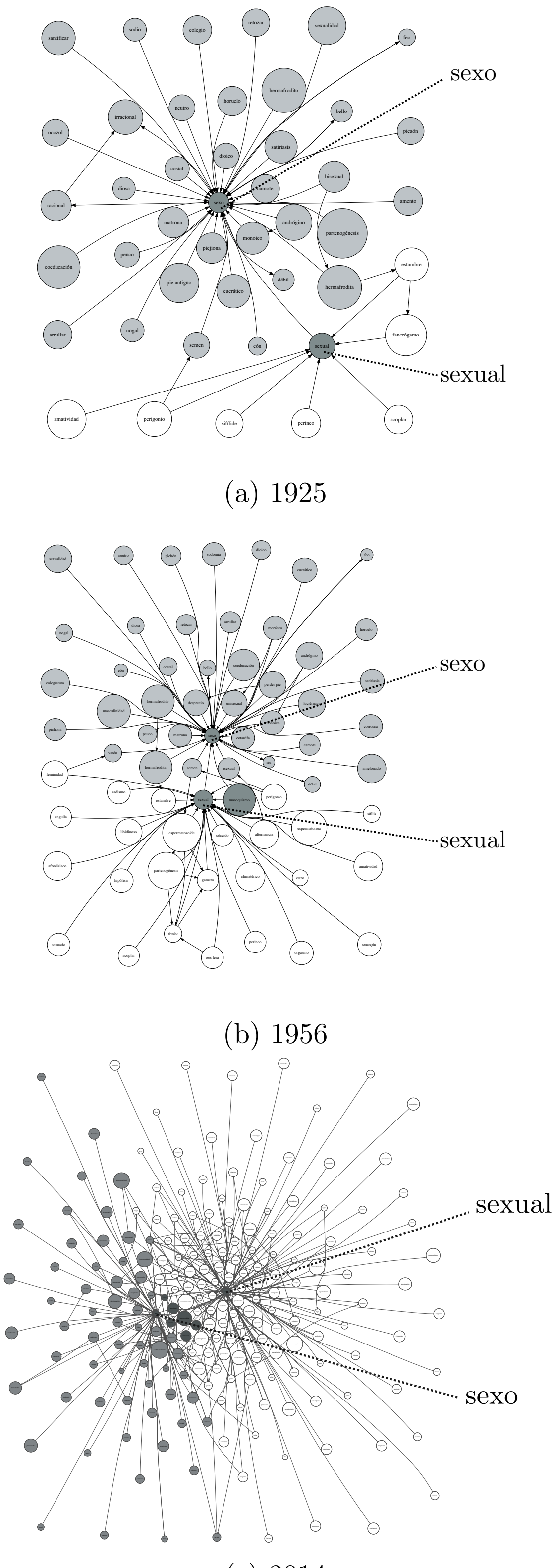


(a) 1925

(b) 1956

(c) 2014

Figure 7.4: Sub-network around the words *homosexual* (homosexual) and *sodomita* (sodomite). White nodes correspond to *homosexual* neighbors and light grey to *sodomita*. Neighbors of both words are dark grey.



(a) 1925

(b) 1956

(c) 2014

Figure 7.5: Sub-network around the words *sexo* (sex) and *sexual* (sexual). White nodes correspond to *sexual* neighbors and light grey to *sexo*. Neighbors of both words are dark grey.

The relationship between *homosexual* (homosexual) and *sodomita* (sodomite) presents a different evolution. In 1925 (Figure 7.4a), *homosexual* was not defined in the dictionary, while *sodomite* occurred as definitional entry. *Sodomite* covered two concepts: a demonym of an old Palestinian city and a person who engages in sodomy. In 1956 (Figure 7.4b), the entry *homosexual* was incorporated into the dictionary as a definitional entry. However, it was not a proper definitional entry. It was incorporated as a synonym of *sodomite*, working as a proxy for other words like *homosexuality* to reach *sodomite*. This situation changed in 2014 (Figure 7.4c), when *homosexual* no longer expressed the meaning of sodomite. It is now defined using concepts such as homosexuality and sexual attraction to persons of the same sex. Its neighborhood grew considerably; more than 50 words use it in their definitions. Lastly, both entries are not connected anymore. Their concepts diverged. *Sodomite* holds the same meaning since 1925 and *homosexual* evolves from not being in the dictionary, passing to be a synonym of *sodomite*, to become an entry with its own meaning. Last but not least, note that in this analysis the use of neighborhoods of the network was essential.

7.5 Conclusion of the Section

In this section, we analyzed the evolution of the Spanish dictionary. One might think that the accumulated changes over one century would have affected the network structure of the dictionary. However, that is not the case. We observe that the changes in the dictionary, such as the incorporation or elimination of definitions and words, do not affect the global structure of the dictionary. In other words, if we remove all the labels from the network, networks from different editions of the Spanish dictionary are practically indistinguishable. On the other hand, local properties are affected by the changes over the years offering insights about the evolution of lexicon, for instance, the relation between the two words and how they change the domains where they reside.

These results show us the kind of linguistic insights that dictionary networks can provide. These insights could not be obtained by just analyzing the text or definitions individually. How words relate to each other play a fundamental role when studying language.

We hypothesize that these results are not limited to the Spanish Dictionary. Dictionary networks in other languages should behave in the same manner, since, as we saw earlier, dictionary networks share a common structure.

Chapter 8

Exploiting Dictionary Networks: A New Class of Proximity Data

In order to show the potentialities of the network methodology over dictionaries, as well as, the semantic information that lies in dictionaries, in this section we show and analyze a particular application of dictionary networks. We combine the collective heritage of dictionaries, a product of the labor of sometimes centuries of evolution of meanings and definitions built by groups of expert linguists (lexicographers), with that of modern techniques of network science, to obtain an automatic method to get clouds of lexical proximity.

Lexical proximity or *proximity data* is the term used in psychology to express the degree of closeness of language notions (entities) in a person's mind. There are many types of proximity and they can be subjective or objective. Examples are synonyms, antonyms, hyponyms, meronyms, categorical and thematic relations, free associations, cause-effect, etc. It is well established how relevant the lexicon and thus, proximity data among lexicon, is for cognition research [79, 80, 27].

There have been many forms in which researchers and linguists have addressed proximity data. Among them are psychological experiments via polls [96, 78]; linguistic analysis [71]; free word associations [65, 30]; development of Thesauri [100]; word embeddings from text corpora [70]. Each of them gives particular types of relations, in most cases reflecting the process used in their construction. Thesauri are built by experts that group words according to similarity of meaning. There is also synonym extraction from dictionary definitions [112] by identifying patterns and using regular expressions that depend on the style and designs of definitions used for different languages by dictionaries. Lexical fields usually are composed manually by experts. Automatic attempts for lexical fields also exploit regularities found in definitions [42]. Word embeddings [70] capture mainly similarity based on context, but they do not capture other lexical relations such as hypernyms and meronyms. They require much bigger corpora than dictionaries. There are also lexical databases such as WordNet [71] and FrameNet [5] that contain proximity data. They are constructed by experts and require dedicated work and time to be built.

The rationale that guides our work is the following. A dictionary can be considered a fair representation (a proxy) of the vocabulary of a language and implicitly represents a network of words. Thus it should be possible to apply current network methods to cluster similar groups of nodes (words) in that network. (The network is small/medium size: A classical dictionary contains about 100,000 nodes). This hypothesis can be validated using the knowledge that speakers of a language have of the semantic relationships among the words in this vocabulary.

Based on this conceptual framework, we present a method to obtain proximity data from the implicit network in a dictionary using classical measures of similarity between the vertices of a graph [56], a technique that has been successful in different applications, e.g. finding similar documents given a query [51], establishing hierarchy in metabolic networks [91] and measuring trophic roles in food webs [60]. To the best of our knowledge, dictionaries have not been systematically used for this goal. It has been proposed in the linguistic field that a dictionary naturally conforms a network of meanings [59, 58]. Based on this, we hypothesize that combining these findings with network theory we can develop an algorithm that will permit us to automatically establish semantic relations among words from a dictionary.

The novelties of the obtained results point to a different class of relations among those known and obtained by purely automatic methods (based on the dictionary) that reflect, as preliminary evaluation shows, relevant associations for people.

The method can be summarized as follows.

Data. We use the *Diccionario de la Lengua Española*[1]. We chose Spanish because the population available to us to validate the results speak Spanish. But none of the steps in the methodology uses or exploit particular features of the Spanish language.

Method. The methodology to obtain clouds of proximity from a dictionary consists of the following steps. First, build a dictionary network from the dictionary and its definitions. Second, obtain the similarity matrix of the vertices using the vertex similarity method proposed by [56]. The principle behind the measure is that two vertices (words) are similar when their immediate neighbors (in the network) are similar, and so recursively. Third, using the similarity matrix so obtained, cluster the words of the dictionary obtaining clouds of proximity.

Validation. We did a preliminary evaluation of the quality of the clouds with 50 people, university students and faculty. We asked them to tell the quality (pertinence, qualitative degree of relation among them, etc.) for 102 different randomly chosen clouds. The experiment shows that the results are highly meaningful for people. The experiment also shows that polysemous words need to be considered to improve the clouds. In the current model, these words often connect different clouds of proximity that should be disconnected.

8.1 Extracting proximity clouds from dictionaries

Our methodology to extract clouds of words that are near each other from the dictionary is divided into three steps. First, we build a dictionary network from the dictionary and its

[1] http://dle.rae.es

definitions. Second, we obtain a similarity matrix of words with a vertex similarity method that only uses the structure of the network. And third, we cluster the words of the dictionary using the similarity matrix, obtaining the clouds of proximity.

8.1.1 Building dictionary network

We use the minimal model of dictionary networks showed in Section 3.2 with some modifications in the cleaning process. The minimal model removes prepositions, conjunctions, interjections, pronouns (personal, demonstrative, possessive, etc.) and article, since they appear too often in any text. In this context, we also consider as stopwords the most frequent words in the dictionary. For example, words such as *be*, *make*, *see*, *state*, and *manner*, appear in more than 3000 definitions in the Open Plain Text English Dictionary. These frequent words introduce much noise into the network, because they appear to be similar to too many words, resulting in wrong clouds of proximity or a giant component that encompasses almost the whole dictionary. This gives a parameter which we had to tune. We determined that removing the 5% most frequent words eliminates a reasonable amount of noise. We can remove the most frequent words after building the dictionary network, removing the vertices with highest degree.

Our methodology is not based on any particular language and does not exploit any linguistic feature inherent to the language of the dictionary being processed (number of words, grammatical information, etc.) at any step. We focus on the structure of the network derived from the dictionary. Since dictionary networks share a common structure [43] our methodology should be able to obtain clouds of proximity regardless of the language of the dictionary.

8.1.2 Getting the similarity between words

To measure the proximity between the words in the dictionary, we compute the similarity between the vertices in the dictionary network. We use the methodology proposed by Leitch in 2006 [56]. They proposed a measure of similarity based on the concept that two vertices are similar if their immediate neighbors in the network are themselves similar. They follow the idea that edges in networks show resemblance or affinity between vertices in the same way we assume that similar entities in social networks are more likely to be connected.

They defined proximity with the following statement: vertex i is similar to j if i has any neighbor k that is itself similar to j. They follow the recursion of the proximity definition, making each vertex similar to itself at the start. The expression for the proximity of the network is:

$$\mathbf{DSD} = \frac{\alpha}{\lambda_1}\mathbf{A}(\mathbf{DSD}) + \mathbf{I} \tag{8.1}$$

where $\mathbf{S}$ is the similarity matrix of the network, $\mathbf{D}$ is the diagonal matrix having the degrees of the vertices in its diagonal elements, $\mathbf{I}$ is the identity matrix, $\mathbf{A}$ is the adjacency matrix ($A_{\mathrm{i}j} = 1$ if there is an edge between i and j, otherwise $A_{\mathrm{i}j} = 0$), λ_1 is the largest eigenvalue of A, and $0 \leq \alpha \leq 1$ is a parameter that regulates the contribution of long paths relative to short ones (with values in the range 0.90 - 0.99 being typical). We compute the similarity matrix iterating the equation until convergence with $\alpha = 0.97$. In order to simplify the computation, we start the iterations with $\mathbf{DSD} = 0$, that being equivalent to start with each vertex only being similar to itself.

There are other vertex similarity methods similar to the one proposed by Leitch in 2006 [56] such as the one proposed by Blondel in 2004 [10] and [51]. However, these methods have the disadvantage that only include paths of even or odd length but not both. This situation may lead in wrong similarity results. In dictionary networks, there is no fundamental difference between vertices connected by a path of odd or even length, thus we would be losing potential similarities between them. Since we use their similarity to join words together, we may fail to find out the connection between two clouds that should be together, extracting smaller and more disperse clouds.

8.1.3 Determining the clouds of proximity

We use the similarity matrix to assemble the clouds. The last step in the methodology is to obtain clouds of proximity. We performed a clustering analysis of the words in the dictionary over the similarity matrix S. We used the hierarchical agglomerative clustering algorithm with average-linkage [116]. We opted for this algorithm because we do not know what are the best sizes of the groups beforehand and because each group could have different sizes. We also opted for average-linkage because we want to cluster words by the sense they represent together, rather than have one representative word. For example, a pair of polysemous words could be close because they are similar in one of their senses, but not in the same sense their respective clusters represents. Single linkage would join these clusters when they should not. A similar phenomenon is discussed later on the evaluation section.

We choose the number of clusters with two objectives. First, we aim to have the least possible number of clusters composed by a single word. We believe every word in the dictionary is related in some way to at least one word. Since a word is defined by other words, it should not be alone in a cluster. Second, we aim to avoid having a giant component holding most of the words in the dictionary. It does not make sense to have most of the dictionary in a cluster. It would not give any information at all.

In practice, for the dictionaries we studied, we found that a number of clusters around $n_c = 2000$ gives better results. With this number of clusters, we have fewer than 20 single word clusters and no cluster has more than 50 words. With a smaller number of clusters we can reduce to zero the number of single word clusters, but the largest clusters rapidly increase their size to over 100 words. On the other hand, a larger number of clusters slowly reduces the size of the largest clusters and rapidly increases the number of single word clusters. For example, in the case of the Spanish dictionary, with a number of clusters $n_c = 2000$ we have 13 single word clusters and the size of the largest cluster is 33. If we reduce the number of clusters to $n_c = 1000$, we have no single word clusters, but the size of the largest cluster increases to 113. If we increase the number of clusters to $n_c = 3000$, the size of the largest cluster is reduced to 28, but we obtain 141 single word clusters.

8.2 Evaluation

Lexical knowledge such as similar words, extended vocabulary, lexical fields, and free association concepts, is knowledge acquired through world and everyday experience. It grows gradually over time [78, 30]. Thus, we need native speakers who speaks the same language as the dictionary used in the evaluation.

Participants. We recruited a total of $n = 50$ native Spanish speakers between ages of 20 and 65. Participants have different backgrounds: 75% of them study or work in Computer Science, 20% of them come from Social Science, and the remaining 5% work in journalism. None of them received compensation of any kind for their participation.

Tasks. To each participant we showed 8 clouds of proximity. For each cloud participants have to complete two tasks.

8.2.1 First Task

The participants were asked to rate the strength of the relation between the words in a proximity cloud. The proximity cloud was presented as the induced subgraph of the words in the dictionary network. Participants were told that a proximity cloud is considered related if they share the same meaning or sense, if they are associated by the context in which they appear or are used, if one word can cue another to come to mind, in other words, if they consider that words are associated in meaning. Participants had to rate a proximity cloud in a discrete scale $\{-2, -1, 0, 1, 2\}$, with -2 meaning that the words of the cloud do not share any relation at all and 2 meaning that the words are highly related.

8.2.2 Second Task

Participants were also asked to mark the words they considered that do not belong to the cloud. A cloud may be considered to have a strong relation between its elements in spite of having one word with little or no relation to the rest of the elements. For example, in the cloud *green, red, blue, hat, cyan, violet* the word *hat* has no relation at all since the relation connecting them is that they are colors. We want to identify these words in the clouds of proximity.

8.3 Materials

We obtained the clouds of proximity from the dictionary *Diccionario de la Lengua Española*[2] (DLE; Dictionary of the Spanish language).

We applied our methodology to DLE using parameter $\alpha = 0.97$ and extracted a total of $n_c = 2000$ clouds of proximity. Then we reduced it to 102 clouds of proximity so each cloud contains one word from a list of randomly selected entries from the dictionary. This list of entries is composed by 51 low frequency words, 37 medium frequency words, and 17 high

[2] http://dle.rae.es

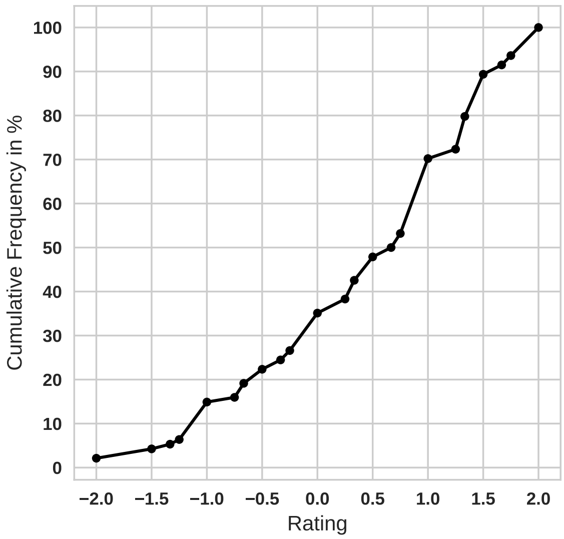


Figure 8.1: Cumulative frequency of proximity clouds obtained from the Spanish dictionary rated by native speakers.

frequency words, so it follows the same frequency distribution as the dictionary. We say that a word is a low frequency word if it appears fewer than 20 times in other definitions, medium frequency if it appears between 20 and 60, and a high frequency if it appears more than 60 times. Since we gave 8 clouds of proximity to each participant, each cloud of words was rated by 4 people.

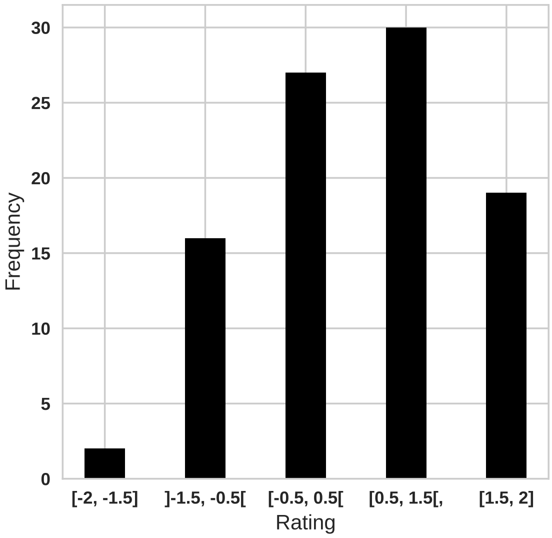


Figure 8.2: Frequency of rating by range of proximity clouds from the Spanish dictionary rated by native speakers.

8.4 First Task Results

There was a high level of agreement in the rating of each cloud. Only 8 clouds had scores with a difference of three or more points between their minimum and their maximum scores. For example, a cloud with scores $\{-1, -1, 0, 2\}$ has a difference of 3 points. Clouds with this level of disagreement were discarded and not considered for the analysis of the evaluation. This restriction left us with 96 well rated clouds of proximity. We consider the average of the scores of a cloud as its final score for the analysis. Participants took an average of 6.5 minutes ($SD = 3.0$) to complete the task.

The primary results of the evaluation are shown in Figure 8.1 and Figure 8.2. Participants rated positively (a score $0 < r$) 64% of the clouds and rated 46% of them with a score $1 \leq r$. On the other hand, 26% of the clouds were rated negatively ($r < 0$), and 15% have scores $r \leq -1$. Clouds had an average score of 0.48 ($SD = 1.02$).

8.4.1 Characteristics of positive rated proximity clouds

The induced subgraphs of positively rated proximity clouds, in contrast to negatively rated clouds, show a more complex structure. The subgraphs have more edges, forming triangles, small cliques and loops. They are composed by smaller clouds of concepts that connect and interact with others. For example, in Figure 8.3, we can observe three smaller clouds. There is a triangle between *redactor, redactar, redacción* (editor, write/edit, and composition/paper respectively). There is another triangle between *impreso, pliego, folleto* (flyer/printout, sheet, and pamphlet/brochure respectively). And there is a small cloud *colaborador, colaborar, colaboración* (collaborator, collaborate, and collaboration respectively). Additionally, we have *periódico* (newspaper) that helps to connect the three concepts.

8.4.2 Characteristics of negative rated proximity clouds

The induced subgraphs of negatively rated proximity clouds share the same network structure. Around 80% of the clouds with a rating $r \leq -1$ have a Path (linear) graph structure or are trees that are almost path graphs (Figure 8.4). In these clouds, neighbor words have relations but they digress quickly because of some senses in definitions. For example, in Figure 8.4 we can start in *remedar* (mimic) and reach *lluvioso* (rainy) following the ideas "mimic" to "not habitual facial expressions" to "rate of occurrence" to "frequent event" to "rainy". However, there is no direct relation between the words in the ends of the paths *remedar* (mimic) and *lluvioso* (rainy). We believe that these sudden changes of context led participants to give a low score to these clouds.

A source for error in the formation of the proximity clouds is polysemy. Since we use a simple model to build the dictionary network, we lose the information about which adjacent words correspond to the same sense. In Figure 8.5, we can observe two clusters that should be disjoint but they are connected by the polysemy of *inerte*. The cloud on the left contains words about "negligence" and the cloud on the right about "silicon". The word *inerte* shares both senses, making it difficult to rate its quality as a whole. Polysemy also causes words to be apparently in the wrong cloud, because sometimes polysemous words can be interpreted as a less frequent sense. A cloud in the evaluation contains the words: mourning, mournful,

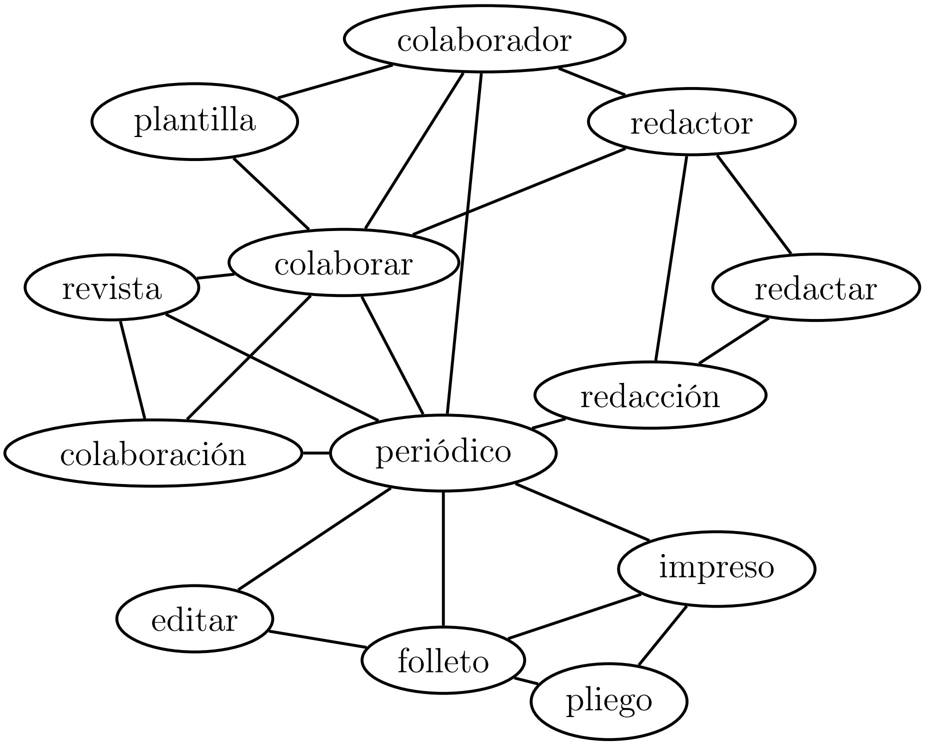


Figure 8.3: Example of a proximity cloud with a high rating by native speakers extracted from the Spanish dictionary.

funeral, pomp, sad, and chia. Chia is an edible seed, but in Spanish chia is also used to refer to a cape used in ancient mourning. This is a low frequent sense of chia, so at first sight it seems as if our methodology gathered together a word totally out of context.

8.5 Second Task Results

The second task asked participants to mark words they consider that do not belong to the cloud. In 47% of the clouds of proximity, at least one word was marked as not belonging to the group. Only a 9.8% of the clouds have more than two words participants consider they do not have any relation to the rest of the words.

We analyzed the induced subgraphs of the clouds that have at least one marked word. We observed that 74% of words marked are terminal vertices or leaves. We can divide these terminal words into two categories:

(I) Polysemous words that are not between the most frequent words of the dictionary or polysemous words with low frequent senses. For example, chia has both mournful and edible senses and the Spanish word *Disco* can mean a discotheque, a piece of furniture for discs, or a collection of music. The multiple senses cause words to be misinterpreted and classified as if they did not belong to the cloud.

(II) Words that are part of a path graph structure. As we mentioned earlier, clouds with a low score usually have the structure of a Path graph. If a cloud of proximity has more than one word that does not belong, in most cases these words form a path graph.

8.6 Discussion

Our main question was: what is the strength of the relationship between the concepts in clouds? Participants showed fair agreement on whether concepts in clouds relate to each other or not. Thus, we can infer that the relationships established by the proximity clouds fit the lexical knowledge of native speakers and correspond to relationships in their language. However, we cannot assume that all the relationships existing in a language show up in the clouds. Native speakers do not manage and are not acquainted with all the vocabulary of their language. Clouds with distant or unrelated words and misplaced words in clouds are easy to identify because they show a particular structure in the induced subgraphs. Often they are path-graphs or leaf vertices, allowing us to exclude these clouds from the outcome of the methodology.

This method to obtain proximity data was presented only to show potentialities of the method and thus argue for the value of the network dictionaries. For example, we did not exploit the distances between vertices (words). We hypothesize that it is possible to use this feature to construct an algorithm that could suggest degrees of proximity among words. Such an algorithm would output a graph with labeled edges indicating the degree of proximity between two words.

The experiments suggest that dictionaries are a rich source of proximity data and that this information can be extracted using standard network analysis techniques over the network of the dictionary. Native speakers show a positive response to the proximity clouds (groups of words related to each other), supporting the hypothesis that dictionaries and their implicit networks are a fair representation of a language. The research and experiments were done using the Spanish Dictionary, but the methods are completely reproducible for any other language dictionary.

This research generated an interesting question that was not answered in this research: From a semantic point of view, what is the relation that keeps these similarity groups of words? Do they represent a new type of "...nym", such as synonyms and meronyms? Or are

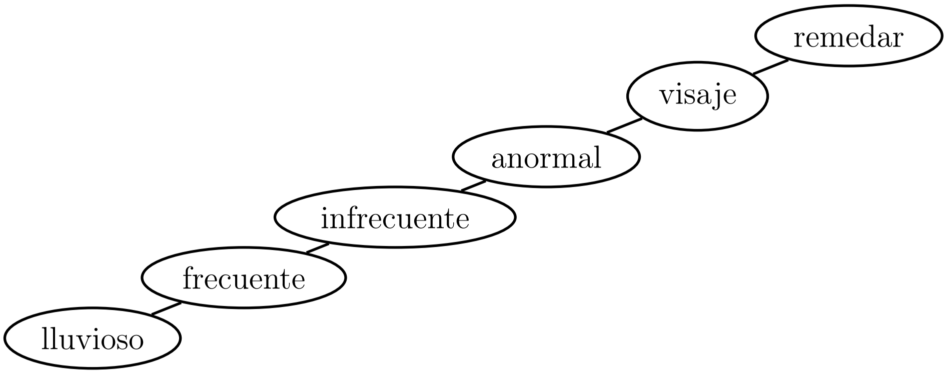


Figure 8.4: Example of a proximity cloud with a low rating by native speakers extracted from the Spanish dictionary.

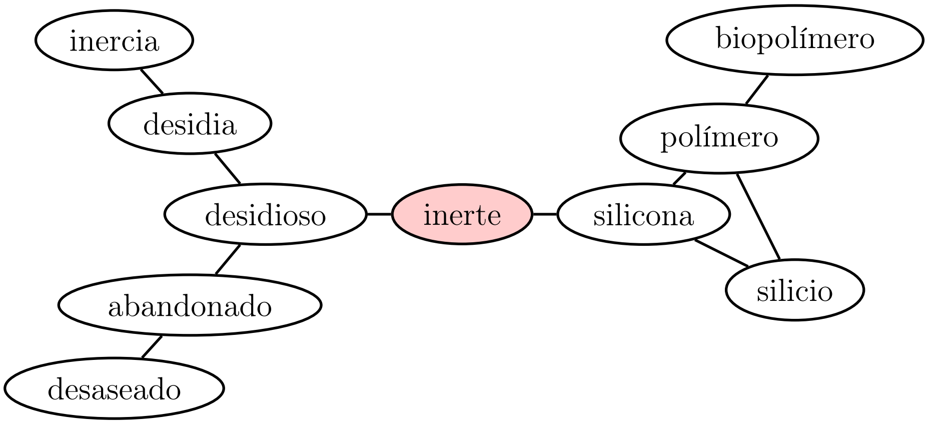


Figure 8.5: Example of cloud of proximity where a polysemous word is connecting two different clouds. The Spanish word *inerte* (inert in English; in red in the graph) has two senses: inactive and lazy. The right side of the cloud is a subcloud composed by words associated to “inactive chemical compounds”. The left side is a subcloud composed by words associated to “lazy”.

they overlapping classical linguistic similarity relations? Or something else?

Chapter 9

Conclusions

In this work, we studied and analyzed dictionary networks to support the hypothesis –formulated by the linguist Kenneth Litkowski [59]– that the underlying lexical network in dictionaries are good sources of material for natural language. The lexical network arises when the semantic information of dictionaries is organized with nodes representing dictionary entries and edges representing the relation "is used to define". In particular, we studied Spanish and English dictionaries. We think that our results provide enough evidence to support this hypothesis. Moreover, it opens a whole area of research as we delineate below.

9.1 Summary of Contributions

Four aspects of dictionary networks captured our attention as a result of this study. First, their basic global and local properties (as compared to other types of networks). Second, the original goal of Litkowski, that is, linguistic features that emerge from them. Third, the regularity in the triadic configuration of dictionaries, allowing us to propose a model to generate such networks. And fourth, the patterns of evolution of dictionary networks over time. These contributions capture one of the main advantages of dictionary networks: their duality between the specific and the whole. They allow us swimmingly observe how local features and changes affect (or do not affect) large pieces of the dictionary.

The structure of dictionaries networks. As for the first contribution, we found that dictionary networks share a common global and local structure. The uniformity of their properties as dictionary networks does not appear to be associated with the language of the dictionary nor the year of publication. This fact points to intrinsic properties of such networks. Their properties are highly noticeable. For example, they show a resilience, much stronger than in most common networks. This is something that could help designers of other types of networks. On the other hand, as a more technical issue, distribution of triads is highly characteristic of dictionary networks. Are there relationships between this observation and the resilience of the networks? This resilience property makes dictionary networks so distinctive that given a plain network (only nodes and edges without labels), one could tell with a high probability if such network corresponds to a network built from a dictionary or not. The converse problem is more difficult: how to design such a network?

We develop a model to generate random networks that produces a similar structure as a dictionary, particularly at a local level. To refine such a model seems to be non trivial and we think is a valuable line of research.

Notion of relevance of meaning. Regarding linguistic features, we found that the most interesting graph notion that naturally interrogates linguistics is centrality. The manifold notions available call for a systematic exploration of them in dictionary networks. Centrality measures have 40% precision for detecting Ogden's words, a good result considering that there are almost 100,000 words in the network. However, there is no tendency for one particular notion. Is that there is no particular notion of "centrality" (as in social networks and other areas) in the lexicon of a language? Or is it that at that level of reduction, less than 1%, there is no difference in their behaviour? We leave the questions.

Model of random dictionary network. As for the third contribution, we presented a model to generate artificial dictionary networks. This is a refinement of classical models for social networks, such as the Barabási–Albert model [7] that performed poorly for dictionaries. It simulates the process performed by expert lexicographers when a word is added to an actual dictionary. This model captures the most salient features of dictionary networks, particularly power law and triad configuration, better than existing models.

Evolution of dictionaries. Regarding evolution of dictionaries, their networks reveal that their structure remains largely unchanged (one could hypothesize that there is a basic structure playing here). The incorporation of new words and the removal of others do not alter this stable structure. What really produces the changes are the evolution of the subgraphs that surround each node (word), evolution that gives each word a subtle and evolving semantic network of meanings.

9.2 Limitations and future guidelines

A clear limitation of this work is that we studied systematically only two dictionaries, *Diccionario de la Lengua Española* (Spanish) and *Online Plain Text English Dictionary*, although we used a few others for particular aspects or comparisons. We hypothesize that the results obtained in this work are valid beyond this limitation as we have not exploited any particular feature to these languages. It would be interesting to replicate this study with dictionaries of other languages of the same branch or family (German, French or Italian) and languages with greater lexical distances (Finish or Russian). Deeper analysis with dictionaries of different languages, especially linguistically distant ones, could allow us to corroborate, or otherwise criticize, the conclusion that the similarity between dictionary network structures does not depend on the specific language but seems to be a property of the lexicon in general. Similarly, the study of other types of dictionaries, such as specialized ones, could help us understand to what extent the properties of dictionary networks are related to the natural language lexicon and not, for example, to all types of lexicon or Lexicographic techniques.

As we would like to expand our study to more dictionaries, we also would like to deepen the analysis on a particular dictionary. Some future work tracks are the following:

Weighted model In this thesis we adopted a unweighted bag of words model for building the network. We would like to compare different weighted graph models. For example, taking into account the order of words in the definitions and words senses, the relationship between same word classes, or using information retrieval measures such as TF-IDF.

Semantic properties We showed that the macrostructure of the dictionary is divided into several components, mainly in a weakly and a strongly connected component. We believe that studying the the difference in the semantic properties of words and relationships between those two components can lead us to a better understanding on how words move from one component to another, and thus, how the dictionary evolves.

Taxonomy of words In the same line, the configuration of the graph makes it possible to distinguish sub-networks that at first do not seem to be based on paradigmatic relationships. From this two questions arise: Is it possible to make a taxonomy of these relationships? Is there a way to compare these relationships with those relationships that we establish between words in our heads?

Mental lexicon Regarding our minds, the study of dictionaries could be further advanced with a finer analysis of word classes. In this thesis we excluded grammatical words, but probably this procedure was not able to remove all words with a syntagmatic function. We hypothesize that focusing on the word classes would help us in the analysis of the mental lexicon and the organization of the concepts in our minds.

All in all, we hope to have convinced the reader that, as the linguist Kenneth Litkowski hypothesized 42 years ago, dictionary networks are are a rich source of linguistic information, and merit further study.

Bibliography

[1] Keith Allan. *Natural language semantics*. Blackwell Publishing, 2001.

[2] Robert A Amsler. A taxonomy for English nouns and verbs. In *Proceedings of the 19th annual meeting on Association for Computational Linguistics*, pages 133–138. Association for Computational Linguistics, 1981.

[3] Robert Alfred Amsler. The structure of the Merriam-Webster pocket dictionary. 1980.

[4] Joseph L. Austerweil, Joshua T Abbott, and Thomas L. Griffiths. Human memory search as a random walk in a semantic network. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 3041–3049. Curran Associates, Inc., 2012.

[5] Collin F Baker, Charles J Fillmore, and John B Lowe. The berkeley framenet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1*, pages 86–90. Association for Computational Linguistics, 1998.

[6] Margarita Balamakova. David Crystal, A Dictionary of Linguistics and Phonetics (5th edn.). Oxford: Blackwell Publishing, 2003. Pp. 508. ISBN 0 631 22664 8. *Journal of the International Phonetic Association*, 34(1):100–101, 2004.

[7] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

[8] Alain Barrat and Martin Weigt. On the properties of small-world network models. *The European Physical Journal B-Condensed Matter and Complex Systems*, 13(3):547–560, 2000.

[9] Vladimir Batagelj, Andrej Mrvar, and Matjaž Zaveršnik. *Network analysis of dictionaries*. University of Ljubljana, Inst. of Mathematics, Physics and Mechanics, Department of Theoretical Computer Science, 2002.

[10] Vincent D Blondel, Anahí Gajardo, Maureen Heymans, Pierre Senellart, and Paul Van Dooren. A measure of similarity between graph vertices: Applications to synonym extraction and web searching. *SIAM review*, 46(4):647–666, 2004.

[11] Vincent D Blondel and Pierre P Senellart. Automatic extraction of synonyms in a

dictionary. *vertex*, 1:x1, 2002.

[12] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *CoRR*, abs/1607.04606, 2016.

[13] Tom Bosc and Pascal Vincent. Learning word embeddings from dictionary definitions only. In *Proceedings of the NIPS 2017 Workshop on Meta-Learning*, 2017.

[14] Nicoletta Calzolari. An empirical approach to circularity in dictionary definitions. *Cahiers de Lexicologie Paris*, 31(2):118–128, 1977.

[15] Nicoletta Calzolari. Detecting patterns in a lexical data base. In *Proceedings of the 10th International Conference on Computational Linguistics and 22nd annual meeting on Association for Computational Linguistics*, pages 170–173. Association for Computational Linguistics, 1984.

[16] Nicoletta Calzolari and Eugenio Picchi. Acquisition of semantic information from an on-line dictionary. In *Proceedings of the 12th conference on Computational linguistics-Volume 1*, pages 87–92. Association for Computational Linguistics, 1988.

[17] Cristian Cardellino. Spanish Billion Words Corpus and Embeddings, March 2016.

[18] José Cañete, Gabriel Chaperon, Rodrigo Fuentes, and Jorge Pérez. Spanish Pre-Trained BERT Model and Evaluation Data. In *to appear in PML4DC at ICLR 2020*, 2020.

[19] Xinxiong Chen, Zhiyuan Liu, and Maosong Sun. A unified model for word sense representation and disambiguation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1025–1035, 2014.

[20] Martin S Chodorow, Roy J Byrd, and George E Heidorn. Extracting semantic hierarchies from a large on-line dictionary. In *Proceedings of the 23rd annual meeting on Association for Computational Linguistics*, pages 299–304. Association for Computational Linguistics, 1985.

[21] Graham Clark. Recursion through dictionary definition space: Concrete versus abstract words. *On WWW at http://www. ecs. soton. ac. uk/Âharnad/Temp/concreteabstract. pdf. Accessed*, 23(06), 2003.

[22] Allan M Collins and M Ross Quillian. Retrieval time from semantic memory. *Journal of verbal learning and verbal behavior*, 8(2):240–247, 1969.

[23] Peter Csermely, András London, Ling-Yun Wu, and Brian Uzzi. Structure and dynamics of core/periphery networks. *Journal of Complex Networks*, 1(2):93–123, 10 2013.

[24] Simon De Deyne and Gert Storms. Word associations: Norms for 1,424 Dutch words in a continuous task. *Behavior research methods*, 40(1):198–205, 2008.

[25] Adriano de Jesus Holanda, Ivan Torres Pisa, Osame Kinouchi, Alexandre Souto Martinez, and Evandro Eduardo Seron Ruiz. Thesaurus as a complex network. *Physica A:*

Statistical Mechanics and its Applications, 344(3):530–536, 2004.

[26] Gerard de Melo. Wiktionary-based word embeddings. *Proceedings of MT Summit XV*, pages 346–359, 2015.

[27] Simon John Dennis. A comparison of statistical models for the extraction of lexical information from text corpora. 2003.

[28] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[29] Sergey N Dorogovtsev and José Fernando F Mendes. Language as an evolving word web. *Proceedings of the Royal Society of London B: Biological Sciences*, 268(1485):2603–2606, 2001.

[30] Haim Dubossarsky, Simon De Deyne, and Thomas T Hills. Quantifying the structure of free association networks across the life span. *Developmental psychology*, 53(8):1560, 2017.

[31] Jennifer A Dunne, Richard J Williams, and Neo D Martinez. Food-web structure and network theory: the role of connectance and size. *Proceedings of the National Academy of Sciences*, 99(20):12917–12922, 2002.

[32] Paul Erdos and Alfred Renyi. On random graphs. *Publicationes Mathematicae Debrecen*, 6:290–297, 1959.

[33] Paul Erdos and Alfred Renyi. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, 5(1):17–60, 1960.

[34] Martin G Everett and Stephen P Borgatti. The centrality of groups and classes. *The Journal of mathematical sociology*, 23(3):181–201, 1999.

[35] Katherine Faust. Comparing social networks: size, density, and local structure. *Metodoloski zvezki*, 3(2):185, 2006.

[36] Katherine Faust. 7. Very Local Structure in Social Networks. *Sociological Methodology*, 37(1):209–256, 2007.

[37] Katherine Faust. A puzzle concerning triads in social networks: Graph constraints and the triad census. *Social Networks*, 32(3):221–233, 2010.

[38] Christiane Fellbaum. Wordnet and wordnets. In Alex Barber, editor, *Encyclopedia of Language and Linguistics*, pages 2–665. Elsevier, 2005.

[39] Ángel Fernández, Emiliano Diez, María Ángeles Alonso, and María Soledad Beato. Free-association norms for the Spanish names of the Snodgrass and Vanderwart pictures. *Behavior Research Methods, Instruments, & Computers*, 36(3):577–583, 2004.

[40] Ramon Ferrer i Cancho and Richard V Solé. The small world of human language. *Proceedings of the Royal Society of London B: Biological Sciences*, 268(1482):2261–2265, 2001.

[41] Charles J Fillmore et al. Frame semantics and the nature of language. In *Annals of the New York Academy of Sciences: Conference on the origin and development of language and speech*, volume 280, pages 20–32, 1976.

[42] Thierry Fontenelle. Automatic extraction of lexical-semantic relations from dictionary definitions. In *Proceedings of the 4th International Congress on Lexicography, EURALEX'90*, pages 89–103, 1990.

[43] Camilo Garrido and Claudio Gutierrez. Dictionaries as Networks: Identifying the graph structure of Ogden's Basic English. In *COLING*, pages 3565–3576, 2016.

[44] Camilo Garrido, Ricardo Mora, and Claudio Gutierrez. Group centrality for semantic networks: a swot analysis featuring random walks. *arXiv preprint arXiv:1601.00139*, 2016.

[45] Bruno Gaume, Nabil Hathout, and Philippe Muller. Word sense disambiguation using a dictionary for sense similarity measure. In *Proceedings of the 20th international conference on Computational Linguistics*, page 1194. Association for Computational Linguistics, 2004.

[46] Felix Hill, Kyunghyun Cho, Anna Korhonen, and Yoshua Bengio. Learning to understand phrases by embedding the dictionary. *Transactions of the Association for Computational Linguistics*, 4:17–30, 2016.

[47] Paul W Holland and Samuel Leinhardt. Local structure in social networks. *Sociological methodology*, 7:1–45, 1976.

[48] Paul W Holland and Samuel Leinhardt. A method for detecting structure in sociometric data. In *Social Networks*, pages 411–432. Elsevier, 1977.

[49] Ignacio Iacobacci, Mohammad Taher Pilehvar, and Roberto Navigli. Embeddings for word sense disambiguation: An evaluation study. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 897–907, 2016.

[50] Shalev Itzkovitz, Ron Milo, Nadav Kashtan, Guy Ziv, and Uri Alon. Subgraphs in random networks. *Physical review E*, 68(2):026127, 2003.

[51] Glen Jeh and Jennifer Widom. SimRank: a measure of structural-context similarity. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 538–543. ACM, 2002.

[52] Hawoong Jeong, Sean P Mason, A-L Barabási, and Zoltan N Oltvai. Lethality and centrality in protein networks. *Nature*, 411(6833):41, 2001.

[53] Hideki Kozima and Teiji Furugori. Similarity between words computed by spreading activation on an English dictionary. In *Proceedings of the sixth conference on European chapter of the Association for Computational Linguistics*, pages 232–239. Association for Computational Linguistics, 1993.

[54] Jérôme Kunegis. Konect: the koblenz network collection. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 1343–1350. ACM, 2013.

[55] Irene Langkilde and Kevin Knight. Generation that exploits corpus-based statistical knowledge. In *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*, 1998.

[56] Elizabeth A Leicht, Petter Holme, and Mark EJ Newman. Vertex similarity in networks. *Physical Review E*, 73(2):026120, 2006.

[57] Jure Leskovec and Andrej Krevl. SNAP Datasets: Stanford large network dataset collection. http://snap.stanford.edu/data. Last accessed 10 December 2016.

[58] David Levary, Jean-Pierre Eckmann, Elisha Moses, and Tsvi Tlusty. Loops and self-reference in the construction of dictionaries. *Physical Review X*, 2(3):031018, 2012.

[59] Kenneth C Litkowski. Models of the semantic structure of dictionaries. *American Journal of Computational Linguistics*, 81:25–74, 1978.

[60] Joseph J Luczkovich, Stephen P Borgatti, Jeffrey C Johnson, and Martin G Everett. Defining and measuring trophic role similarity in food webs using regular equivalence. *Journal of Theoretical Biology*, 220(3):303–321, 2003.

[61] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60, 2014.

[62] Judith Markowitz, Thomas Ahlswede, and Martha Evens. Semantically significant patterns in dictionary definitions. In *Proceedings of the 24th annual meeting on Association for Computational Linguistics*, pages 112–119. Association for Computational Linguistics, 1986.

[63] A Blondin Massé, Guillaume Chicoisne, Yassine Gargouri, Stevan Harnad, Olivier Picard, and Odile Marcotte. How is meaning grounded in dictionary definitions? In *Proceedings of the 3rd Textgraphs Workshop on Graph-Based Algorithms for Natural Language Processing*, pages 17–24. Association for Computational Linguistics, 2008.

[64] John Matta, Jeffrey Borwey, and Gunes Ercal. Comparative resilience notions and vertex attack tolerance of scale-free networks. *arXiv preprint arXiv:1404.0103*, 2014.

[65] Yevgen Matusevych and Suzanne Stevenson. Analyzing and modeling free word associations. In *CogSci 2018*, pages 750–755, July 2018.

[66] Rada Mihalcea and Dan I Moldovan. Extended WordNet: Progress report. In *in Proceedings of NAACL Workshop on WordNet and Other Lexical Resources*. Citeseer, 2001.

[67] Rada Mihalcea and Paul Tarau. TextRank: Bringing Order into Text. In *EMNLP*, volume 4, pages 404–411, 2004.

[68] Rada Mihalcea, Paul Tarau, and Elizabeth Figa. PageRank on semantic networks, with application to word sense disambiguation. In *Proceedings of the 20th international conference on Computational Linguistics*, page 1126. Association for Computational Linguistics, 2004.

[69] Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. Advances in Pre-Training Distributed Word Representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[70] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[71] George A Miller. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41, 1995.

[72] George A Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J Miller. Introduction to WordNet: An on-line lexical database. *International journal of lexicography*, 3(4):235–244, 1990.

[73] George A Miller and Walter G Charles. Contextual correlates of semantic similarity. *Language and cognitive processes*, 6(1):1–28, 1991.

[74] Ron Milo, Shai Shen-Orr, Shalev Itzkovitz, Nadav Kashtan, Dmitri Chklovskii, and Uri Alon. Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824–827, 2002.

[75] Adilson E Motter, Alessandro PS De Moura, Ying-Cheng Lai, and Partha Dasgupta. Topology of the conceptual network of language. *Physical Review E*, 65(6):065102, 2002.

[76] Philippe Muller, Nabil Hathout, and Bruno Gaume. Synonym extraction using a semantic distance on a dictionary. In *Proceedings of the First Workshop on Graph Based Methods for Natural Language Processing*, pages 65–72. Association for Computational Linguistics, 2006.

[77] Roberto Navigli and Simone Paolo Ponzetto. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250, 2012.

[78] Douglas L Nelson, Cathy L McEvoy, and Thomas A Schreiber. The University of South

Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, 36(3):402–407, 2004.

[79] Douglas L Nelson, Vanesa M McKinney, Nancy R Gee, and Gerson A Janczura. Interpreting the influence of implicitly activated memories on recall and recognition. *Psychological review*, 105(2):299, 1998.

[80] Douglas L Nelson and Nan Zhang. The ties that bind what is known to the recall of what is new. *Psychonomic Bulletin & Review*, 7(4):604–617, 2000.

[81] Mark EJ Newman. The structure and function of complex networks. *SIAM review*, 45(2):167–256, 2003.

[82] Mark EJ Newman, Stephanie Forrest, and Justin Balthrop. Email networks and the spread of computer viruses. *Physical Review E*, 66(3):035101, 2002.

[83] Yoshiki Niwa and Yoshihiko Nitta. Co-occurrence vectors from corpora vs. distance vectors from dictionaries. In *Proceedings of the 15th conference on Computational linguistics-Volume 1*, pages 304–309. Association for Computational Linguistics, 1994.

[84] Beth A Ober and Gregory K Shenaut. Semantic memory. In *Handbook of Psycholinguistics*, pages 403–453. Elsevier, 2006.

[85] Charles Kay Ogden. *Basic English: A general introduction with rules and grammar*, volume 29. K. Paul, Trench, Trubner, 1930.

[86] John Olney, Carter Revard, and Paul Zeff. Processor for Machine-Readable Version of Webster's Seventh at System Development Corporation. *The Finite String*, 4(3):1–2, 1967.

[87] John C Olney. To all interested in the Merriam-Webster transcripts and data derived from them. *Systems Development Corporation Document L-13579*, 1968.

[88] Siddharth Patwardhan and Ted Pedersen. Using WordNet-based context vectors to estimate the semantic relatedness of concepts. In *Proceedings of the Workshop on Making Sense of Sense: Bringing Psycholinguistics and Computational Linguistics Together*, 2006.

[89] Olivier Picard, Alexandre Blondin-Massé, Stevan Harnad, Odile Marcotte, Guillaume Chicoisne, and Yassine Gargouri. Hierarchies in dictionary definition space. *arXiv preprint arXiv:0911.5703*, 2009.

[90] Sabine Ploux and Hyungsuk Ji. A model for matching semantic maps between languages (French/English, English/French). *Computational linguistics*, 29(2):155–178, 2003.

[91] Erzsébet Ravasz, Anna Lisa Somera, Dale A Mongru, Zoltán N Oltvai, and A. Barabási. Hierarchical organization of modularity in metabolic networks. *science*, 297(5586):1551–1555, 2002.

[92] Richard Reichert, John Olney, and James Paris. Two dictionary transcripts and programs for processing them. Volume I. The encoding scheme, parsent and conix. Technical report, System Development Corp. Santa Monica California, 1969.

[93] Peter Mark Roget. Roget's Thesaurus of English Words and Phrases. http://www.gutenberg.org/etext/10681, 1911. Last accessed 01 July 2017.

[94] Sascha Rothe and Hinrich Schütze. Autoextend: Extending word embeddings to embeddings for synsets and lexemes. *arXiv preprint arXiv:1507.01127*, 2015.

[95] Thijs Scheepers, Evangelos Kanoulas, and Efstratios Gavves. Improving word embedding compositionality using lexicographic definitions. In *Proceedings of the 2018 World Wide Web Conference*, pages 1083–1093, 2018.

[96] Roger W Schvaneveldt, Francis T Durso, and Donald W Dearholt. Network structures in proximity data. *The psychology of learning and motivation*, 24:249–284, 1989.

[97] John Maynard Smith and Eors Szathmary. *The major transitions in evolution*. Oxford University Press, 1997.

[98] Ricard V Solé, Bernat Corominas-Murtra, Sergi Valverde, and Luc Steels. Language networks: Their structure, function, and evolution. *Complexity*, 15(6):20–26, 2010.

[99] Karen Spärck Jones. Dictionary Circles. Technical report, System Development Corp. Santa Monica California, 1967.

[100] Armando Stellato, Sachit Rajbhandari, Andrea Turbati, Manuel Fiorelli, Caterina Caracciolo, Tiziano Lorenzetti, Johannes Keizer, and Maria Teresa Pazienza. VocBench: a web application for collaborative development of multilingual thesauri. In *European Semantic Web Conference*, pages 38–53. Springer, 2015.

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

[102] Chiraag Sumanth and Diana Inkpen. How much does word sense disambiguation help in sentiment analysis of micropost data? In *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 115–121, 2015.

[103] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.

[104] Kaveh Taghipour and Hwee Tou Ng. Semi-supervised word sense disambiguation using word embeddings in general and specific domains. In *Proceedings of the 2015 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 314–323, 2015.

[105] Julien Tissier, Christophe Gravier, and Amaury Habrard. Dict2vec: Learning word embeddings using lexical dictionaries. 2017.

[106] University of Texas. eXtended WordNet. http://www.hlt.utdallas.edu/~xwn/about.html, 2003. Last accessed 01 July 2017.

[107] Jaan Valsiner and Kevin J Connolly. *Handbook of developmental psychology*. Sage, 2002.

[108] Jean Veronis and Nancy M Ide. Word sense disambiguation with very large neural networks extracted from machine readable dictionaries. In *Proceedings of the 13th conference on Computational linguistics-Volume 2*, pages 389–394. Association for Computational Linguistics, 1990.

[109] Vaira Vikis-Freibergs and Imants Freibergs. Free association norms in French and English: Inter-linguistic and intra-linguistic comparisons. *Canadian Journal of Psychology/Revue canadienne de psychologie*, 30(3):123, 1976.

[110] Philippe Vincent-Lamarre, Alexandre Blondin Massé, Marcos Lopes, Mélanie Lord, Odile Marcotte, and Stevan Harnad. The latent structure of dictionaries. *Topics in cognitive science*, 8(3):625–659, 2016.

[111] Tong Wang and Graeme Hirst. Extracting Synonyms from Dictionary Definitions. In *RANLP*, pages 471–477, 2009.

[112] Tong Wang and Graeme Hirst. Exploring patterns in dictionary definitions for synonym extraction. *Natural Language Engineering*, 18(3):313–342, 2012.

[113] Duncan J Watts. Networks, dynamics, and the small-world phenomenon. *American Journal of sociology*, 105(2):493–527, 1999.

[114] Yorick Wilks, Dan Fass, Cheng-Ming Guo, James E McDonald, Tony Plate, and Brian M Slator. Providing machine tractable dictionary tools. *Machine translation*, 5(2):99–154, 1990.

[115] Eric Yeh, Daniel Ramage, Christopher D. Manning, Eneko Agirre, and Aitor Soroa. WikiWalk: Random walks on Wikipedia for semantic relatedness. In *Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing (TextGraphs-4)*, pages 41–49, Suntec, Singapore, August 2009. Association for Computational Linguistics.

[116] Marie Lisandra Zepeda-Mendoza and Osbaldo Resendis-Antonio. *Hierarchical Agglomerative Clustering*, pages 886–887. Springer New York, New York, NY, 2013.

[117] Ingrid Zukerman, Sarah George, and Yingying Wen. Lexical paraphrasing for document retrieval and node identification. In *Proceedings of the second international workshop on Paraphrasing*, pages 94–101, 2003.